Dual-branch Network for Species Identification via Passive Acoustic Monitoring

Notebook for the <BirdCLEF> Lab at CLEF 2025

Jingyin Tan[1], Aiguo Wang[1,*]

[1]School of Computer Science and Artificial Intelligence, Foshan University, China

Abstract

The BirdCLEF2025 task aims to train a multi-label classifier to infer the presence probabilities of multiple species from the audio signals. In this work, we introduce a dual branch architecture to build an end-to-end passive acoustic monitoring predictor. Specifically, two different types of acoustic features (i.e., Mel-spectrogram and Mel-Frequency Cepstral Coefficients) are first extracted from the raw signals. ResNet and ConvNeXt are then used to learn two-branch features. Afterwards, the features are concatenated and fed into a fully connected layer to output the prediction probabilities. Finally, we conduct comparative experiments on the competition test data. Experimental results show that the proposed model achieves 0.751 and 0.771 macro-mean ROC-AUC on the 34% and 66% test dataset, respectively.

Keywords

species identification, dual-branch, acoustic features

1. Introduction

The identification of under-studied species via passive acoustic monitoring is less affected by weather and more detectable in enhancing biodiversity monitoring[1], compared with conventional observer-based biodiversity surveys[2]. The BirdCLEF2025[3, 4] competition aims to predict the presence of each of 206 target species in every 5-second segment of audio recordings, which is an important application scenario of passive acoustic monitoring. Accordingly, researchers have explored various methods in BirdCLEF2024 towards higher accuracy. For example, the approach in [5] utilizes a transfer learning method based on pseudo multi-labels, demonstrating the effectiveness of leveraging pre-trained embeddings for birdcall classification. The method in [6] designs an ensembled model that is a combination of EfficientNet-B0 and EfficientNet-B1 to leverage the strengths of different models.

Although previous methods have achieved promising results, they overlook the usage of multi-channel features and the impact of pretrained weights on the feature mapping backbones. To this end, we in this work propose a dual-branch neural network that uses two types of acoustic features to better learn the latent meaningful features. The main contributions of this work are outlines are follows.

(1) A dual-branch neural network is proposed to build an end-to-end passive acoustic monitoring predictor to identify species. Two types of features, including Mel-spectrogram (Mel) and Mel-Frequency Cepstral Coefficients (MFCC) are extracted from raw audio signals, which are then fed into two typical pretrained feature representation networks (i.e., ResNet and ConvNeXt).

(2) We conduct comparative experiments to evaluate the effectiveness of the proposed model. Particularly, we evaluate different ways of initializing the parameters of ResNet and ConvNeXt. Results show that ResNet backbone with pretrained weights and ConvNeXt with random weights strategy outperforms others, with scores of 0.751 and 0.771, respectively (Team name: Hathaway Tan, Rank: 1455th).

The structure of this paper is as follows. Section 2 details the proposed model. Section 3 introduces the datasets, preprocessing steps and presents experimental results, followed by the conclusion section.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

tjynicky@gmail.com (J. Tan); agwang@fosu.edu.cn (A. Wang)

2. Methodology

Figure 1 presents the end-to-end species identification model via passive acoustic monitoring. It mainly consists of the training phase and test phase. During the training stage, different types of acoustic features are extracted from the raw signals. Then, the two types of features are fed into two powerful feature mapping networks. Afterwards, the learned features are concatenated and sent to a fully connected layers to generate prediction probabilities. During the test stage, test data is predicted with the model and output prediction probabilities.

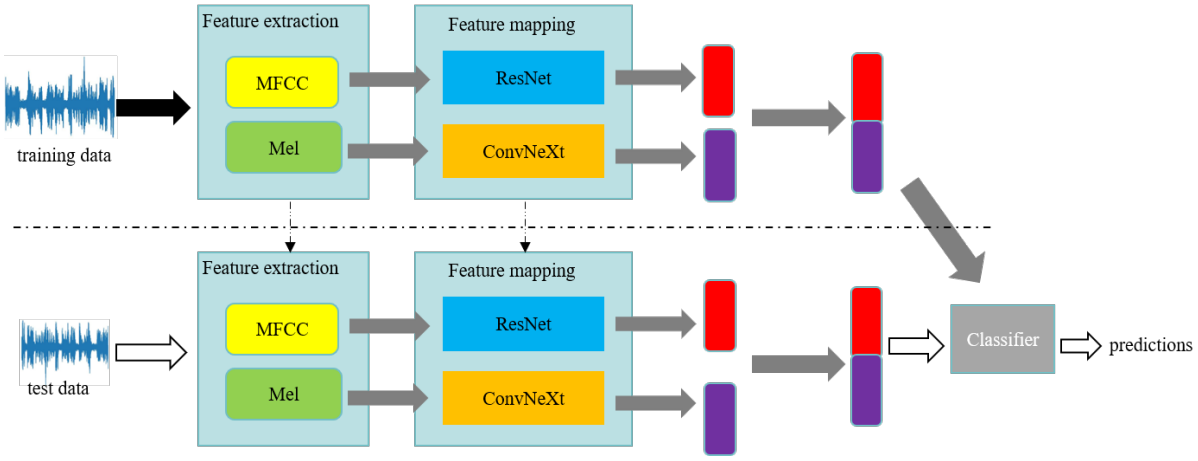

Figure 1: The proposed dual-branch species identification model

2.1. Feature extraction

Considering the synthetic effect of different types of acoustic features in analyzing signals[7], we in this work explore two types of features (Mel-spectrogram and Mel-Frequency Cepstral Coefficients) to take advantage of multi-channel feature representation.

2.2. Pretrained weight analysis

Recent studies indicate that pretrained models in the context of feature mapping have been widely used because of its effectiveness in accelerating training[8], yet they tend to suffer from limited accuracy due to the small size of the fine-tuning dataset in the downstream task[9]. Hence, we also conduct comparative experiments to evaluate the impact of pretrained weights on the feature mapping backbones.

3. Experimental setup and results

3.1. Dataset

The BirdCLEF2025 training audio dataset consists of a total of 28,564 audios, covering 206 unique species across four major taxonomic classes. Each audio recording has a duration ranging from 0.54s to 1774s. Table 1 presents the summary of dataset and Figure 2 displays the top 20 species by training audios.

3.2. Experimental setup

To increase the number of samples for training, we use the sliding window without overlapping to segment the raw audio data into 10 seconds slices, where zero padding is applied to extend the data length if the original audio data is shorter than 10s. Finally, we get 78,579 segments in total. Figure 3 shows an example of data segment from an audio file in training dataset.

Table 1
Summary of training dataset

Taxonomic	Number
Aves	27648
Amphibia	583
Mammalia	178
Insecta	155

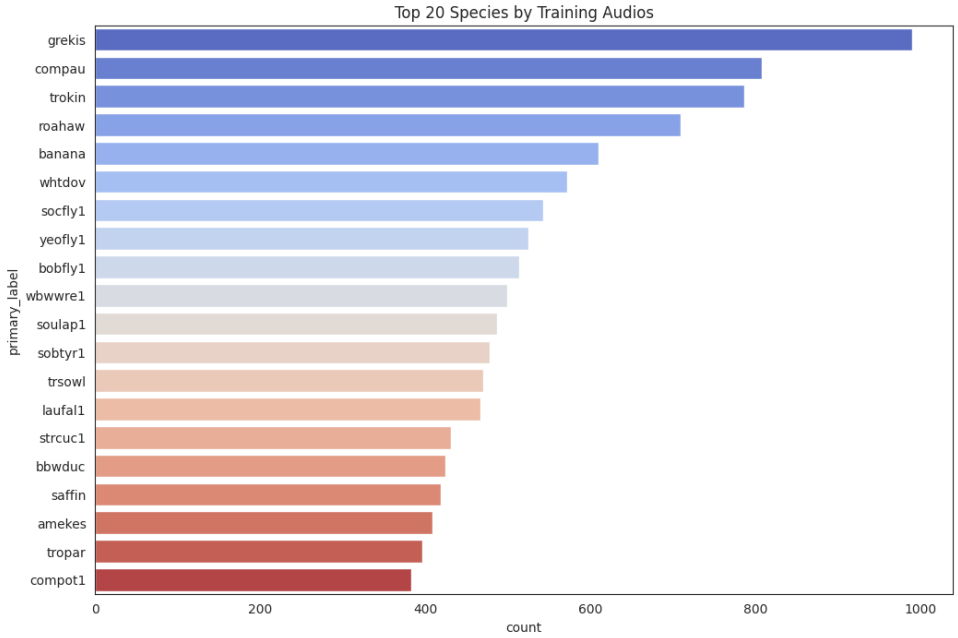

Figure 2: Top 20 species by training audios

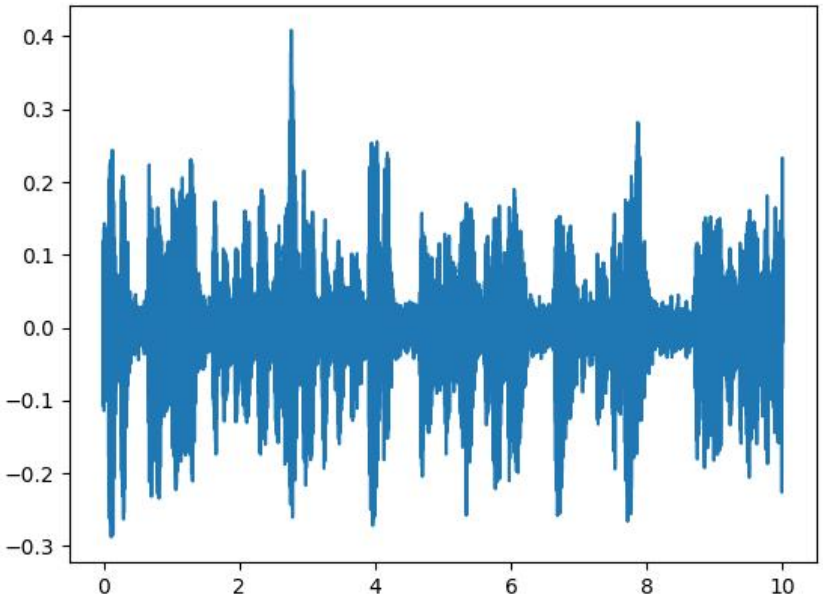

Figure 3: A 10-second segment from audio file 1139490-CSA36385

We then extract the MFCC and Mel features from each of the segments. Specifically, the MFCC features consist of the 13-dimensional MFCCs, first-order and second-order derivatives. For Mel-spectrogram features, we extract 128 Mel frequency bands per frame, resulting in a 128-dimensional feature vector. We set the FFT window size to 1024 and the hop length to 512. These parameters define the time-frequency resolution when computing the MFCC and Mel. We normalize the MFCCs and Mels sample by sample using Z-Score approach. The MFCCs and Mels are finally reshaped to 224 x 224 pixels.

For the training procedure, we utilize ResNet50.a1_in1k pretrained on ImageNet-1k dataset[10, 11], and use ConvNeXtv2_pico.fcmae[12]. The loss function is BCEWithLogitsLoss. The model totally runs 30 epochs with an early-stop strategy avoiding overfitting. We fine-tune the end-to-end model with the AdamW optimizer. An initial learning rate 0.001 is used and a cosine annealing learning rate scheduler is utilized, which adjusts the learning rate following a cosine curve from the initial value down to 1e-6.

As for the performance metric, a version of macro-averaged ROC-AUC that skips classes which have no true positive labels is used. We employ 3-fold stratified cross-validation for training. During training, the model achieving the highest average AUC on the validation set in each fold is saved.

During the test stage, prediction is conducted on the hidden test set, which transforms the model's outputs to multi-class probabilities and calculates AUC scores using predictions and real multi-class labels. The submission format requires that the length of each test audio segment is 5 second, so we concatenate the 5 second segment with itself to create a 10 second slice. We opt not to train directly on 5-second segments due to the limited acoustic context they offer, which can negatively impact classification performance.

3.3. Experimental results

Tables 2 and 3 present the results of different weight-using strategies on 34% and 66% test data respectively. In the table, "fine-tuned" means the feature mapping networks are equipped with pretrained weights; "from scratch" indicates that the parameters of feature mapping networks are randomly initialized.

The experimental results presented in Tables 2 and 3 show the AUC scores of various model combinations on two different test datasets (34% and 66%). Across both datasets, the model **ResNet_ft(fine-tuned)+ConvNeXt_fs(from scratch)** consistently achieves the highest AUC scores — **0.751** on the 34% test set and **0.771** on the 66% test set — indicating superior performance. In contrast, the model ResNet_fs(from scratch)+ConvNeXt_ft(fine-tuned) performs the worst. Interestingly, the combination ResNet_fs(from scratch)+ConvNeXt_ft(fine-tuned) performs worse than using ConvNeXt_fs(from scratch), which may imply that the feature in ConvNeXt has a less significant or even slightly detrimental effect when ResNet is not enhanced.

Table 2
Experimental results on the 34% test data

Model	ROC-AUC score
ResNet_ft(fine-tuned)+ConvNeXt_ft(fine-tuned)	0.738
ResNet_ft(fine-tuned)+ConvNeXt_fs(from scratch)	**0.751**
ResNet_fs(from scratch)+ConvNeXt_ft(fine-tuned)	0.722
ResNet_fs(from scratch)+ConvNeXt_fs(from scratch)	0.732

3.4. Discussion

Our study reveals the following performance ordering: ResNet pretrained only > both pretrained > no pretraining > ConvNeXt pretrained only. These findings shows that visual pretrained models transfer well to spectrogram's low-level texture and edge features but require close alignment between input representation and pretrained domain. Based on our findings, we recommend using pretrained

Table 3
Experimental results on the 66% test data

Model	ROC-AUC score
ResNet_ft(fine-tuned)+ConvNeXt_ft(fine-tuned)	0.763
ResNet_ft(fine-tuned)+ConvNeXt_fs(from scratch)	**0.771**
ResNet_fs(from scratch)+ConvNeXt_ft(fine-tuned)	0.726
ResNet_fs(from scratch)+ConvNeXt_fs(from scratch)	0.731

ImageNet weights only when the input representation retains visual-like structures, such as Mel spectrograms, which benefit from learned low-level convolutional filters. On the contrary, for more abstract representations like MFCC, we advise against using pretrained visual weights.

4. Conclusion

In this work, we proposed a dual-branch architecture that leverages both Mel-spectrogram and MFCC features, processed through ResNet and ConvNeXt backbones, for passive acoustic monitoring in the BirdCLEF2025 task. Comparative experiments on the competition's test datasets demonstrate the effectiveness of our design, with the model achieving macro-mean ROC-AUC scores of 0.751 on the 34% test set and 0.771 on the 66% test set. These results confirm that the proposed end-to-end framework can effectively capture complementary acoustic information and deliver robust multi-label bird species classification performance.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI-GPT-4o in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Thomas, T. A. Marques, Passive acoustic monitoring for estimating animal density, Acoustics Today 8 (2012) 35–44.

[2] H. Klinck, J. S. Cañas, M. Demkin, S. Dane, S. Kahl, T. Denton, Birdclef+ 2025, https://kaggle.com/competitions/birdclef-2025, 2025. Kaggle.

[3] L. Picek, S. Kahl, H. Goëau, L. Adam, et al., Overview of lifeclef 2025: Challenges on species presence prediction and identification, and individual animal identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025.

[4] J. S. Cañas, S. Kahl, T. Denton, M. P. Toro-Gómez, S. Rodriguez-Buritica, J. L. Benavides-Lopez, J. S. Ulloa, P. Caycedo-Rosales, H. Klinck, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF+ 2025: Multi-taxonomic sound identification in the middle magdalena valley, colombia, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.

[5] A. Miyaguchi, A. Cheung, M. Gustineli, A. Kim, Transfer learning with pseudo multi-label birdcall classification for ds@ gt birdclef 2024, arXiv preprint arXiv:2407.06291 (2024).

[6] A. Porwal, Bird-species audio identification, ensembling of efficientnet-b0 and pre-trained efficientnet-b1 model (2024).

[7] Yaseen, G.-Y. Son, S. Kwon, Classification of heart sound signal using multiple features, Applied Sciences 8 (2018) 2344.

[8] K. He, R. Girshick, P. Dollár, Rethinking imagenet pre-training, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4918–4927.

[9] H. Liu, M. Long, J. Wang, M. I. Jordan, Towards understanding the transferability of deep representations, arXiv preprint arXiv:1909.12031 (2019).
[10] R. Wightman, H. Touvron, H. Jégou, Resnet strikes back: An improved training procedure in timm, arXiv preprint arXiv:2110.00476 (2021).
[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[12] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, S. Xie, Convnext v2: Co-designing and scaling convnets with masked autoencoders, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 16133–16142.