Application of Deep Learning techniques in the Classification of Bird Audio for Environmental Monitoring in Doñana

Alba Márquez-Rodríguez[a], Manuel Jesús Marín Jiménez[a], Miguel Ángel Muñoz Mohedano[b]

[a] *Universidad de Córdoba, Córdoba, Córdoba, Spain*
[b] *Estación Biológica de Doñana, Sevilla, Sevilla, Spain*

Abstract

Passive acoustic monitoring (PAM) that uses devices like automatic audio recorders has become a fundamental tool in conserving and managing natural ecosystems. However, this practice generates a large volume of unsupervised audio data, and extracting valid information for environmental monitoring is a significant challenge. It is then critically necessary to use methods that leverage Deep Learning techniques for automating species detection. BirdNET is a model trained for bird identification that has succeeded in many study systems, especially in North America or Europe, but it results inadequate for other regions due to insufficient training and its bias on focal sounds rather than entire soundscapes. Another added problem for species detection is that many audios recorded in PAM programs are empty of sounds of species of interest or these sounds overlap. This study presents a multi-stage process for automatically identifying bird vocalizations that includes first a YOLOv8-based Bird Song Detector, and second, a fine-tuned BirdNET for species classification at a local scale with enhanced detection accuracy. As a study case, we applied this Bird Song Detector to audio recordings collected in Doñana National Park (SW Spain) as a part of the BIRDeep project. We annotated 461 minutes of audio data from three main habitats across nine different locations within Doñana, resulting in 3749 annotations representing 38 different classes. Mel spectrograms were employed as graphical representations of bird audio data, facilitating the application of image processing methods. Several detectors were trained in different experiments, which included data augmentation and hyperparameter exploration to improve the model's robustness. The model giving the best results included the creation of synthetic background audios with data augmentation and the use of an environmental sound library. This proposed pipeline using the Bird Song Detector as a preliminary step, significantly improves BirdNET detections by increasing True Positives by approximately 281.97%, and reducing False Negatives by about 62.03%, thus demonstrating a novel and effective approach for bird species identification. Our findings underscore the importance of adapting general-purpose tools to address specific challenges in biodiversity monitoring. The experimental results show that fine-tuning Deep Learning models that account for the unique characteristics of specific ecological contexts can substantially enhance the accuracy and efficiency of PAM's efforts.

Keywords: Computer Vision, Convolutional Neural Networks, Deep Learning, Ecoacoustics, Passive Acoustic Monitoring

1. Introduction

Natural environments face significant challenges in terms of conservation and monitoring due to habitat loss, the effects of climate change, and anthropogenic pressure. In response to this crisis, biodiversity monitoring and species-interaction assessments have become essential to understanding environmental impacts and developing conservation strategies. Effective biodiversity monitoring is fundamental for conservation efforts, as it provides the data necessary to make informed decisions, but it is challenging to get the data necessary to make those informed decisions.

In this regard, identifying and tracking bird species are crucial, as birds serve as indicators of ecosystem health (1). Although various automatic monitoring technologies, such as cameras and audio recorders, are already in use, efficiently managing and analyzing the large volumes of data generated by these devices remains challenging. One effective technology is Passive Acoustic Monitoring (PAM), which uses audio recorders to continuously capture sounds in an environment. PAM is particularly valuable for monitoring biodiversity as it can operate in remote and inaccessible areas, providing continuous data without disturbing the habitat (2). By leveraging PAM, the monitoring scale can be expanded significantly, allowing for more comprehensive and detailed ecological studies (3).

In recent years, the cost of recording devices has reduced, leading to an increase in the collection of this type of data in the field of ecology (4; 5; 6). However, in many cases, these data are not adequately labelled or classified, making analysis difficult and limiting their utility for decision-making (7). The primary objective is to address the challenge of efficiently managing large data volumes for extracting relevant information for biodiversity conservation by applying Machine Learning techniques. This approach aims to automate data labelling process and improve data analysis efficiency in the context of environmental monitoring. By doing so, human resources can be freed for more complex tasks, allowing for a better understanding of ecological interactions.

The remainder of the paper is organized as follows. First a Section of Related Works in 2 is presented. After presenting the dataset creation in Section 3, the methodology, including the integration of YOLOv8 and BirdNET for bird vocalization detection and species classification and detailing the development and implementation of the Bird Song Detector and the multi-stage pipeline approach, are presented in Section 4. In Section 5, we present the experimental results obtained from applying the proposed methodology in the context of Doñana National Park. Section 6 analyzes and discusses the implications of the results, addressing the challenges and opportunities identified. Finally, Section 7 provides the conclusions drawn from this study and outlines future research directions.

2. Related Works

Historically, early bird vocalization recognition methods relied on basic sound feature analysis, using techniques such as Random Forests for classification. These methods focused on extracting specific audio features, such as frequency, pitch, and duration, to create a feature set that could be used for classification (8). While effective to a certain extent, these approaches were limited by their reliance on manually crafted features and often struggled with complex and overlapping sounds.

Recent advancements in Deep Learning techniques for bird vocalization recognition have emerged as valuable tools, enabling precise and continuous monitoring of bird populations (9). One of the most popular models of bird vocalization recognition is BirdNET (10), which has proved to be successful in many cases (11; 12; 13). The testing of these models has been conducted in environments for which the base model was especially well trained, mainly from Northen America and central-northern Europe. This means that the model was initially trained on a dataset that closely resembles the conditions of the test environment, thereby increasing its predictive accuracy. Nevertheless, this also implies that models trained on global datasets performance may degrade when applied to unfamiliar environments or conditions due to variability in local bird vocalizations and unique environmental conditions (14; 15). BirdNET's performance is suboptimal in real-world contexts, including overlapping bird songs and different acoustic backgrounds (11) as False Positives (FPs) often arise from other vocalizing animals, anthropogenic sounds or weather conditions (16; 10; 17).

Indeed, foundational models like BirdNET, which are trained on global datasets, often struggle to recognize species for which they were not trained on. In theory, it is possible to retrain an existing model in order to add missing species, what is known as fine-tuning (18). Alternatives such as Perch (19) and custom models based on BirdNET with fine-tuning for local conditions have been developed. These fine-tuned models aim to improve accuracy by accounting for the unique characteristics of local bird populations (15). However, this task is quite challenging as it requires Machine Learning expertise similar to having to train models from scratch. As a result, the successful implementation of Machine Learning tools to accelerate the annotation process in ecoacoustics has been achieved by only a small number of organizations (20).

To address these limitations, we propose a multi-stage pipeline approach that combines a generalizable Bird Song Detector based on YOLOv8 (You Only Look Once v8) (21) with a project-specific classifier (Figure 1) inspired by methodologies used in camera trap projects (14; 22). The Bird Song Detector has been trained to detect and temporally locate bird vocalizations, even from species not encountered during training.

As a study case, we apply this Bird Song Detector to the PAM program developed by the BIRDeep project at Doñana National Park (SW Spain) (23), in order to facilitate the automatization of bird species identification in the local soundscapes of Doñana. By using a Bird Song Detector as a preliminary step, we can ensure the presence of bird songs in the audio segments.

The added value of this pipeline lies in its ability to isolate relevant segments of audio that contain bird vocalizations before applying a more computationally intensive species classifier. This process not only reduces the number of non-bird segments incorrectly identified as bird vocalizations but also simplifies the fine-tuning of the species classifier, as the model is optimized to work with segments that are already confirmed to contain bird sounds. This is more than just a simple fine-tuning of an existing model; by separating the detection and classification stages, we minimize background noise interference and focus the classifier's resources on the most relevant audio data, leading to improved overall system performance. The generalization ability of the Bird Song Detector also ensures that it can identify bird vocalizations even from species not present in the training set, making this approach robust for real-world applications with diverse species compositions.

3. Dataset

In this section, we provide a detailed description of the dataset employed in our study. We begin by discussing the field site and acoustic data collection process, followed by the data preparation steps necessary for transforming the raw audio into a format suitable for analysis. Finally, we describe the data distribution strategy.

3.1. Field study site and PAM design

Soundscapes were registered at Doñana National Park (SW Spain). This area corresponds to the marshes of the Guadalquivir delta and it is one of the most important wetlands of Southern Europe, where millions of migrating birds stopover and winter every year (24; 25). Doñana has three main habitats, which are differentiated by their flooding regime and vegetation: scrublands, marshland and the ecotone or transition among them. The deployment design of the BIRDeep project included nine AudioMoth recorders (26) that were distributed among these three habitats: two in the marshland, three in the ecotone and four in the scrubland, differentiating high and low

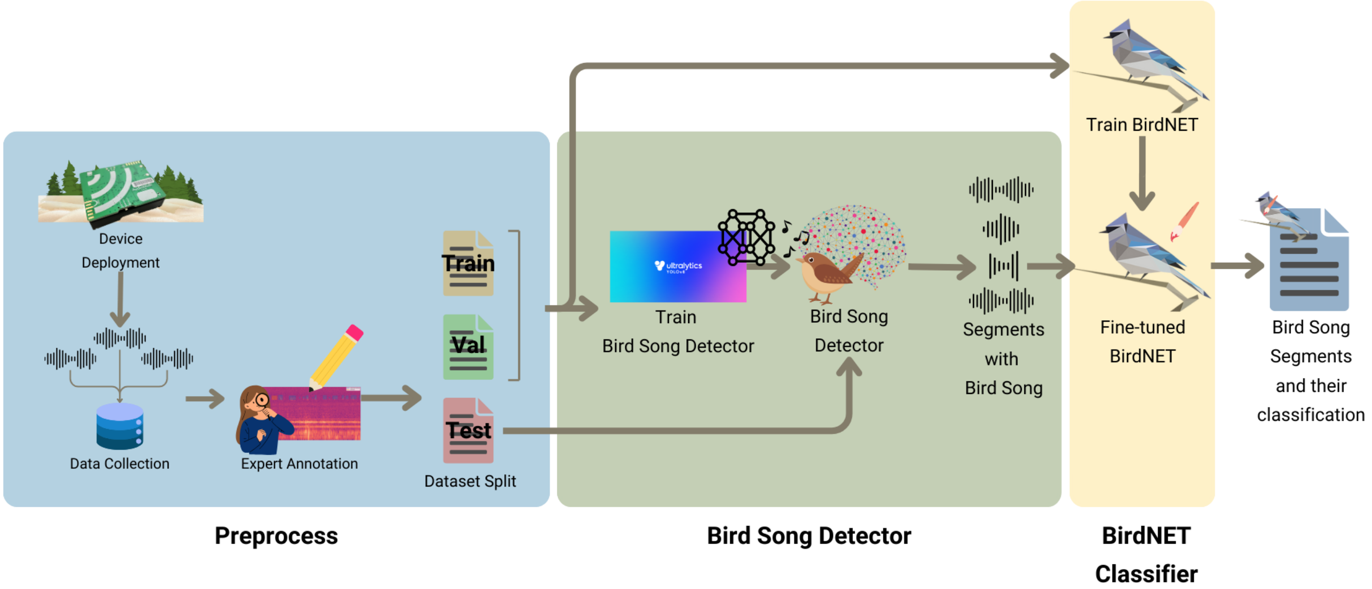


Figure 1: Pipeline for the development of our Bird Song Detector. The process was divided into three main stages: (1) Preprocess: This stage involved deploying recording devices in natural environments to collect audio data. The collected data was then annotated by experts to identify bird vocalizations, followed by splitting the dataset into training, validation, and test sets. (2) Bird Song Detector: In this stage, a Bird Song Detector was trained using the annotated dataset. This detector identified segments containing bird vocalizations from the audio recordings. The trained model was then used to detect bird songs in the audio data, producing segments that contained potential bird vocalizations. (3) BirdNET Classifier: The final stage involved fine-tuning the BirdNET model using the original annotated segments. The model was then validated and tested with the segments identified by the Bird Song Detector. This fine-tuned BirdNET model accurately classified the bird species present in each segment.

scrubland (see Figure 2). AudioMoths are low-cost automatic audio recording devices with open-source hardware. They continuously recorded 1 minute of audio every 10 minutes. Configuration parameters of deployed AudioMoth included a sampling rate of 32 kHz, a medium gain, and a filter band focused on bird frequencies (0.6-16.0 kHz).

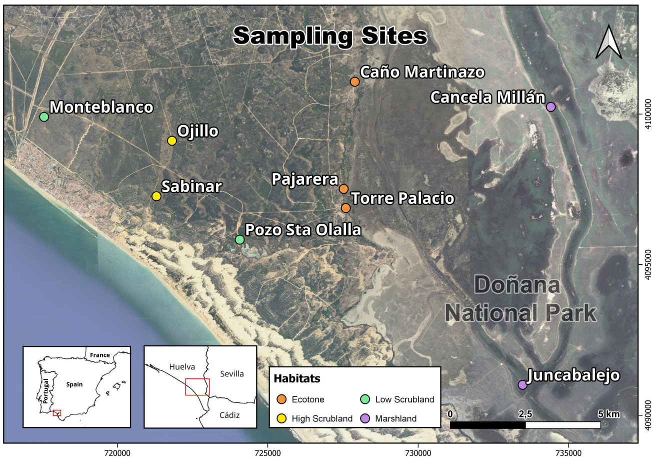


Figure 2: Distribution of the nine passive acoustic monitoring devices (AudioMoths) deployed in Doñana National Park across its three main habitats: marshland, scrubland (high or low) and ecotone.

3.2. Acoustic data annotation

A total of 461 minutes of audio data were annotated by two of the co-authors of this work with ornithological expertise. They used Audacity software to listen to the recordings and annotate the spectrograms (27). When faced with uncertainties, they referred to field censuses to ensure the accuracy of their annotations.

Field census provided a valuable reference by listing the species observed in the area during the recording period. These field census are conducted periodically due to the associated costs and provide real data regarding diversity, which can be used to cross-check the audio labels. This information reduced the ambiguity in the spectrograms and audio, allowing for more precise annotations. By cross-referencing the audio data with field observations, the annotators could confirm species presence and improve the reliability of their annotations at the same time that it helped to narrow down the number of potential species, making it easier to identify the species present in the recordings.

Annotation efforts prioritized periods of high bird activity, mainly at morning chorus, to maximize the number of bird vocalizations (28). Annotations consisted of bounding boxes with the minimum and maximum frequency and start time and end time of each bird vocalization in each spectrogram. In total, and after a standardization process of the annotations, there were 3749 annotations belonging to the 38 different classes shown in Figure 3. In addition to the species-specific classes we have distinguished other general classes: *Genus* (when the species was unknown but the genus of the species was distinguished), a general *Bird* class, and a *No Audio* class for recordings that contain only soundscape without bird songs. As the Bird Song Detector only has two classes, we reclassified labels as *Bird* or *No bird* for recordings that include only soundscape background without biotic sound or whether biotic sounds were non-avian.

It is important to note that the dataset (29) exhibits class imbalance, with varying frequencies of annotations across different bird species classes. Additionally, the dataset contains inherent challenges related to environmental noise, which will be addressed later when discussing the dataset construction in conjunction with complementary datasets.

3.3. Data Preparation

Audio data was transformed into Mel spectrograms for training a Deep Learning model based on image processing techniques, i.e. Convolutional Neural Networks (30). A Mel spectrogram is a variant of spectrogram where the frequency axis is transformed to a Mel scale, which mimics human auditory perception more closely than the linear scale. This graphical representation of audio data displays how the signal's energy is distributed across different frequencies over time, making it suitable for image processing techniques (31).

Figure 4 shows an example of a Mel spectrogram from the Doñana dataset (29), including annotations for temporal windows and the complete frequency spectrum. Although annotations for specific frequency windows were available, the study was simplified by considering the full frequency spectrum due to the limited size of the dataset.

3.4. Data Distribution

The dataset used in this study (29) was divided into training, validation, and test sets, aiming for an 80-10-10 proportion per species (32). However, maintaining independence and avoiding correlation among subsets to prevent overestimation during model evaluation (33) proved challenging as some audios were multilabeled, containing vocalizations from more than one species. This made it difficult to strictly adhere to the desired 80-10-10 ratio. To mitigate these issues, we prioritized ensuring that no audio file appeared in more than one subset, even if it contained multiple species, to maintain the independence of the sets. The final distribution, as depicted in Figure 5, reflects these adjustments, balancing the dataset as much as possible while considering these constraints.

The dataset is available at a Hugging Face repository (29) with this split and the best data augmentations achieved during experimentation as shown in Section 5.2.

4. Methods

In this section we delve into the model development, highlighting the methodologies and techniques used to build and fine-tune our models. At the end we outline our approach for model evaluation to ensure robust performance assessment.

4.1. Bird-Song Detector Model Development

To detect bird vocalizations within audio recordings, we developed a Bird Song Detector using YOLOv8 (21). YOLOv8 is a state-of-the-art real-time object detection model that balances high accuracy and speed, making it ideal for processing large datasets in ecological studies. Unlike its predecessors, YOLOv8 introduces several enhancements, including improved feature extraction through convolutional layers and optimized bounding box regression, which enhance its ability to detect fine-grained details. The model architecture consists of multiple convolutional layers followed by fully connected layers that predict bounding boxes, objectness scores, and class probabilities for detected objects (21).

Given the nature of our audio data, which contains a mix of bird sounds and background noise, YOLOv8's precise detection capabilities are essential for distinguishing relevant events. The model divides each input image into a grid and predicts bounding boxes that indicate the presence of these events, along with a confidence score for each prediction. In our case study, the input images are mel spectrograms, and the relevant events are bird vocalizations represented as sonograms in the images. This approach allows for efficient identification of bird vocalizations even in complex acoustic environments.

We chose YOLOv8 because it has been in development for a longer period, providing proven stability and reliability. At the start of our experiments, YOLOv9 and YOLOv10 were either not yet available or were very recent releases, lacking the extensive validation that YOLOv8 has undergone. Therefore, we opted for YOLOv8, which is widely used in many applications due to its robust performance, offering a good balance of computational efficiency and accuracy. Initial experiments demonstrated that the small-sized YOLOv8 model (`yolov8s.pt`) was particularly effective in maintaining high detection accuracy while minimizing computational load (34).

To further optimize YOLOv8 for our study, we fine-tuned the model using annotated mel spectrograms of bird vocalizations. This fine-tuning involved incorporating data augmentation techniques, such as adding noise and shifting frequencies, to enhance the model's robustness to variations in the dataset (29), as well as adding additional samples from an external dataset (35).

These improvements demonstrate YOLOv8's effectiveness in accurately detecting bird vocalizations within large, diverse datasets, providing a strong foundation for subsequent species classification using fine-tuned BirdNET models.

4.2. Bird-Song Classifier Model Development

Following this, BirdNET V2.4 was fine-tuned to create a classifier adapted to the ecological context of Doñana. BirdNET is a Deep Learning model specifically designed to classify bird species using audio inputs. It segments audio recordings into 3-second clips, transforms the audio into Mel spectrogram images, and performs classification using a deep Convolutional Neural Network (36).

The fine-tuning process involved appending additional training data specific to Doñana's bird species to the BirdNET model. This augmentation aimed to address the bias and enhance the model's performance in the region. This approach

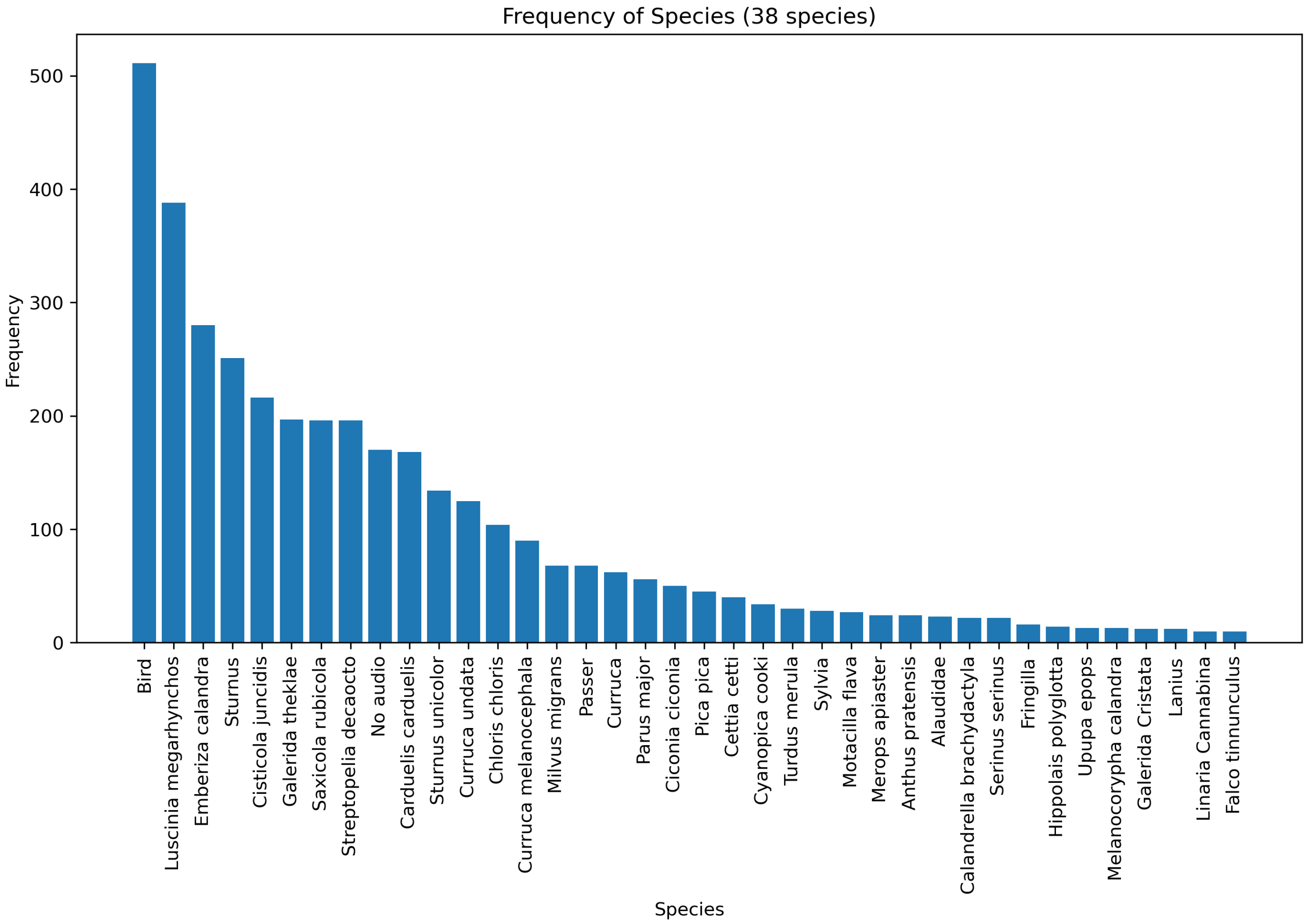


Figure 3: Distribution of the 38 annotated classes in the dataset.

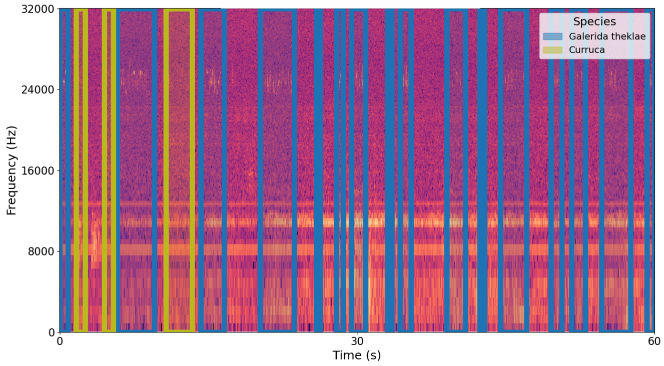


Figure 4: An example of a Mel spectrogram from the Doñana dataset with annotations (i.e. blue and lime green rectangles) for temporal windows and the complete frequency spectrum for the annotated bird vocalizations.

helps the model generalize better to variations in species vocalizations and acoustic environments present in Doñana, thereby improving its accuracy in identifying local bird species (37; 38).

BirdNET V2.4 covers frequencies from 0 Hz to 15 kHz. The model supports a global species selection, encompassing 6,522 classes (including 10 non-event classes), making it suitable for the identification of diverse bird species all over the world. Non-event classes refer to categories that represent sounds or signals that are not related to bird vocalizations, such as background noise, human-made sounds, or other environmental noises.

Technical specifications of BirdNET V2.4 include (36):

- 48 kHz sampling rate, with automatic upsampling and downsampling capabilities to handle artifacts from lower sampling rates.
- Two Mel spectrograms are computed as input for the Convolutional Neural Network:
  - First spectrogram: fmin = 0 Hz, fmax = 3000 Hz, nfft = 2048, hop size = 278, 96 mel bins.
  - Second spectrogram: fmin = 500 Hz, fmax = 15 kHz, nfft = 1024, hop size = 280, 96 mel bins.

Both spectrograms are resized to a final resolution of 96 × 511 pixels after raw audio normalization between -1 and 1, incorporating non-linear magnitude scaling (39).

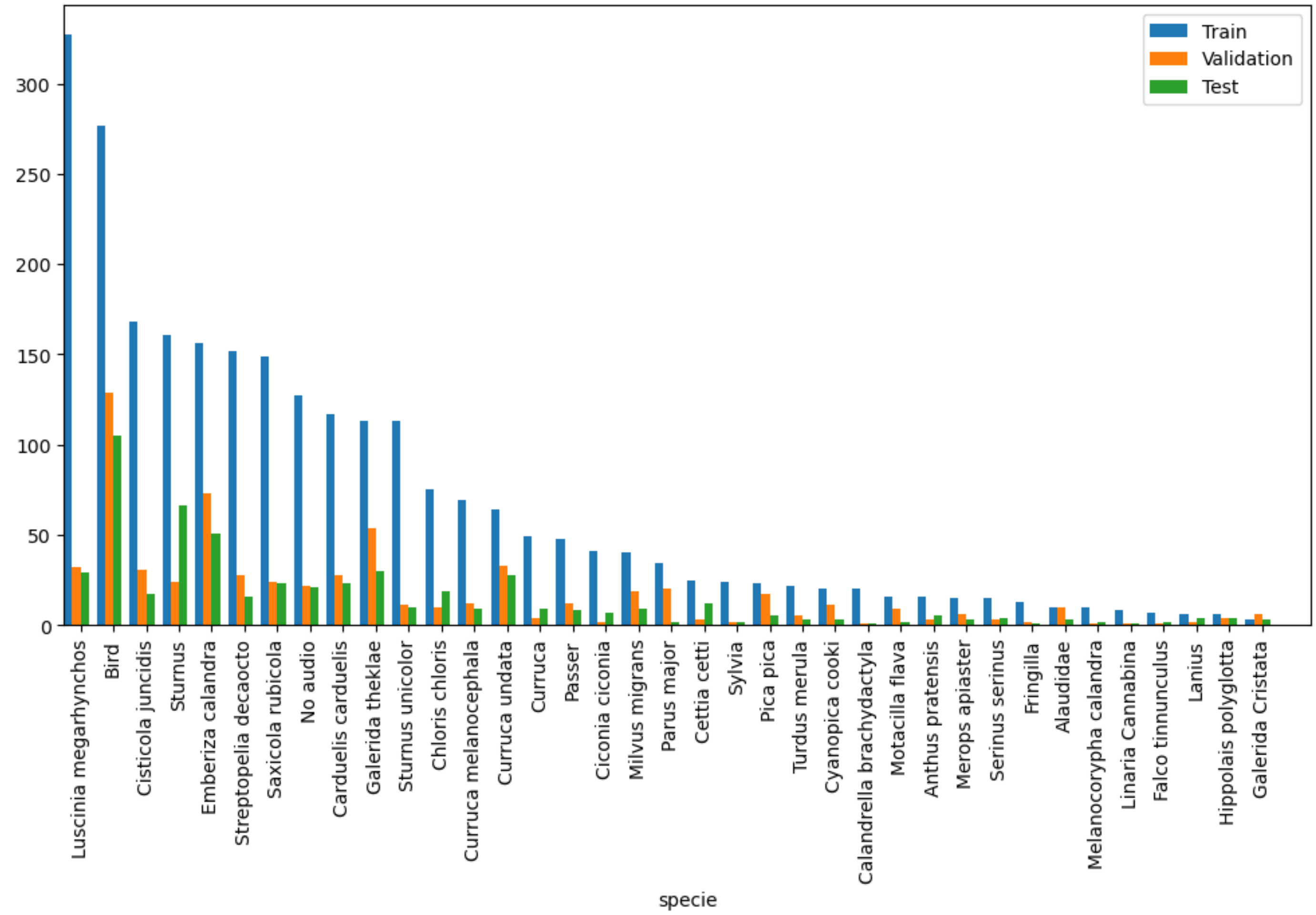


Figure 5: Distribution of the dataset per species across training, validation, and test sets.

- BirdNET V2.4 uses an EfficientNetB0-like backbone with a final embedding size of 1024 for feature extraction and classification.

Despite its capabilities, BirdNET's original training datasets from public repositories like Xeno-canto (40) and Macaulay Library (41) may exhibit biases and limitations for detailed analyses of some local assemblages. These repositories predominantly contain recordings of focal bird songs, which isolate the bird's vocalization from its acoustic environment. This approach results in cleaner audio data that lacks overlapping sounds from multiple species and natural ambient backgrounds. Moreover, these datasets are spatially and temporally biased, often collected from regions where the libraries are heavily used, such as North America. Consequently, species common in these regions are overrepresented compared to those in other geographical locations with different ecological dynamics. These factors can lead to challenges when deploying models trained on such datasets in diverse ecological contexts at regional or local scales.

4.3. Predictions

The predictions of the Bird Song Detector were obtained in the form of bounding boxes, which are given as the image coordinates of the center of the bounding box (x_{center} and y_{center}), the *width*, and the *height* of the bounding box in the YOLOv8 output format. These coordinates are then transformed into temporal coordinates, using the x_{center} value and *width* to calculate the start time and end time of the predicted bird song segment. This transformation is performed using the following equations:

$$x_{center\_d} = x_{center} \times W \tag{1}$$

$$w\_d = w \times W \tag{2}$$

$$start\_sec = \left(x_{center\_d} - \frac{w\_d}{2} \right) \times \frac{60}{W} \tag{3}$$

$$end\_sec = \left(x_{center\_d} + \frac{w\_d}{2} \right) \times \frac{60}{W} \tag{4}$$

where W is the width of the spectrogram image used as input for the Bird Song Detector model. The variables x_{center} and *w* represent normalized coordinates and width of the bounding box, respectively. To interpret these predictions in real-world temporal coordinates, they are deserialized using x_{center_d} and w_d, which are the denormalized counterparts of x_{center} and w, scaled by W. And, *start_sec* and *end_sec* denote the starting and ending seconds of the bounding box segment.

5. Experiments and Results

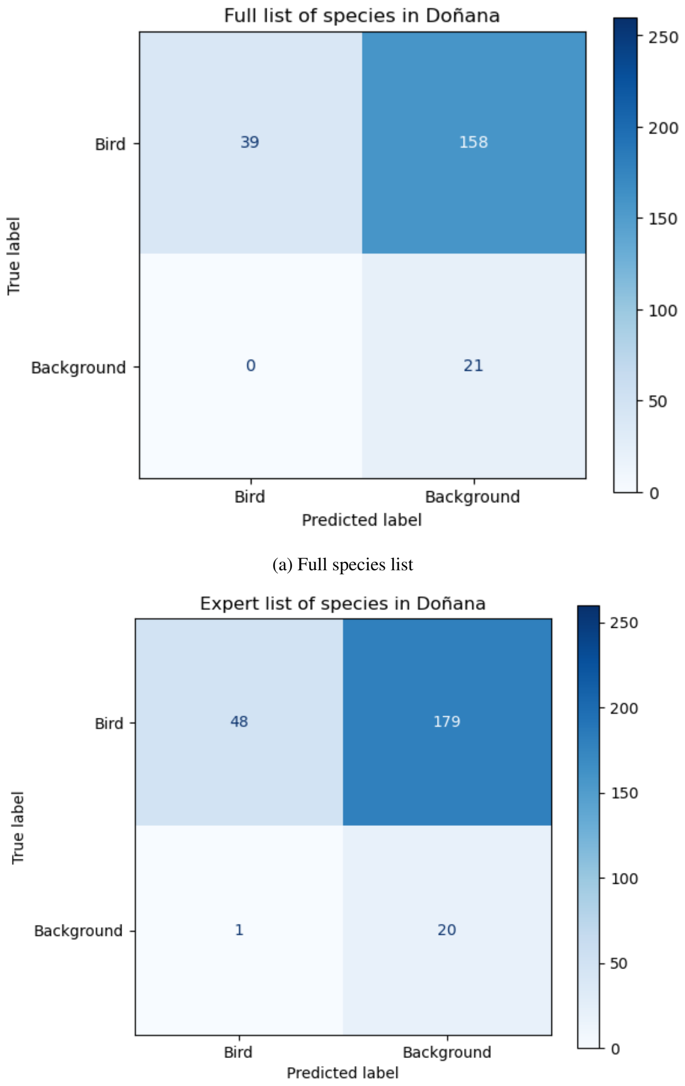
In this section, we delve into the practical evaluations and training processes involved in our study. We first present the evaluation of BirdNET as a bird vocalization detector. We then discuss further evaluations conducted after fine-tuning BirdNET with data from Doñana. Finally, we explore the training process of the Bird Song Detector, highlighting the impact of incorporating background noise and the adjustments made to improve the model's performance. We also present the outcomes of our study on the development and fine-tuning of a Bird Song Detector and classifier tailored for the Doñana ecological context. We begin by examining the performance of the Bird Song Detector. Next, we evaluate the classification of bird species using the fine-tuned BirdNET model.

To evaluate the performance of BirdNET as a Bird Song Detector without species classification, confusion matrices were generated. First, evaluation was done with a full list of species from Doñana that included 412 classes, using a minimum Intersection over Union (IoU) of 0.2 and a confidence score of 0.6 (Figure 6a). Second, a refined evaluation was done using a shorter list of 337 species extracted from expert annotations. This list was created by reviewing all species recorded in Doñana during historical censuses, incorporating references from existing literature (42; 43), even if a species had only been observed once. However, to focus on more reliable detections, only the most common species were retained, along with those that, while not frequent, are known to potentially appear in the region (Figure 6b). Both evaluations used the same IoU and confidence score parameters. IoU measures the overlap between predicted bounding boxes and ground truth annotations, ensuring that the predictions are spatially accurate and aligned with actual bird vocalizations.

5.1. *Evaluation of BirdNET after fine-tuning*

Further evaluations were conducted after fine-tuning BirdNET with data from Doñana. Initially, the confidence score was set at 0.6 (11; 12). This score is derived from the output of the final layer of BirdNET, where logits are processed through a sigmoid activation function with a specified sensitivity (in this case, sensitivity = 1). The resulting values are converted into confidence scores, ranging from 0 to 1, indicating the probability that a prediction is correct. Figure 7a shows the results obtained at this confidence score threshold. However, this setting did not show a significant improvement.

Subsequently, the confidence score was lowered to 0.1 to explore its impact on performance, as depicted in Figure 7b. Despite the improvements in TP observed at the 0.1 confidence score threshold, it is notable that the number of FPs and False Negatives (FNs) increased considerably.



(a) Full species list

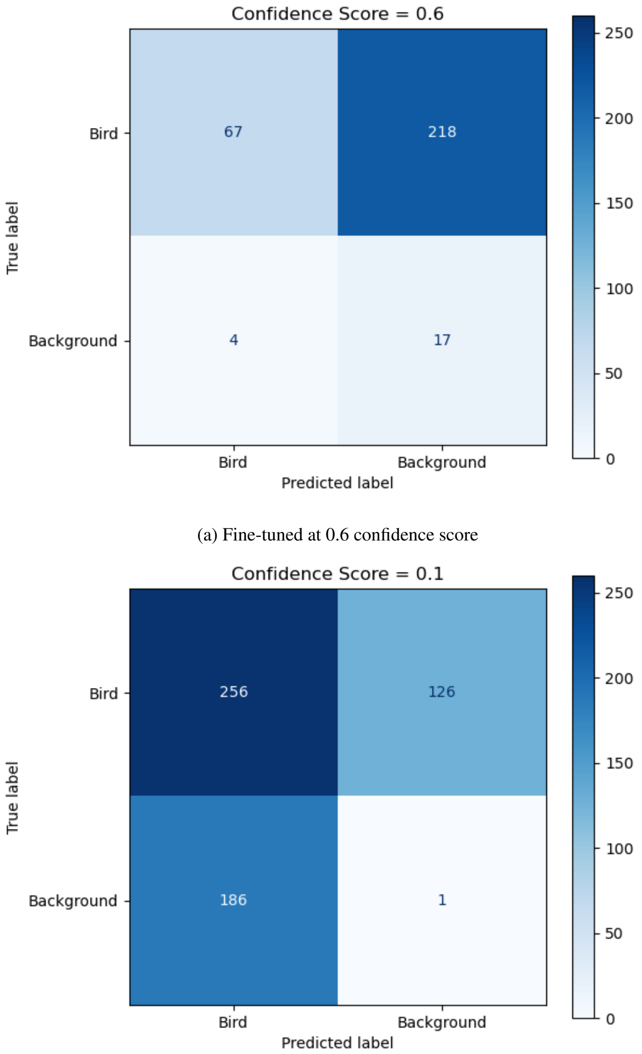
(b) Expert species list

Figure 6: Comparison of confusion matrices for BirdNET without fine-tuning using different species lists: (a) BirdNET evaluated with the full list of species from Doñana, and (b) BirdNET evaluated with the species list from expert annotations. In both cases, BirdNET is evaluated as a detector, i.e., without considering whether the species is correctly identified, only if it corresponds to an actual annotated bird song.

5.2. *Bird Song Detector Training*

To select the Bird Song Detector, multiple models were trained and their performance was evaluated using the mean Average Precision (mAP) metric at an Intersection over Union (IoU) threshold of 50% (mAP50; (44)). Initial experiments, represented by the purple lines in Figure 8 (*Base*, *Hyperparameter Exploration V1*, *Hyperparameter Exploration V2*, *AugmentedBG V1* and *AugmentedBG V2*), showed suboptimal performance, particularly due to a high number of FPs.

Given the complexity and small size of the dataset (29), the



(a) Fine-tuned at 0.6 confidence score

(b) Fine-tuned at 0.1 confidence score

Figure 7: Comparison of confusion matrices for BirdNET fine-tuned with Doñana data at different confidence scores: (a) BirdNET fine-tuned at a confidence score of 0.6, showing not much improvement compared to the base model, and (b) BirdNET fine-tuned at a confidence score of 0.1, where the detection rate improves significantly, but the number of FPs also increases.

bounding boxes, which were initially designed to delimit both the frequency spectrum and the time window, were simplified. This simplification involved taking the full frequency spectrum into account and only delimiting the temporal window (represented by the yellow line *FullFrequencies* in Figure 8). This approach aimed to reduce the complexity of the task for the model, given the limited amount of data available.

To address the issue of the FPs, the ESC-50 dataset (45), which is a large collection of 50 environmental sound classes, was introduced as background noise (negative samples). Bird-related classes were removed from this dataset to prevent confusion. However, when this dataset was fully included, the model primarily learned to recognize background sounds and failed to detect bird songs effectively (green line (*AllESC50*) in Figure 8).

Subsequently, the ESC-50 dataset was reduced to comprise only 25% of the total training data. This adjustment led to significant improvements in the model's performance (orange line (*Best Model*) in Figure 8). This balanced approach allowed the model to better differentiate between bird songs and background noises, improving detection accuracy while minimizing FPs.

The various model configurations employed during the experimentation are summarized in Table 1, which also presents the performance metrics for each configuration. These configurations include different background augmentation techniques and utilizations of the ESC50 dataset. The *Best Model*, which employed synthetic background augmentation of noise and intensity changes and a reduced ESC50 dataset, achieved one of the highest mAP50 scores of 0.29, along with balanced precision and recall. Other configurations, such as *AllESC50*, displayed lower performance metrics. On the other hand, the *Full Frequencies* model without the ESC50 dataset had the best performance during training with a mAP50 score of 0.305. This highlights the importance of specific augmentation strategies and dataset choices in optimizing detection accuracy.

5.3. *Bird Song Detector Evaluation*

To evaluate the performance of the Bird Song Detector,the confidence scores of the detector were converted to logit scores to eliminate linearity introduced by the activation function (12).

The conversion from confidence scores to logit scores is based on the logistic function:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \tag{5}$$

where p represents the confidence score of the detector's prediction. This transformation helps to interpret the confidence scores probabilistically, providing a more nuanced understanding of the detector's performance characteristics.

We evaluated the Bird Song Detector using different probability thresholds (40%, 60%, 80% and 95%) to find the optimal balance between maximizing True Positive (TP) and minimizing False Negatives. This optimization process involved analyzing the trade-off between increasing TP detection (i.e., capturing more bird vocalizations) and the rise in FP or FN errors, as higher thresholds tend to reduce the number of detections, but with fewer FPs. After evaluating these trade-offs, we determined that the 60% threshold provided the best balance: it minimized the loss of TPs while keeping the FN rate at an acceptable level. This threshold was selected as it offers a reasonable compromise between the model's confidence in detecting bird vocalizations and its practical ability to avoid missing significant events. Further, this threshold aligns with the performance metrics relevant to real-time applications, where both detection accuracy and speed are critical factors (see conclusions, Section 5.4).

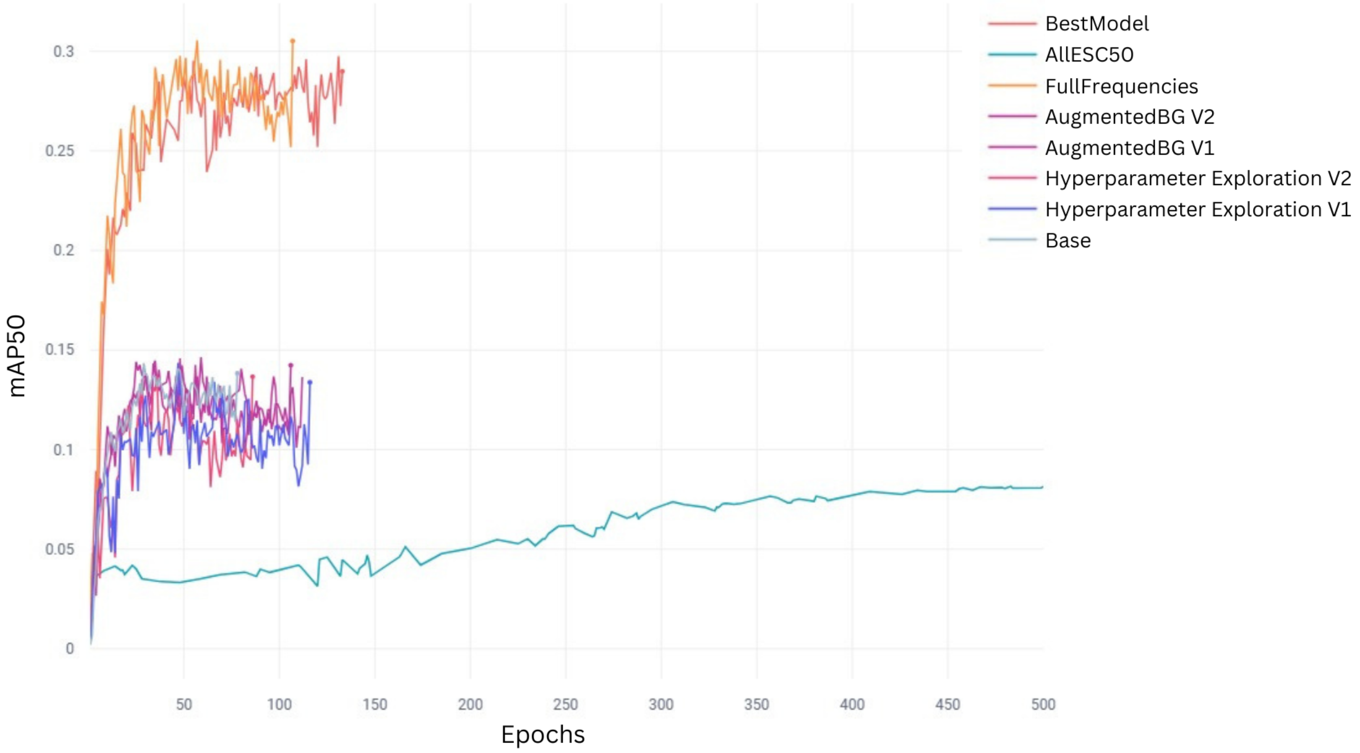


Figure 8: Mean Average Precision at 50% Intersection over Union (mAP50) during training for different experiments. Initial models (purple lines) showed high False Positive (FP) rates. The inclusion of ESC-50 data (orange lines) improved model performance significantly.

Name	Background Augmentation		Frequency Spectrum	mAP50	Precision	Recall
	Synthetic Background Augmentation	ESC50 Dataset				
BestModel	Add Noise + Intensity Change	Reduced	Full Spectrum	0.29	0.412	0.308
AllESC50	Add Noise + Intensity Change	Full	Full Spectrum	0.082	0.142	0.201
FullFrequencies	Add Noise + Intensity Change	-	Full Spectrum	0.305	0.399	0.302
AugmentedBG V2	Add Noise + Intensity Change	-	Range Bounded	0.142	0.291	0.163
AugmentedBG V1	Add Noise	-	Range Bounded	0.136	0.232	0.17
Hyperparameter Exploration V2	-	-	Range Bounded	0.137	0.272	0.172
Hyperparameter Exploration V1	-	-	Range Bounded	0.134	0.258	0.172
Base	-	-	Range Bounded	0.138	0.275	0.174

Table 1: Experimental results and configurations of the Bird Song Detector. The ‘-’ symbol indicates that no synthetic augmentation or dataset was applied in that experiment.

5.4. Selection of confidence score threshold for the Bird Song Detector

Table 2 summarizes the different values obtained for various detection probabilities for the *Best Model* chosen after experimentation (see Section 5) in our data and case study, including their respective logit scores, confidence scores, and TP losses.

As illustrated in Figure 9, the 60% probability threshold corresponds to a logit score of -1.77 and a confidence score of 0.15, with a TP loss of 22.08%. This threshold is strategically chosen to balance detection accuracy with the practical limitations of the model. In Figure 9 each black dot (with some transparency so overlapping can be seen clearer) represents a detection made by the Bird Song Detector, with the confidence score transformed into a logit score on the X-axis. If a dot has a Y-axis value of 0, it means the detection was incorrect (i.e., it did not correspond to a bird vocalization). Conversely, a Y-axis value of 1 indicates that the detection was correct, based on the ground truth annotations provided by experts in our dataset. The blue line represents the logistic regression model fitted to these data points. This line allows us to estimate the probability of a correct prediction for any given logit score, which can then be transformed back into a confidence score. The orange lines in the figure highlight the intersection of this 60% threshold with the blue logistic regression line, showing the corresponding logit score, which is approximately -1.79. When transformed back into a confidence score, this results in a threshold of 0.14.

Lower thresholds, such as 40%, result in a 0% loss of TPs but occur before the logistic regression model’s effects begin to improve performance significantly (logit = -2.75, confidence = 0.06; Table 2). At this threshold, the model fails to utilize the logistic regression adjustments effectively, as evidenced by the

Probability Threshold	Logit Score	Confidence Score	TP Loss (%)
40%	-2.75	0.06	0.00
60%	-1.77	0.15	22.08
80%	-0.58	0.36	74.35
95%	1.30	0.79	99.03

Table 2: Comparison of different probability thresholds for detection with their respective logit scores, confidence scores, and TP losses.

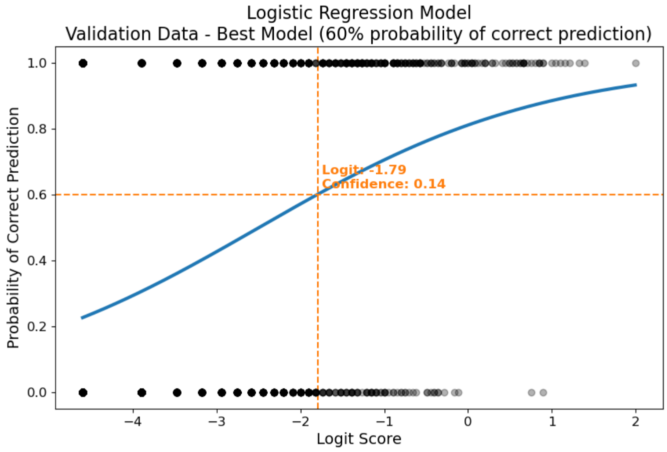


Figure 9: Logistic Regression Model with a 60% probability threshold for correct prediction.

extremely low confidence score.

On the other hand, higher thresholds, like 80% and 95%, lead to excessive losses of TPs (74.35% and 99.03%, respectively). These thresholds result in an impractical trade-off, where the reduction in FPs comes at the cost of almost complete loss of the TPs (which need to have the high probability threshold to be retained), significantly degrading the detector's performance.

In this case and for our case study the 60% threshold is high enough to ensure that the logistic regression model's adjustments are actively enhancing detection performance, while also preserving a manageable amount of TPs. This threshold ensures the model remains both reliable and practical for detecting bird songs, thereby offering an optimal balance between confidence and detection accuracy.

5.5. Bird Vocalization Detection Comparison

To further illustrate the improvements, we compared the averaged results of BirdNET with the same confidence score threshold of 0.6 (Figures 6a, 6b and 7a) to our Bird Song Detector (Figure 11) which will be selected and tested in the next section of Results (see Section 5).

Next, we compared the fine-tuned BirdNET results with a low confidence score threshold of 0.1 (Figure 7b) to our Bird Song Detector (Figure 11).

The percentage improvements in TPs, FNs, FPs were calculated based on changes observed between both methods and were calculated using the following general formula:

$$\text{Percentage Change} = \left(\frac{\text{New Value} - \text{Old Value}}{\text{Old Value}} \right) \times 100 \quad (6)$$

Where *New Value* refers to the value obtained from our Bird Song Detector and *Old Value* refers to the value obtained from BirdNET with the specific confidence score threshold.

An increase in TPs indicates an improvement in detection accuracy, as more bird vocalizations are correctly identified. A decrease in FNs is also a sign of improved performance, as fewer bird vocalizations are missed by the system. Conversely, a decrease in FPs represents a reduction in erroneous detections of non-bird sounds. For all metrics, positive percentage changes in TPs and negative changes in FN and FP signify improvements, while negative changes in TPs or positive changes in FNs and FPs indicate a decline in performance.

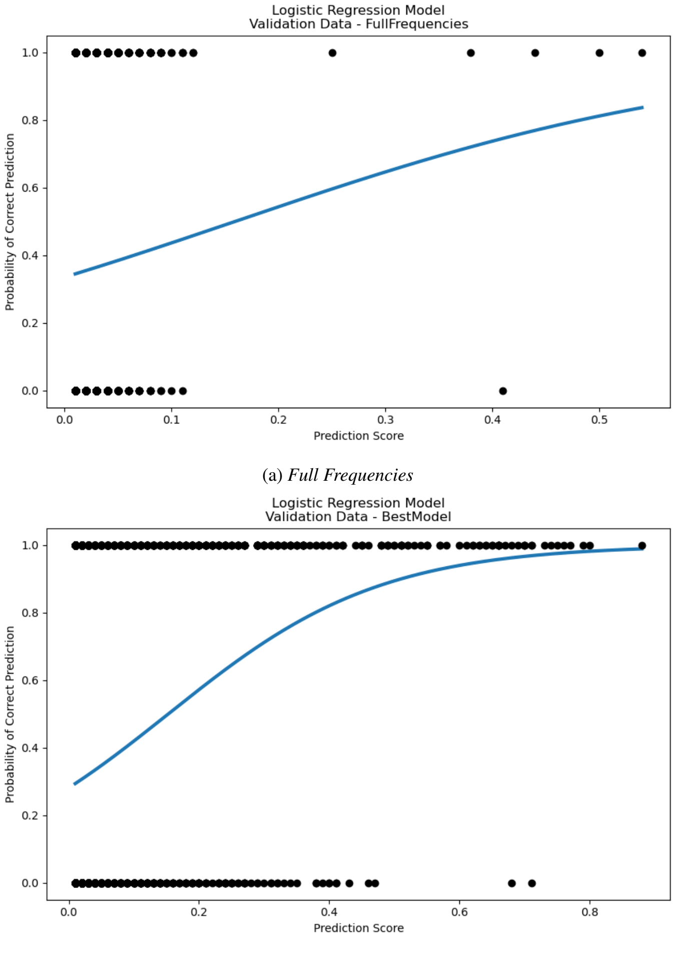
When evaluating percentage changes in the results, it is important to consider them in absolute terms. A large percentage increase or decrease might appear significant at first glance, but its actual impact depends on the scale of the values involved. For instance, changes from small baseline values can result in high percentage variations, even if the absolute difference is relatively minor. Conversely, changes in metrics with larger absolute values might show smaller percentage shifts but represent a more substantial impact on overall performance. Therefore, it is crucial to interpret these percentages within the context of the absolute figures to avoid misinterpreting the true extent of the changes.

5.6. Bird Song Detector Election

To further investigate the performance of the top models, logistic regression curves were plotted for *Best Model* and *Full Frequencies* for Validation data and a confidence score threshold of 0.15 was chosen.

For *Full Frequencies* (Figure 10a), the probability of correct prediction increases gradually with the prediction score. The curve appears more linear and less steep, suggesting that the model's predictions are less confident but more consistent across the range of scores. This linearity indicates that *Full Frequencies* provides a moderate level of certainty in its predictions, reflecting a balanced but somewhat cautious approach to identifying bird songs. The gradual increase in probability demonstrates a steady improvement in prediction accuracy as the logit score rises.

In contrast, the *Best Model* (Figure 10b) exhibits a logistic regression curve where the probability of correct prediction increases more rapidly with the prediction score, indicating a steeper and more pronounced curve. This steepness suggests that *Best Model* makes more confident predictions, with higher scores correlating strongly with correct predictions. The rapid ascent of the curve means that as the logit score increases, the model quickly becomes more certain about its predictions. This behavior implies that *Best Model* is better at distinguishing between correct and incorrect predictions, offering higher confidence in its detections.



(a) *Full Frequencies*

(b) *Best Model*

Figure 10: Logistic Regression Curves for the two best models (a) *Full Frequencies* and (b) *Best Model*.

Given the steeper slope of the logistic regression curve for *Best Model*, it can be inferred that this model is more effective at identifying bird songs accurately. Despite having slightly fewer total predictions compared to *Full Frequencies*, *Best Model* demonstrates superior confidence and accuracy in its predictions. This enhanced performance can be attributed to its specific augmentation strategy, which involves synthetic background augmentation combined with noise and intensity changes on a reduced ESC50 dataset. These techniques likely contribute to the model's ability to make more decisive predictions.

5.7. Bird Song Detector Performance

Using the confidence threshold of 0.15 obtained with the logistic regression, we executed the Bird Song Detector on the test dataset. The resulting binary confusion matrix for these detections is shown in Figure 11. This matrix shows the effectiveness of the detector in identifying temporal windows with bird songs and shows the improvement in detecting bird songs compared to previous confusion matrices.

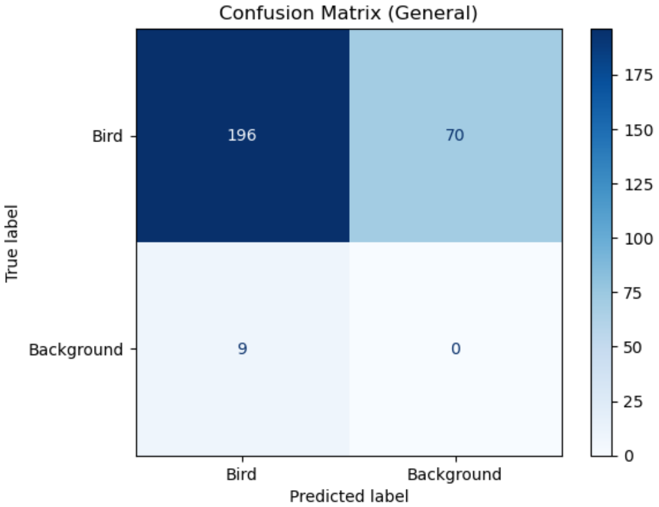


Figure 11: Binary confusion matrix for the Bird Song Detector on the test dataset with a confidence threshold of 0.15.

Using BirdNET as a detector, including the fine-tuned model, resulted in 80% of the positive samples being FPs in almost all experiments. When the confidence score of the fine-tuned model was lowered, the number of FNs decreased to 30%. However, this adjustment also led to a substantial increase in the number of FPs. Nearly 50% of the detected positives (Bird Songs) were false, representing background noise rather than actual Bird Songs. In contrast, our custom Bird Song Detector showed significantly better performance. It missed only 20% of the actual Bird Songs, resulting in minimal loss, and less than 1% of all predicted positive samples were FPs. This demonstrates the effectiveness of our detector in accurately identifying bird songs while minimizing false detections.

To further illustrate the performance of the detector, Figure 12 presents an example of the detector's predictions on a Mel spectrogram. The predictions are highly accurate and correspond to the annotated spectrogram segments.

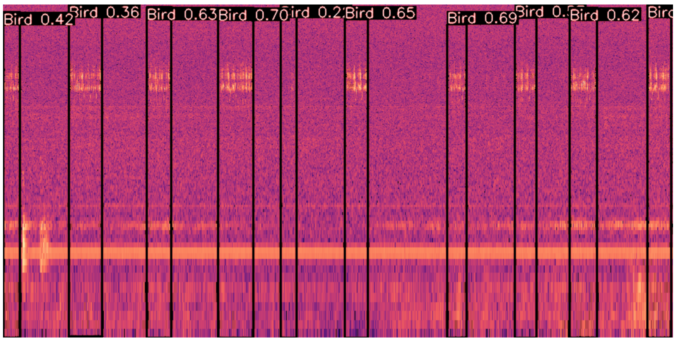


Figure 12: Predictions made by the Bird Song Detector on a Mel spectrogram. The predictions are highly accurate, corresponding to the annotated spectrogram segments in Figure 4.

5.8. Bird Song Detector vs BirdNET as a Bird Vocalization Detector

The comparison of the different models as bird vocalization detectors in Calculations (see Section 5.5) highlight a significant enhancement in TP and FN, with TP increasing by approximately 282% and FN decreasing by about 62%. However, there is a notable increase in FP, worsening by approximately 439%. While the FPs show a significant increase of approximately 439%, it is important to put this figure into context. The actual values we are comparing are 1.67 and 9, which are relatively small numbers, especially when considering the total duration of 200-300 minutes of audio data in the test dataset. In this timeframe, there could potentially be an infinite number of incorrect detections. However, in the case of BirdNET, only an average of 1.67 incorrect detections were made, and for our Bird Song Detector, this number was 9. Therefore, even though the percentage increase seems high, the actual impact on the overall performance of the detector is minimal. This demonstrates the effectiveness of our Bird Song Detector in minimizing false detections while accurately identifying bird songs.

The comparison of the Bird Song Detector and BirdNET fine-tuned with a low confidence score threshold shows that, while lowering the confidence score in BirdNET, our Bird Song Detector reduces the number of FN by approximately 44%, and outperforms the FP metric by reducing them about 95%. In contrast, the TP of our Bird Song Detector decreases by approximately 23%.

5.9. Bird Species Classification

With the temporal windows containing bird songs identified and extracted, these segments were then processed by the fine-tuned BirdNET model. The improved detection rates from the fine-tuned model allowed us to compute an species-specific confusion matrix. Figure 13 presents the species-specific confusion matrix obtained for the test segments obtained from the Bird Song Detector and classified by BirdNET fine tuned. This confusion matrix highlights the model's varying performance across different bird species even when BirdNET was fine tuned.

This approach minimizes effort and time since the temporal windows and cropped audio are already provided by the Bird Song Detector. In addition, BirdNET also provides a preliminary species classification. To ensure the accuracy and reliability of the species classification, even when BirdNET's confidence score is low, the identified segments can be reviewed by expert ornithologists.

6. Discussion

The results of our case study demonstrate the feasibility and effectiveness of using a combination of a custom Bird Song Detector and a fine-tuned BirdNET for bird identification. The application of YOLOv8 (21), a state-of-the-art detection model, allowed us to accurately identify specific temporal windows where bird songs occurred, detecting not only the presence of bird vocalizations but also precisely locating them within the audio recordings.

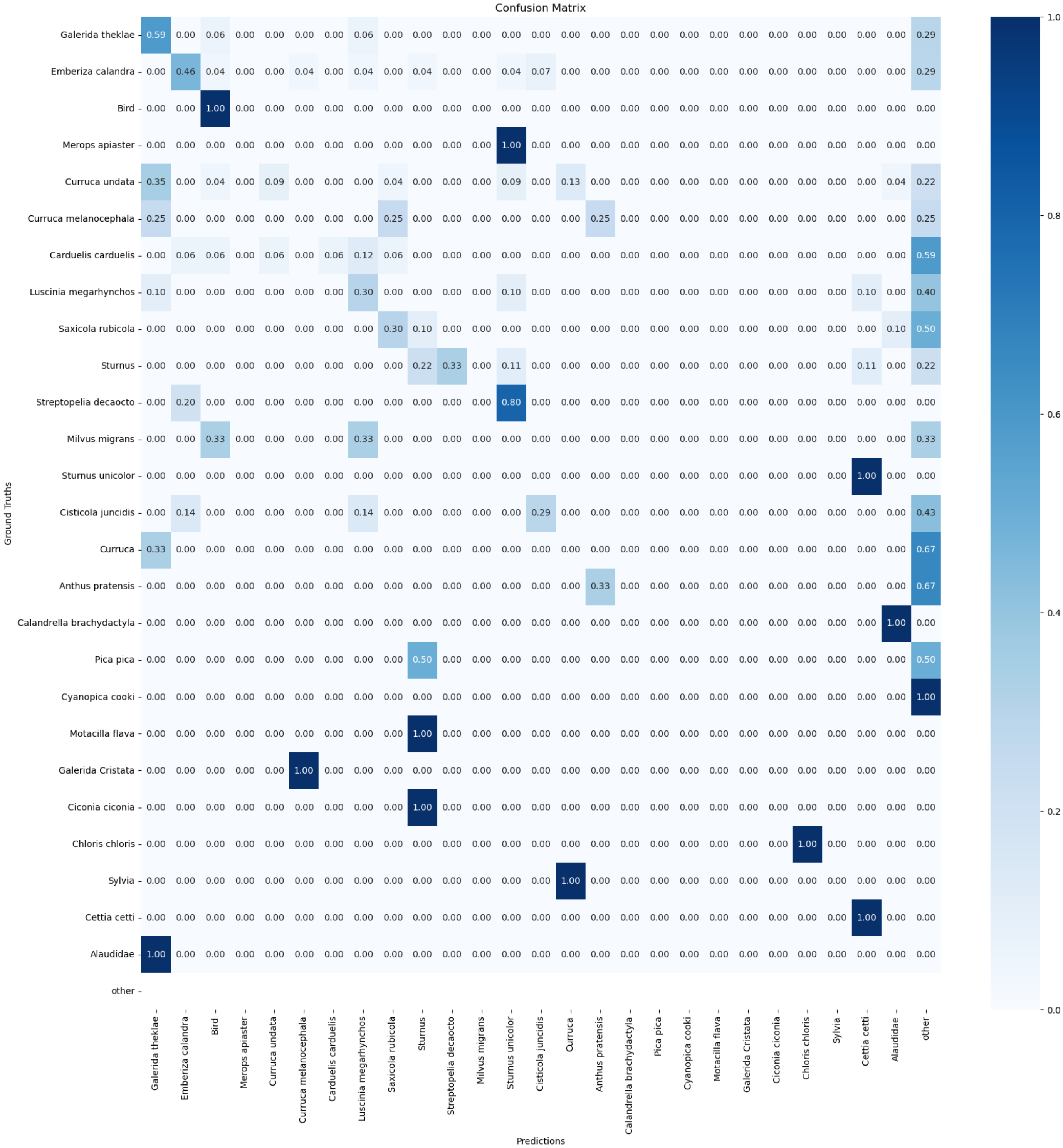
BirdNET, capable of identifying species within 3-second windows, faces challenges in accurately pinpointing the exact duration of bird vocalizations (11; 12). What is more, according to the original BirdNET's paper (10), the model achieved a mean average precision of 0.791 for single-species recordings, an F0.5 score of 0.414 for annotated soundscapes, and an average correlation of 0.251 with hotspot observations (areas with a high diversity of bird species). These results suggest that BirdNET performs inadequately when applied to real-world scenarios, such as accurately detecting bird species in diverse and dynamic environments (11).

BirdNET assigns confidence scores to its predictions. If a high confidence score threshold is set, the recall is good, but the sensitivity is very low. With a high confidence score, a lot of FNs are obtained, and a lot of bird vocalizations are lost. On the other side, lowering the confidence score results in too many FPs (segments without bird vocalizations) (12). However, by applying the Bird Song Detector first, we can ensure that the audio segments contain bird vocalizations, allowing us to lower the confidence score in BirdNET without increasing the FPs, as the initial detector filters out non-bird sounds.

Using BirdNET as a detector, including the fine-tuned model, resulted in about 80% of the positive predictions being FNs in almost all experiments. When the confidence score of the fine-tuned BirdNET model was lowered, the number of FN decreased to about 30%. However, this adjustment also led to a substantial increase in the number of FPs. Nearly 50% of the detected positives (bird vocalizations) were FP, representing background noise rather than actual bird songs (46). In contrast, the two-step process we employed, consisting of our custom Bird Song Detector followed by the fine-tuned BirdNET, showed significantly better performance. The Bird Song Detector effectively filtered out non-bird sounds, reducing the impact of background noise. As a result, the fine-tuned BirdNET worked with cleaner data, missing only 20% of the actual bird vocalizations, with less than 1% of all predicted positive samples being FPs. This demonstrates the effectiveness of our two-step approach in accurately identifying bird songs while minimizing false detections.

These findings emphasize the importance of using a high confidence score threshold with BirdNET to ensure a balance between detection accuracy and the minimization of FPs. However, our custom Bird Song Detector still outperforms BirdNET by effectively reducing FNs and maintaining a manageable FP rate and gives the possibility of applying BirdNET with a lower confidence score without having a high amount of FP and FN as it effectivily filters the segments that potencially have a bird vocalization.

For our Bird Song Detector, the logistic regression transformation of the confidence scores to logits provides a more nuanced thresholding approach, reducing the number of FNs and enhancing overall detection performance. The impact of the confidence threshold on detection performance was significant, it was possible to balance sensitivity and recall effectively.



	Galerida theklae	Emberiza calandra	Bird	Merops apiaster	Curruca undata	Curruca melanocephala	Carduelis carduelis	Luscinia megarhynchos	Saxicola rubicola	Sturnus	Streptopelia decaocto	Milvus migrans	Sturnus unicolor	Cisticola juncidis	Curruca	Anthus pratensis	Calandrella brachydactyla	Pica pica	Cyanopica cooki	Motacilla flava	Galerida Cristata	Ciconia ciconia	Chloris chloris	Sylvia	Cettia cetti	Alaudidae	other
Galerida theklae	0.59	0.00	0.06	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29
Emberiza calandra	0.00	0.46	0.04	0.00	0.00	0.04	0.00	0.04	0.00	0.04	0.00	0.00	0.04	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29
Bird	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Merops apiaster	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Curruca undata	0.35	0.00	0.04	0.00	0.09	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.09	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.22
Curruca melanocephala	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25
Carduelis carduelis	0.00	0.06	0.06	0.00	0.06	0.00	0.06	0.12	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.59
Luscinia megarhynchos	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.40
Saxicola rubicola	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.50
Sturnus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.33	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.22
Streptopelia decaocto	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Milvus migrans	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33
Sturnus unicolor	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Cisticola juncidis	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.43
Curruca	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67
Anthus pratensis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67
Calandrella brachydactyla	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Pica pica	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50
Cyanopica cooki	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
Motacilla flava	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Galerida Cristata	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ciconia ciconia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Chloris chloris	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Sylvia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cettia cetti	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Alaudidae	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
other																											

Figure 13: Species-specific normalized by rows confusion matrix for the fine-tuned BirdNET on predicted segments in the test dataset with the Bird Song Detector.

The Bird Song Detector included various abiotic and other animal sounds as background noise. Integrating these sounds into the training data enables the detector to generalize effectively, allowing it to detect bird vocalizations even for species it was not specifically trained on (14). By incorporating noises such as environmental sounds and other animal vocalizations into the training process, the detector becomes more robust and less prone to FPs. This approach helps mitigate instances where sounds from other animals like frogs or insects might be misclassified as bird vocalizations due to their overlap and common occurrence in natural environments (47). Overall, the results shown in Table 1 provide valuable insights into how different factors influence the performance of the Bird Song Detector, guiding future improvements for enhanced bird song identification.

Reviewing segments identified as bird songs but not confirmed by experts often revealed anthropogenic noises such as short fence hits. This highlights the complexity of accurately identifying bird vocalizations and underscores the importance of expert verification, especially in scenarios involving ambiguous sounds or background noises.

The fine-tuning of BirdNET for the Doñana ecological context improved the classification accuracy of bird species, addressing the specific acoustic environment and species composition of the area. BirdNET's built-in oversampling techniques were supplemented with data augmentation techniques like Mixup to mitigate class imbalance, where some species were underrepresented. Mixup involves linearly interpolating pairs of spectrogram images and their corresponding labels to create new training samples.

However, performance varied across different species, highlighting the influence of data availability and quality on model performance even after applying data augmentation techniques. Species with more abundant and diverse training data exhibited better classification results, underscoring the need for balanced and comprehensive datasets. Given that BirdNET's classification accuracy is currently the limiting factor in the overall performance of the system, future efforts should primarily focus on improving the fine-tuning and training of BirdNET. Enhancing the model's ability to handle underrepresented species and fine-tuning it further to the specific ecological context of Doñana will likely yield the most significant improvements in the pipeline's overall accuracy.

The varying performance among different species is likely due to the disproportionate amount of data available for each species in the training dataset. Different species may require distinct confidence thresholds even after fine-tuning (11; 12). This suggests that adaptive thresholding, where thresholds are dynamically adjusted based on species or context, could further improve classification accuracy.

Despite advances, the classifier still faces challenges in accurately identifying certain species. Expert review remains crucial to verify and refine classifications, especially for species with less robust data representation. The involvement of ornithologists ensures the reliability of the model's outputs and provides valuable feedback for further refinement (46).

Overall, the combination of the Bird Song Detector and the fine-tuned BirdNET for species classification has shown promising results. The proposed pipeline effectively identifies bird songs and provides a preliminary species-specific identification, with the flexibility to incorporate expert reviews for low-confidence detections, ensuring high accuracy and reliability in monitoring bird species in the Doñana region.

This work not only advances Deep Learning-based bird vocalization detection techniques but also underscores the importance of adapting these models to specific local contexts. The results demonstrate significant improvements in detection accuracy, paving the way for broader applications in ecological monitoring and conservation efforts.

7. Conclusions

In this section, we summarize the main findings and implications of our study on bird vocalization identification. We discuss the improvements in detection and classification accuracy, highlight the importance of species-specific considerations, and outline the continuous efforts needed to enhance model performance. Additionally, we explore the broader impact of our work on ecological monitoring and conservation, and propose future research directions to further refine and expand our approach.

This study presents a novel approach to bird vocalization identification through the integration of a custom Bird Song Detector with a fine-tuned BirdNET model. Our research focuses on enhancing the accuracy and efficiency of bird song detection and classification within specific ecological contexts, such as Doñana National Park, a biodiversity hotspot that requires detailed monitoring of avian species. By combining a YOLOv8-based Bird Song Detector with BirdNET, we aimed to overcome challenges associated with background noise and the misclassification of non-bird sounds. This approach allows for more precise identification of bird vocalizations, reducing the impact of irrelevant acoustic events and improving overall model performance.

The key advantage of our pipeline is its significant reduction of FPs and FNs compared to standalone BirdNET models. In a real-world application, our Bird Song Detector reduced FNs by approximately 62% while increasing TPs by 282% compared to a BirdNET-only approach. Although there was a 439% increase in FPs, this rise occurred from a very low baseline (from 1.67 to 9 FPs), making the overall increase in FPs minimal. Furthermore, when compared to a low-confidence fine-tuned BirdNET model, our Bird Song Detector achieved a 44% reduction in FNs and a 95% reduction in FPs, while maintaining only a modest 23% decrease in TPs. This demonstrates the effectiveness of our system in distinguishing bird sounds from background noise.

Despite the advances, some challenges remain in accurately identifying species with limited training data. Our findings emphasize the ongoing need for expert verification in refining classifications, especially for ambiguous sounds or species with fewer training samples. This expert validation will be crucial for improving the accuracy of species identification, particularly in complex soundscapes with overlapping vocalizations.

The practical implications of this research for ecological monitoring and conservation are significant. The integration of Deep Learning models allows for the cost-effective and scalable monitoring of bird populations, which is essential for tracking species trends and identifying critical habitats.

Looking forward, we aim to explore the potential benefits of newer YOLO models, such as YOLOv9 and YOLOv10, which were released after our initial experiments. These models could potentially enhance the detection capabilities of our pipeline, leading to further improvements in performance. Additionally, we are developing real-time detection systems to enable continuous monitoring of bird populations. This will be complemented by automated tools that assist ornithologists in reviewing low-confidence detections, streamlining the validation process.

Another long-term goal is to develop a foundational Bird Song Detector model using extensive libraries such as Macaulay and XenoCanto, as well as other datasets that include realistic scenarios with noise and various species. This foundational model could then be fine-tuned with specific data from projects like BirdNET, allowing for a more generalized pipeline that can be applied across various ecological contexts. Such a model would significantly enhance the scalability and adaptability of our approach, making it applicable to a wide range of ecosystems for global biodiversity monitoring.

In our case study, the Bird Song Detector demonstrated remarkable improvements, with True Positives increasing by approximately 282% and False Negatives decreasing by 62% compared with different BirdNET-only models. Although FPs increased by 439%, this increase was minimal, rising from just a medium of 1.67 to 9 FP detections. These results highlight the effectiveness of our pipeline in improving bird vocalization detection accuracy while minimizing erroneous detections, providing a powerful tool for ecological monitoring.

Ultimately, these efforts aim to create a robust and versatile bird vocalization detection system that can be deployed across diverse ecological contexts, thereby broadening the applicability and impact of our research in ecological monitoring and conservation.

Acknowledgements

This study has received financial support from the BIRDeep project (TED2021-129871A-I00), which is funded by MICIU/AEI/10.13039/501100011033 and the 'European Union NextGenerationEU/PRTR', as well as grants PID2020-115129RJ-I00 from MCIN/AEI/10.13039/501100011033.

We thank Giulia Bastianelli and Eduardo Santamaría García for annotating the audio. Eduardo Santamaría did the map presented in Figure 2.

References

[1] R. D. Gregory and A. van Strien, “Wild bird indicators: using composite population trends of birds as measures of environmental health,” *Ornithological Science*, vol. 9, no. 1, pp. 3–22, 2010.

[2] L. S. M. Sugai, T. S. F. Silva, J. W. Ribeiro Jr, and D. Llusia, “Terrestrial passive acoustic monitoring: review and perspectives,” *BioScience*, vol. 69, no. 1, pp. 15–25, 2019.

[3] R. Gibb, E. Browning, P. Glover-Kapfer, and K. E. Jones, “Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring,” *Methods in Ecology and Evolution*, vol. 10, no. 2, pp. 169–185, 2019.

[4] S. S. Farley, A. Dawson, S. J. Goring, and J. W. Williams, “Situating ecology as a big-data science: current advances, challenges, and solutions,” *BioScience*, vol. 68, no. 8, pp. 563–576, 2018.

[5] O. Metcalf, C. Abrahams, B. Ashington, E. Baker, T. Bradfer-Lawrence, E. Browning, J. Carruthers-Jones, J. Darby, J. Dick, A. Eldridge, *et al.*, “Good practice guidelines for long-term ecoacoustic monitoring in the uk,” 2023.

[6] K. Darras, P. Batáry, B. J. Furnas, I. Grass, Y. A. Mulyani, and T. Tscharntke, “Autonomous sound recording outperforms human observation for sampling birds: a systematic map and user guide,” *Ecological Applications*, vol. 29, no. 6, p. e01954, 2019.

[7] D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. Van Langevelde, T. Burghardt, *et al.*, “Perspectives in machine learning for wildlife conservation,” *Nature communications*, vol. 13, no. 1, pp. 1–15, 2022.

[8] S. C. Keen, K. J. Odom, M. S. Webster, G. M. Kohn, T. F. Wright, and M. Araya-Salas, “A machine learning approach for classifying and quantifying acoustic diversity,” *Methods in ecology and evolution*, vol. 12, no. 7, pp. 1213–1225, 2021.

[9] J. Xie, Y. Zhong, J. Zhang, S. Liu, C. Ding, and A. Triantafyllopoulos, “A review of automatic recognition technology for bird vocalizations in the deep learning era,” *Ecological Informatics*, vol. 73, p. 101927, 2023.

[10] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, “Birdnet: A deep learning solution for avian diversity monitoring,” *Ecological Informatics*, vol. 61, p. 101236, 2021.

[11] C. Pérez-Granados, “Birdnet: applications, performance, pitfalls and future opportunities,” *Ibis*, vol. 165, no. 3, pp. 1068–1075, 2023.

[12] C. M. Wood and S. Kahl, “Guidelines for appropriate use of birdnet scores and other detector outputs,” *Journal of Ornithology*, pp. 1–6, 2024.

[13] G. E. Schuster, L. J. Walston, and A. R. Little, “Evaluation of an autonomous acoustic surveying technique for grassland bird communities in nebraska,” *PloS one*, vol. 19, no. 7, p. e0306580, 2024.

[14] S. Beery, D. Morris, and S. Yang, “Efficient pipeline for camera trap image review,” *arXiv preprint arXiv:1907.06772*, 2019.

[15] P. Lauha, P. Somervuo, P. Lehikoinen, L. Geres, T. Richter, S. Seibold, and O. Ovaskainen, “Domain-specific neural networks improve automated bird sound recognition already with small amount of local data,” *Methods in Ecology and Evolution*, vol. 13, no. 12, pp. 2799–2810, 2022.

[16] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, “Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge,” *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368–380, 2019.

[17] M. L. Clark, L. Salas, S. Baligar, C. A. Quinn, R. L. Snyder, D. Leland, W. Schackwitz, S. J. Goetz, and S. Newsam, “The effect of soundscape composition on bird vocalization classification in a citizen science biodiversity monitoring project,” *Ecological Informatics*, vol. 75, p. 102065, 2023.

[18] J. Lalor, H. Wu, and H. Yu, “Improving machine learning ability with fine-tuning,” 02 2017.

[19] J. Hamer, E. Triantafillou, B. van Merrienboer, S. Kahl, H. Klinck, T. Denton, and V. Dumoulin, “Birb: A generalization benchmark for information retrieval in bioacoustics,” *arXiv preprint arXiv:2312.07439*, 2023.

[20] K. Martin, O. Adam, N. Obin, and V. Dufour, “Rookognise: Acoustic detection and identification of individual rooks in field recordings using multi-task neural networks,” *Ecological Informatics*, vol. 72, p. 101818, 2022.

[21] G. Jocher, A. Chaurasia, and J. Qiu, “Ultralytics YOLO,” Jan. 2023.

[22] N. Rigoudy, G. Dussert, A. Benyoub, A. Besnard, C. Birck, J. Boyer, Y. Bollet, Y. Bunz, G. Caussimont, E. Chetouane, *et al.*, “The deepfaune initiative: a collaborative effort towards the automatic identification of european fauna in camera trap images,” *European Journal of Wildlife Research*, vol. 69, no. 6, p. 113, 2023.

[23] Birdeep.org, “Birdeep.” `https://birdeeporg.github.io`, 2024. Accedido el 14 de abril de 2024.

[24] M. A. Rendón, A. J. Green, E. Aguilera, and P. Almaraz, “Status, distribution and long-term changes in the waterbird community wintering in doñana, south–west spain,” *Biological Conservation*, vol. 141, no. 5, pp. 1371–1388, 2008.

[25] A. J. Green, J. Bustamante, G. Janss, R. Fernández-Zamudio, C. Díaz-Paniagua, *et al.*, “Doñana wetlands,” 2016.

[26] A. P. Hill, P. Prince, J. L. Snaddon, C. P. Doncaster, and A. Rogers, “Audiomoth: A low-cost acoustic device for monitoring biodiversity and the environment,” *HardwareX*, vol. 6, p. e00073, 2019.

[27] T. Audacity, "Audacity," *The name audacity (R) is a registered trademark of dominic mazzoni retrieved from http://audacity. sourceforge. net*, 2017.

[28] C. S. Robbins, "Effect of time of day on bird activity," *Studies in avian biology*, vol. 6, no. 3, pp. 275–286, 1981.

[29] A. Márquez-Rodríguez, M. Á. Muñoz-Mohedano, M. J. Marín-Jiménez, E. Santamaría-García, G. Bastianelli, and I. Mendoza, "Birdeepaudioannotations (revision 4cf0456)," 2024.

[30] S. Carvalho and E. F. Gomes, "Automatic classification of bird sounds: using mfcc and mel spectrogram features with deep learning," *Vietnam Journal of Computer Science*, vol. 10, no. 01, pp. 39–54, 2023.

[31] L. Wyse, "Audio spectrogram representations for processing with convolutional neural networks," *arXiv preprint arXiv:1706.09559*, 2017.

[32] M. Hardy, "Pareto's law," *The Mathematical Intelligencer*, vol. 32, pp. 38–43, 2010.

[33] T. Kattenborn, F. Schiefer, J. Frey, H. Feilhauer, M. D. Mahecha, and C. F. Dormann, "Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks," *ISPRS Open Journal of Photogrammetry and Remote Sensing*, vol. 5, p. 100018, 2022.

[34] K. Pasupa and W. Sunhem, "A comparison between shallow and deep architecture classifiers on small dataset," in *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pp. 1–6, IEEE, 2016.

[35] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

[36] S. Kahl, "Birdnet-analyzer." `https://github.com/kahst/BirdNET-Analyzer`, 2021.

[37] K. M. Brunk, R. Gutiérrez, M. Z. Peery, C. A. Cansler, S. Kahl, and C. M. Wood, "Quail on fire: changing fire regimes may benefit mountain quail in fire-adapted forests," *Fire Ecology*, vol. 19, no. 1, pp. 1–13, 2023.

[38] D. Sossover, K. Burrows, S. Kahl, and C. M. Wood, "Using the birdnet algorithm to identify wolves, coyotes, and potentially their interactions in a large audio dataset," *Mammal Research*, vol. 69, no. 1, pp. 159–165, 2024.

[39] J. Schlüter, "Bird identification from timestamped, geotagged audio recordings.," *CLEF (Working Notes)*, vol. 2125, 2018.

[40] X. canto Foundation, "Xeno-canto: Bird sounds from around the world," 2024. Accessed: 2024-07-13.

[41] Macaulay Library, "The world's premier scientific archive of natural history audio, video, and photographs." `https://www.macaulaylibrary.org/about/history/`, 2024. Accessed: 2024-07-13.

[42] L. García, F. Ibáñez, H. Garrido, J. Arroyo, M. Máñez, and J. Calderón, "Prontuario de las aves de doñana. anuario ornitológico de doñana, nº 0, diciembre 2000," 2000.

[43] H. Garrido, J. Arroyo, L. García, F. Ibáñez, M. Máñez, and M. Vázquez, "Anuario ornitológico de doñana, nº 1 (septiembre 1999–agosto 2001)," *Estación Biológica de Doñana y Ayuntamiento de Almonte, Almonte (Huelva)(in Spanish)*, 2004.

[44] R. Padilla, W. L. Passos, T. L. Dias, S. L. Netto, and E. A. Da Silva, "A comparative analysis of object detection metrics with a companion open-source toolkit," *Electronics*, vol. 10, no. 3, p. 279, 2021.

[45] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pp. 1015–1018, ACM Press.

[46] G. Bota, R. Manzano-Rubio, L. Catalán, J. Gómez-Catasús, and C. Pérez-Granados, "Hearing to the unseen: Audiomoth and birdnet as a cheap and easy method for monitoring cryptic bird species," *Sensors*, vol. 23, no. 16, p. 7176, 2023.

[47] A. Robey, H. Hassani, and G. J. Pappas, "Model-based robust deep learning: Generalizing to natural, out-of-distribution data," *arXiv preprint arXiv:2005.10247*, 2020.