Audio vs. Visual Approach to Monitor the Critically Endangered Species *Atlapetes blancae*: Developing Deep Learning Models with Limited Data

Julian D. Santamaria P
SISTEMIC, Engineering Faculty
Universidad de Antioquia-UdeA
Cl. 67 No. 53-108
Medellín, Colombia
email: julian.santamaria@udea.edu.co

Jhony H. Giraldo
LTCI, Télécom Paris
Institut Polytechnique de Paris
Palaiseau
Paris, France
email: jhony.giraldo@telecom-paris.fr

Angélica Diaz-Pulido
Alexander Von Humboldt Institute
Neotropical Innovation Corporation
Cl. 28a No. 15-09
Bogotá, Colombia
email: adiaz@humboldt.org.co

Claudia Isaza
SISTEMIC, Engineering Faculty
Universidad de Antioquia-UdeA
67 No. 53-108
Medellín, Colombia
email: victoria.isaza@udea.edu.co

***Abstract*—Using artificial intelligence algorithms for animal passive monitoring is a cost-effective tool. This kind of data analysis permits detailed and efficient tracking of species, as exemplified by the case of the endemic Antioquia brushfinch (*Atlapetes blancae*). *Atlapetes blancae* is from the high-elevation plateau of Santa Rosa de Osos in Antioquia Colombia. These birds are currently listed as critically endangered by the International Union for Conservation of Nature (IUCN). Their population is estimated at approximately 108 individuals. Sound recorders and camera traps are important tools for long-term monitoring as they provide extensive registers of data. However, analyzing this data is a labor-intensive process that requires experts to manually process the extensive amount of information. Additionally, identifying acoustic patterns for the *Atlapetes blancae* species based on artificial intelligent algorithms is problematic due to the lack of labeled data and the complexity of the vocalizations. This study introduces a novel methodology for real-environment audio analysis, addressing the challenge of unlabeled registers using a semi-automatic approach. We leverage the Learning Algorithm for Multivariate Data Analysis (LAMDA) and KiwiNet convolutional network architecture for audio recognition. Additionally, we analyze the videos using Multi-Layer Robust Principal Component Analysis (Multi-layer RPCA) to obtain cropped images from the video, which are then processed using a ResNet-18 architecture for classification. Finally, we compare both models to identify strengths and limitations. With a collection of 7,147 audio recordings and 17,159 videos, only 11 audio and 48 video recordings contain *Atlapetes blancae* presence. Our approach achieves F-measure average scores of 0.823 and 0.562 for audio and video analysis, respectively. Notably, in this case, the audio model is more robust than the video model.**

Keywords- Atlapetes blancae identification; Computer vision; Bioacoustics; Passive monitoring.

I. INTRODUCTION

The *Atlapetes blancae* is an endemic bird from the Santa Rosa de Osos high elevation plateau in the Department of Antioquia-Colombia [1]. Currently, it is on the International Union for Conservation of Nature (IUCN) Red List as "critically endangered" [2]. The first *Atlapetes blancae* description was made in 2007 [3]. In this description, it was listed as "possibly extinct" due to deforestation in its locality. However, rediscovery of *Atlapetes blancae* was reported in 2018 [4], supported by photographic evidence confirming its presence. Efforts by organizations such as the Neotropical Innovation Corporation (Neotropical Innovation link) and the Alexander von Humboldt Institute have been crucial in developing conservation strategies for this species. Neotropical Innovation Corporation's latest research reveals that they have estimated a population of only 108 individuals.

Conservation plans require implementing species monitoring to estimate population state variables, such as occupancy. A cost-efficient alternative to studying species is passive monitoring. Audio and video monitoring uses sensors, such as camera traps and sound recorders to make registers over the long term in different geographic locations and throughout the day [5]. Passive Acoustic Monitoring (PAM) offers an alternative method for studying and monitoring wildlife with audio recorders [6], while camera traps serve as the alternative when seeking visual data through images or videos. The ease of data collection is an exceptional advantage since it is a non-invasive technique that does not disrupt the natural behavior of the observed species. Furthermore, a substantial volume of registers are acquired for monitoring with acoustic recorders and camera traps over long periods [7], [8]. Nevertheless, the majority of the registers do not contain the presence of the target species. Therefore, it becomes necessary to have computational tools to assist in the analysis of the obtained video and audio recordings [9], [10]. In recent analyses, supervised artificial intelligence techniques have demonstrated impressive performance in the identification of specific species based on audio data [11], [12] as well as in videos [7], [13]. However, it is worth noting that these methods rely heavily on expert-labeled registers [14], which can be a significant

challenge, particularly when dealing with endemic and critically endangered species [8] due to their low probability of occurrence and limited recorded instances.

To address the challenge of spending too much time listening to audio, analyzing spectrograms, and labeling datasets to train models, we propose a semi-automatic methodology. In our approach, we incorporate Guerrero's unsupervised method [15] to uncover potential patterns in relevant vocalizations. The expert's task is simplified to analyzing and confirming the presence of *Atlapetes blancae* within the patterns identified during the unsupervised analysis, instead of manually labeling. To enhance the initial analysis, we recommend using a limited set of species songs as examples, ensuring a more thorough examination of the species' acoustic repertoire. This method enables the identification of distinctive acoustic patterns of the target species, expediting the process. The second part of our methodology employs Arbimon software's pattern-matching algorithm [16] to evaluate the entire dataset. By leveraging the acoustic pattern established in the earlier analysis as a template, this process significantly reduces the need for manual audio analysis and permits experts to automatically label registers. Finally, we employ transfer learning to train our classification model using a pre-trained Convolutional Neural Network (CNN) - KiwiNet [17]. This method is explained in detail in Section III-A.

Recent research has highlighted the capability of CNNs in identifying animal species in camera trap images [18]. Furthermore, adopting segmentation as a preliminary step is an alternative approach that enhances the performance of the model [14]. In our work, we employ Multi-Layer Robust Principal Component Analysis (Multi-layer RPCA) for camera-trap image segmentation [19] to process the videos of our dataset as a preliminary step. Subsequently, we utilize the segmentation images obtained to train a CNN, more specifically a ResNet-18 [20], to classify *Atlapetes blancae* images. This approach is described in Section III-B.

To our knowledge, there is no proposal that leverages unsupervised methods to analyze acoustic patterns from a species with little information and then uses this knowledge to employ a semi-automatic methodology for labeling the recordings. The computational tool developed can be downloaded from [21].

The structure of this article is organized as follows: Section II provides an overview of the related work in the field. Section III outlines the methodology employed in our study. Section IV presents the data base used in our analysis. The results obtained from our analysis are presented in Section V. Finally, conclusions and future work are presented in Section VI.

II. Related Work

A. Audio recognition

In the specific field of *Atlapetes blancae* recognition, Diaz-Vallejo *et al.* [1] used the Raven Pro software (Raven Pro is a software for the visualization, measurement, and acoustic analysis of sound recordings) [22] to estimate occupancy of the *Atlapetes blancae* from audio recordings. However, the annotation process involves manual listening of audio recordings and visualization of audio spectrograms to identify and label different animal vocalizations within them.

For the recognition of other bird species, there are specialized tools available, such as BirNet [23], Merlin Bird ID [24] and KiwiNet [17]. BirNet is a Deep artificial Neural Network (DNN) that uses sound data to identify North American and European bird species. It is trained to recognize 984 bird species, excluding specific species like *Atlapetes blancae*. On the other hand, Merlin Bird ID is a mobile application that incorporates a sound identification feature (Sound ID). It is trained to identify 1,054 species of birds, focusing primarily on birds found in the United States, Canada, Europe, and the Western Palearctic region. Similarly, *Atlapetes blancae* is not included in the species list covered by Merlin Bird ID [24]. Finally, KiwiNet [17] is a CNN specifically trained to identify bird calls, focusing on the Kiwi, a native New Zealand bird species.

In current models for bird species recognition, there is no model that already knows about *Atlapetes blancae*. However, it is possible to enhance the *Atlapetes blancae* sound classification task by utilizing pre-trained representations [8], [25]. Pre-trained CNNs offer starting points for audio-based recognition tasks and can be adaptable for *Atlapetes blancae* with transfer learning. This technique presents a viable solution for mitigating the challenge of limited labeled registers available for training CNNs that normally require a huge amount of data.

Another practical method to handle the problem of *Atlapetes blancae* recognition is clustering. This technique is particularly useful when we are working with unlabeled data because it helps group similar data. This approach provides a different perspective on the dataset and can help us identify interesting connections between data points [15]. The acoustic animal identification method proposed by Guerrero [15] is a clustering-based alternative that can identify sound groups without requiring prior knowledge of the number of different animal sounds. The approach consists of two parts: the first identifies sonotypes and matches them with the cluster that best represents them, while the second attempts to match sounds to specific animals. However, the second part requires a large number of examples of the sounds made by each animal. Unfortunately, we do not have many sound examples of *Atlapetes blancae*, which makes it unsuitable for bioacoustic monitoring and analysis of this bird species.

On the other hand, multispecies sound recognition software is also commonly used for this task. One of the most famous is the Arbimon software [16]. This supervised model is based on Random Forest [26], a technique that combines multiple decision trees to analyze bioacoustic data. Nevertheless, as a supervised model, it requires labeled registers to train a specialized classifier in *Atlapetes blancae*.

B. Image recognition

Research on species image recognition has been limited, especially about bird identification. Even fewer studies have attempted to identify bird species based on images [27], [28].

Generally, birds are treated as a broad category by classifiers [13], [29].

There are several options available for animal image identification, such as Conservation AI [30], Merlin Bird ID [24], MLWIC2 [31], and Wildlife Insights AI model [32]. However, these are supervised models trained on bird datasets that do not include *Atlapetes blancae*. They require a significant number of examples for training as an *Atlapetes blancae* classifier.

Object detection models like MegaDetector [18] and DeepWILD [14] have become crucial for automating wildlife monitoring from camera trap images. While MegaDetector [18] is an image detection model that is capable of detecting images without animals, people, and vehicles in camera-trap images, it requires a significant amount of labeled data and annotated bounding boxes for training examples [33]. Additionally, the model's performance may vary depending on the size of the animal, making it less useful for identifying certain species, such as *Atlapetes blancae*. On the other hand, DeepWILD [14] is used to detect, classify, and count species in camera trap videos with a primary focus on monitoring the wolf's presence.

III. Methodology

This section presents our proposal for recognizing *Atlapetes blancae* in audio recordings and camera-trap videos.

A. Audio analysis proposal

This work proposes a semi-automatic methodology to analyze unlabeled registers, addressing the issue of the unlabeled presence or absence of the *Atlapetes blancae* in audio data. This stage involves identifying vocalization patterns within the complex song of the *Atlapetes blancae*, enabling subsequent labeling of the audio recordings. Following this, a model is trained to recognize the *Atlapetes blancae* in new audio recordings, as illustrated in Figure 1.

1) Preprocessing: In the preprocessing stage, we employ a technique known as acoustic animal identification to extract acoustic data in an unsupervised manner. This technique is based on the research conducted by Guerrero et al. [15], which utilizes segmentation and clustering to extract acoustic data from soundscapes. The segmentation process is based on a modified version of the Acoustic Event Detection (AED) algorithm [34]. The clustering stage utilizes the LAMDA algorithm [35] to make clusters that describe possible sonotypes in the soundscape. The lack of relevant acoustic information of *Atlapetes blancae* makes this technique particularly useful for the purposes of learning about the variable vocal repertory of the endemic species. Using this preprocessing, we identify a representative acoustic pattern for *Atlapetes blancae*.

2) Audios labeling - Pattern Matching: The acoustic representative patterns of *Atlapetes blancae* identified in the last step are used to recognize possible vocalizations in the entire not labeling dataset. We use the pattern-matching tool within the Arbimon software [16]. This tool performs a pattern-matching algorithm by comparing a given pattern with elements in the dataset to identify matching occurrences. The pattern-matching tool provides potential segments in the audio along with their scores. This pattern-matching process does not have high performance when it comes to recognizing the presence of *Atlapetes blancae* in audio recordings. Consequently, we only use this pattern-matching as the previous step in the labeling procedure. A manual validation process is conducted to verify only the segments that match the given audio representative patterns. The labeled audios are subsequently used to create a training and testing set for training a classification model.

3) Supervised training: We use transfer learning to train KiwiNet [17] for our *Atlapetes blancae* recognition problem. KiwiNet is a Convolutional Neural Network (CNN) based on VGG19 [36] architecture and is used for supervised training in acoustic data analysis and identifying individuals based on their calls. VGG19 was modified to improve the regularization of the latent space when used as a feature extractor [17]. KiwiNet and VGG19 differ in that KiwiNet has a convolutional layer before the fully connected layers to reduce the number of filters from 512 to 32 and a global average pooling layer to embed the call characteristics into a 32-element feature set (latent space). Moreover, the KiwiNet analyzes the input data utilizing a colormap (KRGB - Black-Red-Green-Blue) to correlate the image colors with the levels of intensity in the spectrogram. Additionally, the model applies a median equalizer after spectrogram estimation to noise-reduce the data [17]. Furthermore, the backbone of KiwiNet (VGG19) was pre-trained with the ImageNet dataset [37].

In this work, the KiwiNet [17] is trained using the Stochastic Gradient Descent (SGD) optimizer [38] with a learning rate of 0.0001 for 15 epochs on 1-minute recordings with the primary objective of classifying the input record in target class (presence of *Atlapetes blancae*) or noisy class (absence of *Atlapetes blancae*). Additionally, the spectrogram calculation parameters are configured as follows: the discrete Fourier transform utilizes 1024 sampling points, the spectrogram's window length is set to 1024, and the overlap between consecutive windows is 768. This supervised approach enhances the performance of recognizing *Atlapetes blancae* in new audio recordings.

4) Recognition: In the training stage, we trained a model for *Atlapetes blancae* recognition (KiwiNet [17]). This model is used to classify new audios of real-environment and determine the presence or absence of *Atlapetes blancae*. The preprocessing use in the training phase is not necessary for this stage. The KiwiNet is trained on 1-minute recordings (as our audio recordings). Therefore, the recordings can be directly passed to the model for classification.

Based on this methodology, our approach introduces several novel elements. First, we propose a semi-automatic method for analyzing unlabeled audio recordings, specifically targeting the detection of *Atlapetes blancae*. Unlike traditional methods, our approach combines unsupervised acoustic data extraction with supervised learning. By utilizing the segmentation and clustering techniques from the acoustic animal identification

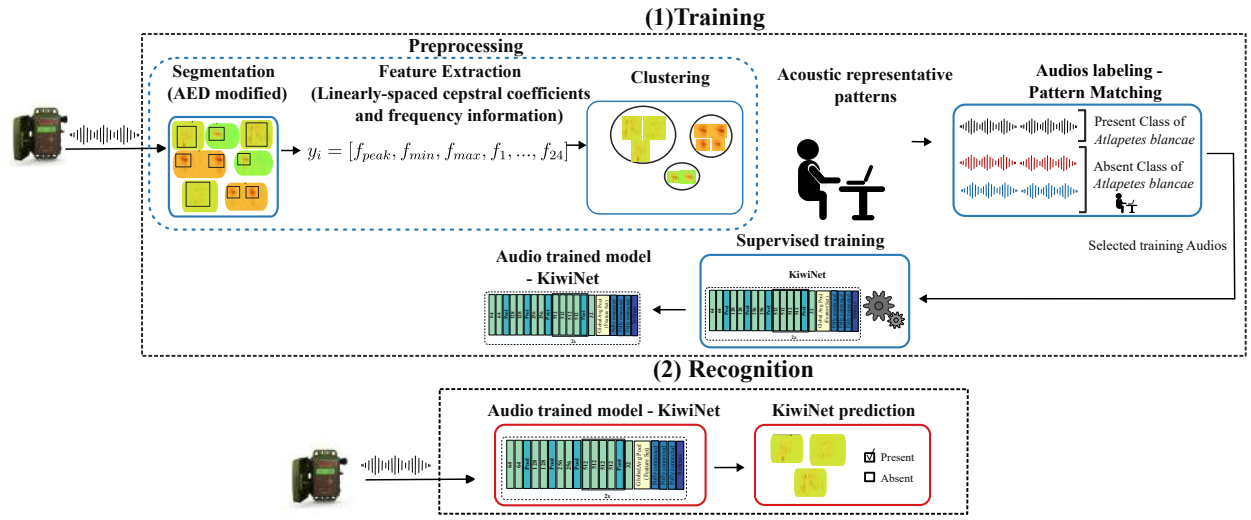

Figure 1. Proposed audio methodology schema: Spectrograms are segmented, features extracted, and clustered using acoustic animal identification. Patterns are identified and labeled to train a KiwiNet model. For recognition, the model classifies new recordings.

method [15], we identify representative vocalization patterns of *Atlapetes blancae*. These patterns are then used in conjunction with the Arbimon software's pattern-matching tool to label potential segments in the dataset, followed by a manual validation process to ensure accuracy. Furthermore, our use of KiwiNet [17], a modified VGG19 architecture, leverages transfer learning to enhance the recognition of *Atlapetes blancae*. The use of a colormap (KRGB) in KiwiNet, along with median equalization for noise reduction, ensures robust performance even in noisy environments. This integrated methodology not only improves the accuracy of *Atlapetes blancae* detection in new audio recordings but also addresses the challenge of working with unlabeled data.

B. Video analysis proposal

In the video analysis methodology, as shown in Figure 2, the first step is to extract the frames. Then, we use the Multi-layer RPCA method [19] to segment these frames and obtain a bounding box, which facilitates the cropping of the original frame and extraction of the image background. We refer to the results of the segmentation stage as cropped images. Afterward, we manually label the cropped images where the *Atlapetes blancae* is present to train the ResNet-18 architecture [20].

1) Segmentation: The Multi-Layer RPCA, proposed by [19], is utilized for camera-trap image segmentation [39] and incorporates texture and color descriptors. This approach decomposes an image into a low-rank matrix representing the background and a sparse matrix representing the foreground in background subtraction. The algorithm employed in this study involves the computation of Multi-layer RPCA, followed by a post-processing step.

During Multi-layer RPCA computation, the sparse and low-rank matrices are calculated for background subtraction. To evaluate the impact of texture descriptors on the entire image, we utilize a parameter called $\beta \in [0, 1]$. In this work, we chose $\beta = 0.6$, based on the best performance observed with our dataset. [19] tested nine algorithms to solve the RPCA problem, and among them, we select the Non-Smooth Augmented Lagrangian v1 (NSA1) algorithm [40] due to its effectiveness with our dataset [19], [39]. The post-processing step involves the application of morphological filters, as described by [19], [39].

2) Image cropping: After segmentation, a binary image is obtained with the bounding box of the segmented object. This bounding box is used to locate and crop the original frame, enabling background subtraction.

3) Image feature-based categorization: The cropped images are classified into four distinct classes: the target class (*Atlapetes blancae*) and three other classes (other birds, animals, and background). The segmentation stage enables us to isolate the moving objects in the videos, which may consist *Atlapetes blancae*, other animals, or noise. Consequently, a classification model is necessary to learn distinguishing patterns and accurately differentiate *Atlapetes blancae* from other moving animals and objects within the videos.

To improve the model's accuracy in distinguishing *Atlapetes blancae* from other bird species, we introduce an additional bird class. This inclusion enhances the model's specificity and enables more precise differentiation. The dataset is divided manually into training and test sets, comprising the target class (*Atlapetes blancae*) and three other classes (other birds, animals, and background).

4) Supervised training: ResNet-18 is a CNN architecture introduced by [20]. It belongs to the ResNet family of models, specifically designed to tackle the issue of vanishing gradients in deep neural networks. This architecture consists of a series of convolutional layers followed by residual blocks [20]. Furthermore, the architecture resizes the image with its shorter side randomly sampled in the range [256, 480] for scale

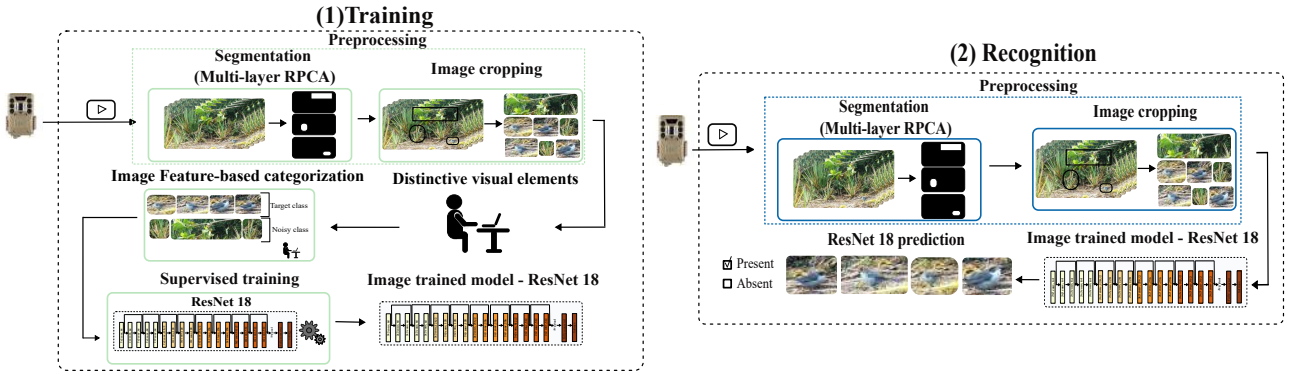

Figure 2. Proposed video methodology schema: Frames are segmented with Multi-layer RPCA, cropped, and used to train a ResNet-18 model. For recognition, new frames are similarly processed and classified by the trained model.

augmentation as part of the preprocessing of the input images. Next, a 224 x 224 crop with per-pixel mean subtraction is randomly selected from the scaled image or its flipped form horizontally [20].

The ResNet-18 architecture was employed with its original parameters and pre-trained weights from the ImageNet dataset [37]. In this work, we trained the ResNet-18 with the principal aim to classify input cropped images into four different classes: one target class (*Atlapetes blancae*) and three other classes (other birds, other animals, and background). As a result, we adjust the output size of the fully connected layer from 1000 to 4, reflecting the number of classes in our dataset. The model was trained using the SGD [38] optimizer with a learning rate of 0.0001 and a momentum of 0.9 for 15 epochs.

5) Recognition: In this stage, we apply the pre-processing step to extract cropped images and use the previously trained model to determine the presence or absence of *Atlapetes blancae*.

Based on this methodology, our approach introduces several novel elements in the field of video analysis for wildlife detection. Firstly, we implement a semi-automatic process that combines Multi-layer RPCA segmentation with manual labeling to create a robust training set. This allows for the precise extraction of frames containing *Atlapetes blancae* from complex backgrounds. Unlike traditional segmentation techniques, Multi-layer RPCA effectively decomposes frames into background and foreground components, enabling accurate isolation of the target species. Additionally, by training the ResNet-18 model on these segmented and manually labeled images, we ensure that the model learns to distinguish *Atlapetes blancae* from other birds, animals, and noise with high specificity. Our inclusion of an additional bird class further enhances the model's precision in identifying *Atlapetes blancae* amidst similar species. This comprehensive methodology not only improves detection accuracy but also addresses the challenges of working with unlabeled and complex video data, providing a significant advancement in automated wildlife monitoring.

C. Evaluation Metrics

The F-measure with macro averaging is chosen as the metric to evaluate the performance of the different models using the test data featuring N classes. Our evaluation involves a comparison of our audio methodology proposal with two software solutions, Arbimon [16] and the acoustic animal identification method [15], alongside one CNN architecture, ResNet-18 [20]. Similarly, we assess our video methodology proposal against two distinct ResNet architectures, ResNet-50 and ResNet-101 [20]. We use the F-measure as the metric of performance, which is given as follows:

First, calculate precision and recall for each $class_i$:

$$\begin{aligned} \text{precision}_i &= \frac{\text{True Positives}_i}{\text{True Positives}_i + \text{False Positives}_i}, \\ \text{recall}_i &= \frac{\text{True Positives}_i}{\text{True Positives}_i + \text{False Negatives}_i}, \end{aligned} \tag{1}$$

Then, average precision and recall across all classes:

$$\begin{aligned} \text{Precision avg} &= \frac{1}{N}\sum_{i=1}^{N}\text{precision}_i, \\ \text{Recall avg} &= \frac{1}{N}\sum_{i=1}^{N}\text{recall}_i, \end{aligned} \tag{2}$$

Finally, calculate the F-measure average:

$$\text{F-measure avg} = 2\frac{\text{Precision avg} \times \text{Recall avg}}{\text{Precision avg} + \text{Recall avg}}, \tag{3}$$

where precision (or confidence) denotes the proportion of predicted positive cases that are correctly real positives. On the other hand, recall (or sensitivity) represents the proportion of real positive cases that are correctly predicted positive [41]. Additionally, to evaluate the performance of the audio and video recognition model, we set aside 56 audio samples and 11,105 frames as test data for the audio and video recognition model.

IV. Data base

A. Study site

The study was conducted in the Yarumal and Santa Rosa de Osos municipalities on the northern highlands of Antioquia

of the central Andes mountain in Colombia. Two areas were sampled from November 2021 to February 2022. These areas are Batallón BITER IV (Tactical Instruction Battalion of the Colombian National Army) and El Vergel (cattle farm); in both cases with natural areas of *Atlapetes blancae* habitat.

B. Study case

1) Audios: We installed three acoustic sensors (SM4, Wildlife Acoustics) in each of the two sampling locations (Batallón and El Vergel), where sporadic observations of *Atlapetes blancae* had been previously reported. One acoustic sensor was placed in the El Vergel location and two in the Batallón location, with a separation of 500 meters between them. We sampled recordings for two months (November - December 2021). The sensors were set to record 60-second audio clips every 15 minutes, covering sound registers from 0:00 to 24:00 UTC. We recorded the audio in uncompressed WAV format, with a sampling rate of 48 kHz and a bit rate of 768 kbps. In total, we collected 7,147 audios, out of which *Atlapetes blancae* song was present in only 11 recordings. For the training phase, we selected 124 audio samples of absences and 8 of presences, while 53 audio samples of absences and 3 of presences were reserved for testing.

2) Videos: For two months (December 2021 - February 2022), we deployed 13 camera traps (Bushnell Trophy cam), seven in the Batallón and six in the El Vergel locations. The mean distance between sample sites was 200 m. Each camera trap was placed around 50 cm above the ground, recording videos of 15 seconds in response to the activation of a passive infrared sensor. Out of the 17,159 videos that were collected, only 48 have the presence of *Atlapetes blancae*. A total of 30,759 frames, which is an individual image captured from the video sequence, were randomly selected for the training dataset, while an additional 11,105 frames were set aside for testing data.

V. Results

In order to compare the audio and the video approaches, this section presents the results of both methodologies described in Section III. To assess the performance of the models, we applied the metric described in Section III-C to the test data.

A. Audio trained models

The audio recognition model (see Fig. 1, part 2-bottom), based on KiwiNet [17], analyzes each input audio to identify the specific acoustic pattern of *Atlapetes blancae*. In order to evaluate the proposed audio model results, we compared our audio model with software applications designed for recognizing multiple species classes, such as the acoustic animal identification method [15] and Arbimon software [16]. In each case, the model analyses the performance considering the presence of *Atlapetes blancae* in the whole audio and not for segments. Table I compares the F-measure average, precision average, recall average and accuracy obtained from the audio recognition models. The results reveal that the best-performing model is our audio model based on KiwiNet with a F-measure average of 0.823 and an Accuracy of 0.964. The performance of our audio recognition model is attributed to one of the primary functions of KiwiNet, which is to identify individuals based on their calls. In this study, we leverage the pre-existing knowledge of the KiwiNet architecture as a starting point to search for *Atlapetes blancae* by utilizing the acoustic patterns identified through the acoustic animal identification method. The software Arbimon has a F-measure average of 0.794 and an accuracy of 0.964, which matches the accuracy of our audio recognition model but has a lower F-measure average. Furthermore, when comparing recall average, Arbimon performs significantly worse than our model. In contrast to Arbimon software, which requires the user to specify a Region Of Interest (ROI), our audio model based on KiwiNet architecture, automatically identifies patterns by searching within its database. The acoustic animal identification method proposed in [15], used as a classifier and not like acoustic patterns identification, achieved a F-measure average of 0.743 and an accuracy of 0.929. While the accuracy shows only a slight decrease compared to our audio model, the F-measure average is significantly lower. Additionally, when comparing precision average, their method performs worse than our model. Unlike supervised models, which require labeled data to learn specific acoustic patterns of the target class, the unsupervised model is trained without labels and identifies various acoustic patterns from different species. However, due to the nature of our target species, a more tailored model for *Atlapetes blancae* recognition is required in this case.

TABLE I
COMPARING THE PERFORMANCE OF AUDIO MODELS ON TESTING DATA.

Model	F-measure avg	Precision avg	Recall avg	Accuracy
Our audio recognition model	0.823	0.823	0.823	0.964
Acoustic animal identification [15]	0.743	0.690	0.805	0.929
Arbimon [16]	0.794	0.981	0.667	0.964
ResNet-18 [20]	0.653	0.580	0.748	0.821

When using CNN architectures as classifiers in audio processing, the most common approach involves the analysis of audio spectrograms, specifically focusing on species vocalization rather than the entire audio spectrogram [9]. For this reason, we utilized the ResNet-18 architecture pre-trained on ImageNet [37]. We trained the ResNet-18 with output segments of the AED algorithm, which capture the part of the spectrogram where the animal vocalization is present. It is worth noting that this ResNet-18 is not trained with the whole spectrogram as the models before, therefore, we take into account the number of segments corresponding to each audio recording to calculate the result. The ResNet-18 achieves a F-measure average of 0.653.

The audio analysis results contribute to scientific and engineering knowledge by demonstrating the efficacy of combining unsupervised and supervised learning techniques for species-

specific audio recognition. The integration of the acoustic animal identification method with KiwiNet's architecture has yielded a highly accurate model for detecting *Atlapetes blancae* vocalizations. This approach leverages pre-existing acoustic patterns and enhances them with supervised learning, significantly improving detection performance compared to traditional methods. Other experts in the field can use this methodology to develop and refine bioacoustic monitoring systems for various species, facilitating more precise and automated wildlife tracking and conservation efforts.

B. Image trained models

In the field of image analysis, there is currently no specific approach available to recognize *Atlapetes blancae*. Furthermore, pre-trained image-based animal identification systems do not include *Atlapetes blancae* in their datasets.

Initially, we attempted to train a ResNet-18 architecture using the whole image as input (frame), but the results were unsatisfactory, with a F-measure average of 0.473 and Accuracy of 0.467. This led us to realize the significance of giving a better context to the input image, which greatly improved the network's ability to recognize and learn the relevant patterns. Therefore we include a previous stage, which limits the input image of the network and facilitates the learning of distinctive patterns by CNN. Table II presents a comparison of the ResNet architecture [20] with different depths, including ResNet-18, ResNet-50, and ResNet-101. This table evaluates the classification performance of these models in detecting the presence or absence of *Atlapetes blancae* in cropped images obtained by the Multi-Layer RPCA algorithm [19]. The increase in the performance of all three ResNet architecture variations can be observed in Table II in comparison to initially ResNet-18 trained on frames. We selected the ResNet-18 architecture because it presents the best performance and we called the hold methodology as RPCA ResNet-18 model. We adjust the evaluation metric of the RPCA ResNet-18 model, taking into account the number of cropped images corresponding to each video to calculate the result. This analysis shows that the F-measure average decreases from 0.940 to 0.495.

TABLE II
COMPARING THE PERFORMANCE OF IMAGE MODELS ON TESTING DATA.

CNN	F-measure avg	Precision avg	Recall avg	Acc	Type of Data
RPCA ResNet-18 (ours)	0.940	0.953	0.928	0.967	Cropped images
RPCA ResNet-50 (ours)	0.937	0.954	0.921	0.966	Cropped images
RPCA ResNet-101 (ours)	0.926	0.947	0.905	0.956	Cropped images
RPCA ResNet-18 (ours)	0.495	0.512	0.882	0.889	Videos
ResNet-18 [20]	0.473	0.475	0.472	0.467	Frames

The video analysis results enhance the understanding of effective segmentation and classification methods for species detection in camera-trap footage. By utilizing the Multi-layer RPCA method for accurate image segmentation and subsequently training a ResNet-18 model on the cropped images, the study demonstrates a novel approach to isolating and recognizing *Atlapetes blancae*. This methodology's significant improvement in detection accuracy underscores its potential application in similar ecological and wildlife monitoring projects. Researchers and engineers can adopt these techniques to improve the specificity and accuracy of automated image-based species identification systems, thus advancing the capabilities of remote sensing and conservation technologies.

VI. CONCLUSION AND FUTURE WORK

In this paper, we present a methodology for recognizing *Atlapetes blancae*, an endemic bird species in a critically endangered state. Previous works have not included *Atlapetes blancae* in their list of recognized species. Furthermore, *Atlapetes blancae* identification with artificial intelligence algorithms is a big challenge due to this species having small data for training and less data labeled. In our proposal, we employ a novel semi-automatic methodology to acquire acoustic information about the target species and to label the audio registers. Additionally, we conduct a comparative analysis between an audio model and a video model, with our findings indicating that the audio model is the preferred choice for processing the data. However, this model represents just the initial step in the development of a sufficiently robust tool for *Atlapetes blancae*. For future work, we believe that integrating sensor information is crucial to the creation of more robust models, rather than relying on separate models for each sensor. Sensor information integration entails the utilization of data from various heterogeneous sources, often with asynchronous data streams, to extract more robust and informative features. By combining data from multiple sensors, we can enhance the accuracy and reliability of our recognition system for *Atlapetes blancae*. Furthermore, there is a pressing need to reduce the computational cost associated with image preprocessing, as this is essential for streamlining the image analysis process. Developing multi-modal algorithms will alleviate the computational burden and expedite image analysis. These advancements will significantly enhance the practicality and scalability of our methodology for large-scale monitoring and conservation efforts. It is important to note that while the analysis of multi-modal sequential data has gained significant traction in recent machine learning research, it has yet to address the specific domain of animal monitoring. As a result, further research and development are required to adapt these approaches to the challenges posed by animal monitoring.

ACKNOWLEDGMENT

This work was supported by Universidad de Antioquia - CODI and Alexander von Humboldt Institute for Research on Biological Resources [code project: 2020-33250].

REFERENCES

[1] M. Diaz-Vallejo *et al.*, "Use of acoustic monitoring to estimate occupancy of the antioquia brushfinch (atlapetes blancae), a critically endangered species, in san pedro de los milagros, antioquia," *Journal of Field Ornithology*, vol. 94, 2023.

[2] BirdLife International, *Atlapetes blancae*, The IUCN Red List of Threatened Species 2021: e.T22735460A181746724, https://dx.doi.org/10.2305/IUCN.UK.2021-3.RLTS.T22735460A181746724.en. Accessed on 2024.03.05., 2021.

[3] T. Donegan, "A new species of brush finch (emberizidae: Atlapetes) from the northern central andes of colombia," *Bulletin of the British Ornithologists' Club*, vol. 127, p. 255, 2007.

[4] R. Correa Peña, S. Chaparro-Herrera, A. Lopera-Salazar, and J. Parra, "Rediscovery of the antioquia brush finch atlapetes blancae. redescubrimiento del gorrión-montés paisa atlapetes blancae," *Cotinga*, vol. 41, pp. 101–108, 2019.

[5] R. T. Buxton, P. E. Lendrum, K. R. Crooks, and G. Wittemyer, "Pairing camera traps and acoustic recorders to monitor the ecological impact of human disturbance," *Global Ecology and conservation*, vol. 16, e00493, 2018.

[6] J. Xie, S. Zhao, X. Li, D. Ni, and J. Zhang, "Kd-cldnn: Lightweight automatic recognition model based on bird vocalization," *Applied Acoustics*, vol. 188, p. 108 550, 2022.

[7] D.-Y. Meng *et al.*, "A method for automatic identification and separation of wildlife images using ensemble learning," *Ecological Informatics*, vol. 77, p. 102 262, 2023.

[8] M. Zhong *et al.*, "Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling," *Applied Acoustics*, vol. 166, p. 107 375, 2020.

[9] A. Noumida and R. Rajan, "Multi-label bird species classification from audio recordings using attention framework," *Applied Acoustics*, vol. 197, p. 108 901, 2022.

[10] E. Dufourq, C. Batist, R. Foquet, and I. Durbach, "Passive acoustic monitoring of animal populations with transfer learning," *Ecological Informatics*, vol. 70, p. 101 688, 2022.

[11] H. Xiao, D. Liu, K. Chen, and M. Zhu, "Amresnet: An automatic recognition model of bird sounds in real environment," *Applied Acoustics*, vol. 201, p. 109 121, 2022.

[12] X. Han and J. Peng, "Bird sound classification based on ecoc-svm," *Applied Acoustics*, vol. 204, p. 109 245, 2023.

[13] H. Böhner, E. F. Kleiven, R. A. Ims, and E. M. Soininen, "A semi-automatic workflow to process images from small mammal camera traps," *Ecological Informatics*, p. 102 150, 2023.

[14] F. Simões, C. Bouveyron, and F. Precioso, "Deepwild: Wildlife identification, localisation and estimation on camera trap videos using deep learning," *Ecological Informatics*, vol. 75, p. 102 095, 2023.

[15] M. J. Guerrero, C. L. Bedoya, J. D. López, J. M. Daza, and C. Isaza, "Acoustic animal identification using unsupervised learning," *Methods in Ecology and Evolution*, vol. 14, 2023.

[16] T. M. Aide *et al.*, "Real-time bioacoustics monitoring and automated species identification," *PeerJ*, vol. 1, e103, 2013.

[17] C. Bedoya and L. Molles, "Acoustic censusing and individual identification of birds in the wild," *bioRxiv Preprint*, 2021.

[18] S. Beery *et al.*, "Efficient pipeline for automating species id in new camera trap projects," *Biodiversity Information Science and Standards*, vol. 3, 2019.

[19] J.-H. Giraldo-Zuluaga, A. Salazar, A. Gomez, and A. Diaz-Pulido, "Camera-trap images segmentation using multi-layer robust principal component analysis," *The Visual Computer*, vol. 35, pp. 335–347, 2019.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[21] J. D. Santamaria P, *Blancaenet: A computational tool*, https://github.com/Julian075/BlancaeNet. Accessed on 2024.06.025, 2024.

[22] K. Lisa Yang Center for Conservation Bioacoustics, *Raven Pro: Interactive Sound Analysis Software (Version 1.6.1)*, https://ravensoundsoftware.com/. Accessed on 2024.03.05., 2019.

[23] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101 236, 2021.

[24] Cornell Lab of Ornithology, *Merlin Bird ID*, https://merlin.allaboutbirds.org. Accessed on 2024.03.05., 2023.

[25] C. Zhang, Q. Li, H. Zhan, Y. Li, and X. Gao, "One-step progressive representation transfer learning for bird sound classification," *Applied Acoustics*, vol. 212, p. 109 614, 2023.

[26] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[27] A. C. Ferreira *et al.*, "Deep learning-based methods for individual recognition in small birds," *Methods in Ecology and Evolution*, vol. 11, pp. 1072–1085, 2020.

[28] J. Guo *et al.*, "Graph knows unknowns: Reformulate zero-shot learning as sample-level graph recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 7775–7783.

[29] M. Favorskaya and A. Pakhirka, "Animal species recognition in the wildlife based on muzzle and shape

features using joint cnn," *Procedia Computer Science*, vol. 159, pp. 933–942, 2019.

[30] C. Chalmers, P. Fergus, S. Wich, and A. C. Montanez, "Conservation ai: Live stream analysis for the detection of endangered species using convolutional neural networks and drone technology," *arXiv Preprint arXiv:1910.07360*, 2019.

[31] M. A. Tabak *et al.*, "Improving the accessibility and transferability of machine learning algorithms for identification of animals in camera trap images: Mlwic2," *Ecology and Evolution*, vol. 10, pp. 10 374–10 383, 2020.

[32] J. A. Ahumada *et al.*, "Wildlife insights: A platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet," *Environmental Conservation*, vol. 47, pp. 1–6, 2020.

[33] S. Leorna and T. Brinkman, "Human vs. machine: Detecting wildlife in camera trap images," *Ecological Informatics*, vol. 72, p. 101 876, 2022.

[34] J. Xie, M. Towsey, M. Zhu, J. Zhang, and P. Roe, "An intelligent system for estimating frog community calling activity and species richness," *Ecological Indicators*, vol. 82, pp. 13–22, 2017.

[35] J. Aguilar-Martin and R. L. De Mantaras, "The process of classification and learning the meaning of linguistic descriptors of concepts," *Approximate reasoning in decision analysis*, vol. 1982, pp. 165–175, 1982.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv Preprint arXiv:1409.1556*, 2014.

[37] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[38] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, pp. 541–551, 1989.

[39] J.-H. Giraldo-Zuluaga, A. Salazar, A. Gomez, and A. Diaz-Pulido, "Recognition of mammal genera on camera-trap images using multi-layer robust principal component analysis and mixture neural networks," in *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2017, pp. 53–60.

[40] N. S. Aybat, D. Goldfarb, and G. Iyengar, "Fast first-order methods for stable principal component pursuit," *arXiv preprint arXiv:1105.2126*, 2011.

[41] D. M. Powers, "Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv Preprint arXiv:2010.16061*, 2020.