UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Automated Detection and Classification of Wildlife in the Chocó Forest (Canandé) using Camera Traps

Proyecto de Titulación

Edwin David Montenegro Benavides

Felipe Grijalva, Ph.D.

Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito para la obtención del título de Magíster en Inteligencia Artificial

Quito, 02 de diciembre de 2024

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ
COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Automated Detection and Classification of Wildlife in the Chocó Forest (Canandé) using Camera Trap

Edwin David Montenegro Benavides

Nombre del Director del Programa:	Felipe Grijalva
Título académico:	Ph.D. en Ingeniería Eléctrica
Director del programa de:	Inteligencia Artificial
Nombre del Decano del colegio Académico:	Eduardo Alba
Título académico:	Doctor en Ciencias Matemáticas
Decano del Colegio:	Ciencias e Ingenierías
Nombre del Decano del Colegio de Posgrados:	Dario Niebieskikwiat
Título académico:	Doctor en Física

Quito, diciembre 2024

© DERECHOS DE AUTOR

Nombre del estudiante:	Edwin David Montenegro Benavides
Código de estudiante:	00338985
C.I.:	1723433742
Lugar y fecha:	Quito, 02 de Diciembre de 2024.

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following graduation project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

DEDICATORIA

A mi familia, a mi padre y a mi madre, quienes siempre han estado a mi lado, brindándome su apoyo incondicional y motivándome a seguir adelante ante cualquier situación que se me ha presentado; a mi hermano, por ser mi compañero de vida y mi mayor inspiración; y a mis abuelitos, por su sabiduría y cariño, que han sido pilares fundamentales en mi formación. Sin duda, el esfuerzo y la dedicación que he puesto en este trabajo son fruto de su amor y guía constante. Ustedes son los autores principales de este logro tan importante para mí. Gracias por creer en mí y acompañarme en este camino.

AGRADECIMIENTOS

Quiero expresar mi más profundo agradecimiento a mi familia, por su apoyo incondicional y constante motivación, que han sido fundamentales para mi desarrollo tanto personal como profesional, impulsándome siempre a alcanzar mis metas y a convertir en realidad los sueños que me he propuesto. Asimismo, agradezco infinitamente a la bióloga Eliana Montenegro, por su generosidad al proporcionarme los valiosos datos recolectados que hicieron posible este proyecto, y a la Fundación Jocotoco, por su confianza y apoyo en el uso legal de los datos y por brindar el espacio donde se realizó la recolección de esta información única en los ecosistemas de nuestro querido Ecuador. Sin su colaboración, este trabajo no habría sido posible.

RESUMEN

El monitoreo manual de fauna silvestre mediante cámaras trampa representa un desafío significativo por el volumen de datos y el tiempo requerido para su análisis, particularmente en regiones megadiversas como el bosque del Chocó. Para abordar esta problemática, se presenta un sistema automatizado basado en aprendizaje profundo para la detección y clasificación de seis especies objetivo (*Agutí Centroamericano*, *Ardillas*, *Armadillo de Nueve Bandas*, *Paca de Tierras Bajas*, *Roedores* y *Tinamú Grande*) en la Reserva Jocotoco Canandé del bosque del Chocó en Ecuador. El sistema incorpora un enfoque de dos etapas: detección de objetos utilizando YOLO para identificar la presencia de animales en videos de cámaras trampa (con umbral de confianza >55% para la extracción inicial de frames), seguido de la clasificación de especies utilizando arquitecturas ResNet50 y MobileNetV3. Los resultados demuestran que ResNet50 alcanza un rendimiento superior con un F1-score ponderado de 0.951 al ser entrenado con frames que superan un umbral de confianza del 60%, mientras que MobileNetV3 alcanza un F1-score de 0.946. El análisis comparativo de ambas arquitecturas sugiere diferentes escenarios de aplicación: ResNet50 destaca en aplicaciones que requieren máxima precisión en la clasificación, mientras que MobileNetV3 se presenta como una alternativa eficiente para implementaciones con recursos computacionales limitados, manteniendo un rendimiento competitivo.

Palabras clave: YOLO, deep learning, conservation, Resnet, Movilenet, Transfer learning, camera traps.

ABSTRACT

Manual wildlife monitoring using camera traps represents a significant challenge due to the volume of data and the time required for its analysis, particularly in megadiverse regions such as the Chocó forest. To address this issue, we present an automated system based on deep learning for the detection and classification of six target species (*Central Agouti*, *Squirrels*, *Nine-banded Armadillo*, *Lowland Paca*, *Rodents* and *Great Tinamú*) in the Jocotoco Canandé Reserve of the Chocó forest in Ecuador. The system incorporates a two-stage approach: object detection using YOLO to identify the presence of animals in camera trap videos (with confidence threshold >55% for initial frame extraction), followed by species classification using ResNet50 and MobileNetV3 architectures. The results show that ResNet50 achieves superior performance with a weighted F1-score of 0.951 when trained with frames exceeding a 60% confidence threshold, while MobileNetV3 achieves an F1-score of 0.946. The comparative analysis of both architectures suggests different application scenarios: ResNet50 excels in applications requiring maximum classification accuracy, while MobileNetV3 is presented as an efficient alternative for implementations with limited computational resources, while maintaining competitive performance.

Key words: YOLO, deep learning, conservation, Resnet, Mobilenet, Transfer learning, camera traps.

TABLA DE CONTENIDO

I Introduction 12
I-A Related Works . . . 13

II Material and Methods 13
II-A Dataset Description . . . 13
II-B Data Processing Pipeline . . . 13
II-B1 Video Processing . . . 13
II-B2 Frame Detection and Extraction . . . 14
II-B3 Final Dataset Distribution . . . 14
II-C Model Architectures . . . 14
II-C1 Detection Models . . . 14
II-C2 Classification Models . . . 14
II-D Training Setup . . . 14
II-D1 Preprocessing and Data Augmentation . . . 14
II-D2 Handling Class Imbalance . . . 14
II-D3 Transfer Learning Strategy . . . 15
II-D4 Hyperparameter Configuration . . . 15
II-D5 Implementation Environment . . . 15
II-E Evaluation Metrics . . . 15
II-E1 Class-specific Metrics . . . 15
II-E2 Global Metrics . . . 15
II-E3 Visual and Statistical Analysis . . . 15

III Results And Discussion 15

IV Conclusion 18

Appendix A: Examples of how frame detector works with YOLO 19

References 20

ÍNDICE DE TABLAS

I Detailed Comparison of MobileNetV3 Model by Class 16
II Detailed Comparison of ResNet50 Model by Class 16

ÍNDICE DE FIGURAS

1 Location of the Canandé Reserve in Esmeraldas province, Quinindé canton, 00°31'33.8" N, 79°12'46.9" W 12
2 Species distribution in metadata. 13
3 Frequency of captured instances per target species. 14
4 Block diagram of the proposed processing pipeline 15
5 Comparative learning curves: ResNet50 vs MobileNetV3 16
6 ROC curves for MobileNetV3 with a 60% confidence threshold. 17
7 ROC curves for MobileNetV3 with a 70% confidence threshold. 17
8 ROC curves for ResNet50 with a 60% confidence threshold. 17
9 ROC curves for ResNet50 with a 70% confidence threshold. 17
10 Confusion matrix: MobileNetV3 with a 60% confidence threshold. 18
11 Confusion matrix: MobileNetV3 with a 70% confidence threshold. 18
12 Confusion matrix: ResNet50 with a 60% confidence threshold. 18
13 Confusion matrix: ResNet50 with a 70% confidence threshold 18
14 YOLOV5 Detector for Class 0(Central American Agouti) 19
15 YOLOV5 Detector for Class 1 (Squirrels) 19
16 YOLOV8 Detector for Class 2 (Nine-banded Armadillo) 19
17 YOLOV5 Detector for Class 3 (Lowland Paca) 19
18 YOLOV8 Detector for Class 4 (Rodents) 19
19 YOLOV5 Detector for Class 5 (Great Tinamou) 19

Automated Detection and Classification of Wildlife in the Chocó Forest (Canandé) using Camera Traps

Felipe Grijalva, *Senior Member, IEEE,* Edwin Montenegro, *Member, IEEE*

***Abstract*—The manual monitoring of wildlife using camera traps presents a significant challenge due to the volume of data and the time required for its analysis, particularly in megadiverse regions such as the Chocó forest. To address this issue, an automated system based on deep learning is proposed for the detection and classification of six target species (*Central American Agouti*, *Squirrels*, *Nine-Banded Armadillo*, *Lowland Paca*, *Rodents*, and *Great Tinamou*) in the Jocotoco Canandé Reserve of the Chocó forest in Ecuador. The system incorporates a two-stage approach: object detection using YOLO to identify the presence of animals in camera trap videos (with a confidence threshold >55% for initial frame extraction), followed by species classification using ResNet50 and MobileNetV3 architectures. The results show that ResNet50 achieves superior performance with a weighted F1-score of 0.951 when trained with frames exceeding a 60% confidence threshold, while MobileNetV3 achieves an F1-score of 0.946. The comparative analysis of both architectures suggests different application scenarios: ResNet50 excels in applications requiring maximum classification accuracy, whereas MobileNetV3 emerges as an efficient alternative for implementations with limited computational resources, maintaining competitive performance.**

***Index Terms*—YOLO, deep learning, conservation, Resnet, Mobilenet, Transfer learning, camera traps.**

I. Introduction

The conservation and management of wildlife communities require precise and extensive data on population status to minimize human-wildlife conflicts and ensure long-term survival. Camera traps have become essential tools in biodiversity monitoring, enabling continuous data collection in remote areas. However, traditional methods, such as tracking footprints or collecting hunting data, remain inefficient and difficult to scale, limiting their applicability in large-scale wildlife studies [1].

Despite their utility, camera traps generate vast amounts of data that require manual processing, creating a significant bottleneck for research and conservation decision-making [2]. This challenge often restricts the scope of studies, reducing sampling intensity and limiting the geographic and temporal extent of wildlife monitoring efforts [1].

The Canandé Reserve, managed by the Jocotoco Foundation, is located in the Ecuadorian Chocó, one of the most biodiverse regions on the planet and a priority for conservation. Home to over 400 bird species and iconic mammals, this ecosystem faces severe threats from deforestation, mining, and agribusiness, which have drastically reduced forest cover and fragmented habitats [3], [4]. Addressing these challenges requires efficient tools for species identification and monitoring to inform conservation policies.

Recent advances in deep and artificial learning offer a powerful solution to this bottleneck, enabling fast and accurate processing of camera trap data. Norouzzadeh [5] demonstrated their ability to process millions of images, while Willi [6] showed these techniques could achieve species identification accuracy comparable to human experts. These innovations enable large-scale studies and provide actionable tools for ecologists.

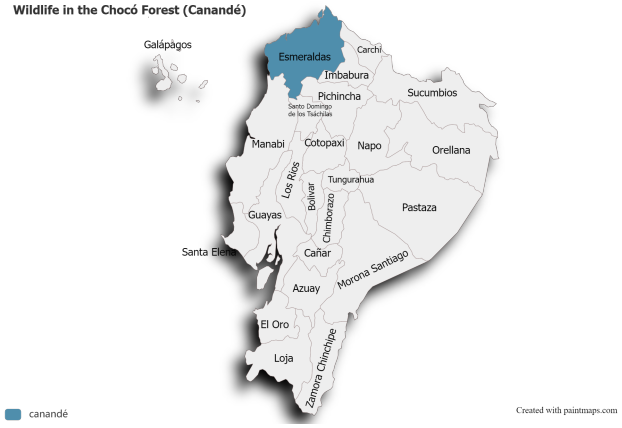


Figure 1: Location of the Canandé Reserve in Esmeraldas province, Quinindé canton, 00°31'33.8" N, 79°12'46.9" W

In this context, the Jocotoco Foundation has developed initiatives to collect wildlife data through camera traps strategically installed in the Canandé Reserve. While these cameras enable non-invasive video capture of local species, the volume of data generated exceeds the capacity for manual analysis, limiting the speed and effectiveness of conservation decisions.

Therefore, this project focuses on automating species identification and monitoring through advanced artificial intelligence technologies, implementing detection and classification models that not only significantly reduce data analysis time but also provide tools for informed decision-making in protecting the Chocó ecosystems. This

F. Grijalva and E. Montenegro are with Universidad San Francisco of Quito USFQ

approach facilitates the identification of emerging threats and supports the implementation of more effective and targeted conservation policies.

A. Related Works

Recent advancements in deep learning have enabled the development of automated systems for wildlife detection and classification. Among these, various versions of the YOLO model have been extensively used for real-time animal detection. For example, in [7], YOLOv8 was applied to mitigate conflicts between humans and wildlife near forests. Three model variations (YOLOv8m, YOLOv8l, and YOLOv8x) were trained on a dataset of 1,619 annotated images of lions, tigers, leopards, and bears. The results demonstrated that YOLOv8x achieved the highest performance, with a mean Average Precision (mAP) of 94.3%, proving its effectiveness in challenging environments.

Similarly, YOLOv5 has been utilized for wildlife detection and alert systems. In [8], real-time videos captured with sensors and drones were processed using YOLOv5, which successfully identified animals and issued alerts to improve the safety of communities near wildlife reserves. This implementation highlighted YOLOv5's efficiency in areas with high wildlife activity.

Beyond object detection, convolutional neural networks (CNNs) have also been employed for species classification. In Colombia, AlexNet was used to create a classification system for endangered animals, achieving a validation accuracy of 97.52%. This demonstrates the model's capacity for species identification and monitoring in conservation efforts [9]..

Another relevant study was conducted at the Tiputini Biodiversity Station (TBS), where YOLOv5 and Faster R-CNN were applied for detecting and classifying peccaries (Tayassu pecari and Dicotyles tajacu). Using a dataset of 7,733 images, YOLOv5 outperformed Faster R-CNN, achieving higher mAP and lower loss rates. These results underscore YOLOv5's robustness for species monitoring in Amazonian environments [10].

In larger-scale implementations, Tabak [11] employed ResNet-18 to process over 3 million camera trap images in Yellowstone National Park. The system achieved 98% accuracy across 28 species, showcasing the scalability and reliability of deep learning models for biodiversity research.

Building on these advances, our work introduces an automated monitoring and classification system tailored to the wildlife of the Ecuadorian Chocó. Utilizing camera trap videos collected in the Canandé Reserve from 2020 to 2021, the proposed system employs a dual-stage approach: YOLOv5l and YOLOv8x are used for frame detection in low-resolution videos, and ResNet50 and MobileNetV3 architectures are applied for precise species classification. Unlike previous studies, our approach accounts for the computational constraints typical of nature reserves, offering adaptable solutions for varying technological resources and environmental conditions. By addressing these challenges, our system enhances conservation strategies and contributes to the long-term protection of the Chocó's unique biodiversity.

II. Material and Methods

A. Dataset Description

This study was conducted using data collected in the Jocotoco Canandé Reserve, located in the Chocó rainforest of Ecuador, a region of great ecological importance and biodiversity. The initial dataset comprises 2,347 wildlife videos captured through camera traps, covering 51 species of mammals and birds. Based on the frequency of videos in the metadata and considering computational limitations, we focused on the six most frequently captured species, which represent approximately 35% of the total dataset and serve as representative target species of the ecosystem:

- Central American Agouti (*Dasyprocta punctata*)
- Squirrels (*Family Sciuridae*)
- Nine-banded Armadillo (*Dasypus novemcinctus*)
- Lowland Paca (*Cuniculus paca*)
- Rodents
- Greater Tinamou (*Tinamus major*)

The videos were captured with a standard resolution of 640x368 pixels at 30 frames per second (fps), with durations ranging from 20-30 seconds. The diversity in capture conditions significantly enriches the dataset, including variations in lighting from daylight to nighttime captures, different weather conditions such as rain, cloudiness, and direct sunlight, as well as varying capture angles, including frontal, lateral, and oblique views, and different distances between the camera and the subject.

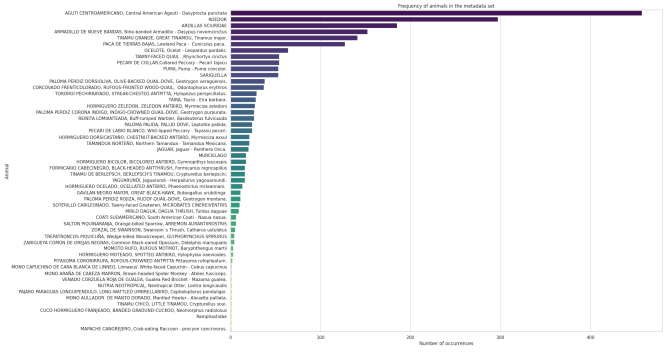


Figure 2: Species distribution in metadata.

B. Data Processing Pipeline

1) Video Processing: Data processing was structured through a robust pipeline designed to maximize the quality and traceability of the analysis. The initial preparation phase involved the careful selection of 130 videos per species, implementing a unique SHA-256 hash system and an ID to ensure data integrity and facilitate tracking throughout the analysis process. The dataset was divided using Scikit-Learn, implementing stratification to ensure

the representativeness of each species: 70% for training, providing a solid foundation for model learning; 15% for validation, allowing fine-tuning of hyperparameters; and 15% for testing, for final performance evaluation.

2) Frame Detection and Extraction: Frame extraction was executed using a dual detection approach, employing two complementary architectures. YOLOv5l was configured with a 55% confidence threshold for general detection, while YOLOv8x was used in cases requiring higher precision due to challenging conditions or complex species patterns. The extraction process was optimized by limiting the analysis to 350 frames per video, where each detection generated specific bounding boxes that frame the animal in the scene. For each successful detection, the precise coordinates of the bounding box, along with their respective confidence scores, were stored, creating a detailed record that allows for traceability and subsequent evaluation of detection quality.Each detection generated bounding boxes and confidence scores, which were stored for traceability. To ensure the quality of the extracted frames, visual inspections were conducted to confirm the presence and clarity of the target species before proceeding to classification. Although manual, this process ensured the reliability of the dataset for subsequent stages.

3) Final Dataset Distribution: The extraction process resulted in a specific distribution per species:

- ID 0 (Central American Agouti): 10,227 frames for training, 1,069 for validation, and 2,275 for testing.
- ID 1 (Squirrel): 2,200 frames for training, 327 for validation, and 451 for testing.
- ID 2 (Nine-banded Armadillo): 4,800 frames for training, 970 for validation, and 509 for testing.
- ID 3 (Lowland Paca): 1,690 frames for training, 419 for validation, and 129 for testing.
- ID 4 (Rodent): 586 frames for training, 86 for validation, and 708 for testing.
- ID 5 (Greater Tinamou): 7,955 frames for training, 1,985 for validation, and 1,938 for testing.

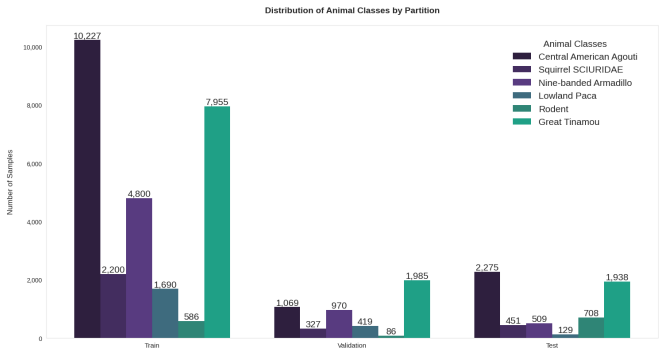


Figure 3: Frequency of captured instances per target species.

C. Model Architectures

1) Detection Models: In this study, two architectures from the YOLO family were implemented, chosen for their efficiency and accuracy in real-time object detection. YOLOv5 was used as the primary detector, taking advantage of its pretraining on the COCO dataset, which includes 200,000 images and 80 categories, providing a solid foundation for wildlife detection [12]. On the other hand, YOLOv8 represents a significant evolution in the YOLO architecture, incorporating substantial improvements in its optimization techniques and feature extraction capabilities. This newer version was specifically used in situations that required superior precision in detection, particularly in challenging lighting conditions or when species presented complex camouflage patterns [13].

2) Classification Models: For the classification task, two complementary architectures were selected:

MobileNetV3 and ResNet50. MobileNetV3 is optimized for resource-constrained environments and incorporates innovations like "squeeze-and-excitation" attention blocks. ResNet50, a deeper architecture with 50 layers and residual connections, efficiently handles the vanishing gradient problem, making it ideal for extracting complex features. Both models were pretrained on ImageNet to leverage transfer learning for the target species[14], [15].

D. Training Setup

1) Preprocessing and Data Augmentation: Image preprocessing was implemented using PyTorch Lightning, establishing a modular and organized structure. The extracted images were normalized to standardize pixel values and cropped according to the bounding box dimensions provided by the detectors. To increase the robustness of the model and prevent overfitting, data augmentation techniques were applied, including RandomCrop for framing variations, RandomHorizontalFlip for horizontal symmetry, RandomRotation for orientation variations, and ColorJitter for modifications in brightness and contrast, thereby enriching the variability of the training dataset.

2) Handling Class Imbalance: To address the inherent class imbalance in the dataset, an adaptive weighting strategy was implemented in the cross-entropy loss function. Class weights were calculated using the following formula from Scikit-Learn:

$$w_i = \frac{n_{samples}}{n_{classes} \times n_{samples_i}} \tag{1}$$

where w_i represents the weight assigned to class i, $n_{samples}$ is the total number of samples, $n_{classes}$ is the total number of classes, and $n_{samples_i}$ is the number of samples in class i.

This formulation ensures that minority classes receive a proportionally higher penalty during training, compensating for their lower representation in the dataset. The implementation was carried out using the PyTorch CrossEntropyLoss function, configured with the weight parameter that accepts a tensor of class weights. This approach is particularly effective in our scenario, where species such as

Rodents are significantly underrepresented compared to the Central American Agouti. Adaptive weighting ensures that the gradient during training is adjusted proportionally, facilitating a more balanced learning of discriminative features for each species [16], [17].
3) Transfer Learning Strategy: The implementation of transfer learning was optimized specifically for each architecture, leveraging their distinct structural characteristics [18]. For MobileNetV3, a selective freezing strategy was adopted, where the first 10 convolutional layers were kept unchanged, preserving the low-level features learned during pretraining on ImageNet (basic shapes, edges, and textures). The upper layers were kept trainable to adapt the model to the specific features of local wildlife, allowing for an optimal balance between transferred knowledge and the necessary specialization.

For ResNet50, the strategy was based on the hierarchical nature of its residual architecture. All layers were frozen except for the fourth residual block and the final classification layer, based on the assumption that high-level features are more specific to the target task.
4) Hyperparameter Configuration: The hyperparameters were carefully selected to optimize the training process:

- Learning rate of 0.0001, chosen to allow stable learning and avoid divergence.
- Batch size of 32, balancing computational efficiency and training stability.
- Maximum of 100 epochs with early stopping implemented at 10 epochs without improvement, preventing overfitting.
- Confidence Threshold of 60% and 70% for filtering frames during classification, ensuring the quality of predictions.

The selection of the **Confidence Threshold** plays a crucial role in the processing pipeline. This hyperparameter determines which frames are considered for training and evaluation of the classification model, ensuring that only detections with high confidence contribute to the learning process.
5) Implementation Environment: The system was implemented using PyTorch 1.10.12 as the main deep learning framework. The experiments were conducted on a DGX workstation equipped with 256GB of RAM and 128 GPUs, enabling efficient model training and parallel processing of large volumes of data. The complete source code of the project, including training scripts, analysis notebooks, and detailed documentation, is available in the public repository: GitHub.

E. Evaluation Metrics

1) Class-specific Metrics: The detailed evaluation by species was performed using a set of complementary metrics that include precision, to assess the proportion of correct positive predictions made by the model; recall, which quantifies the model's ability to correctly identify all existing positive cases; and F1 score, which provides a balanced measure by calculating the harmonic mean between precision and recall. Additionally, support was recorded for each class, indicating the number of instances available in the evaluation set.
2) Global Metrics: To assess the overall model performance, metrics were implemented to provide a comprehensive view of the system. Global accuracy quantifies the total proportion of correct predictions over the total predictions made, while the macro average calculates the unweighted average of individual class metrics, offering a fair evaluation independent of class imbalance. The weighted average, on the other hand, weights the metrics based on the number of instances in each class, offering a more representative perspective of the actual system performance.
3) Visual and Statistical Analysis: For a deeper evaluation, the following were generated:

- Confusion matrices: A detailed visualization of the distribution of correct and incorrect predictions.
- ROC-AUC curves: Evaluation of the model's discriminative ability at different decision thresholds.
- Learning curves: Monitoring the model's behavior during training, visualizing the evolution of loss and accuracy metrics.

These metrics were calculated for both the training and validation sets, allowing a complete evaluation of the model's behavior and early detection of issues like overfitting or underfitting.

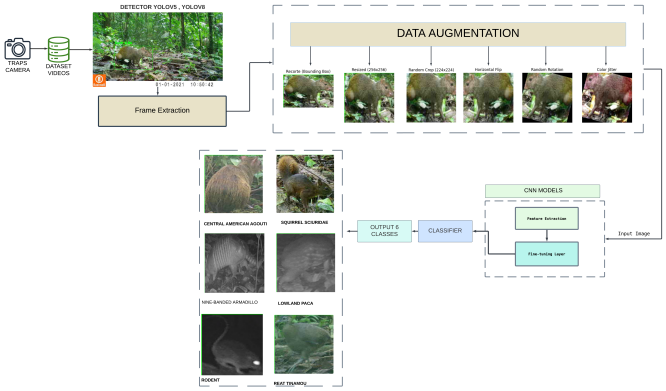


Figure 4: Block diagram of the proposed processing pipeline

III. Results And Discussion

The results of this project are divided into two parts: the first focuses on the efficiency of the object detector, and the second centers on the multiclass classification of different species.

In the detector analysis, YOLOv5l proved to be a robust tool for general frame extraction, efficiently processing most species under favorable lighting conditions with a confidence threshold of 55%. However, for the Nine-Banded

Armadillo and Rodents, the implementation of YOLOv8x, a more advanced architecture, was required. These species were predominantly captured under nighttime conditions and exhibited complex camouflage patterns that made them difficult to distinguish from their surroundings, particularly in the presence of dense vegetation and low light. This dual detection strategy maximized the quality of extracted frames for all species, ensuring a solid foundation for the subsequent classification phase.

Table I: Detailed Comparison of MobileNetV3 Model by Class

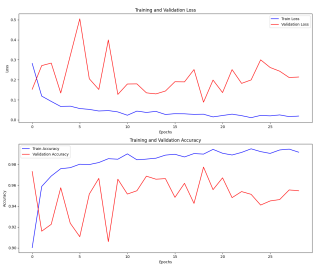
Modelo	Conf.	Clase	Precision	Recall	F1-score	Support
MobileNetV3	60%	0	0.991	0.955	0.973	2012.0
		1	0.550	0.750	0.635	44.0
		2	0.964	0.930	0.947	316.0
		3	0.516	0.926	0.663	121.0
		4	0.912	0.989	0.949	378.0
		5	0.958	0.926	0.942	1834.0
		accuracy	-	-	0.942	4705.0
		macro avg	0.815	0.913	0.851	4705.0
		weighted avg	0.954	0.942	0.946	4705.0
MobileNetV3	70%	0	0.989	0.964	0.984	1555.0
		1	0.759	1.000	0.863	22.0
		2	0.875	0.980	0.925	100.0
		3	0.484	0.859	0.619	71.0
		4	0.7	0.988	0.954	166.0
		5	0.938	0.938	0.956	1459.0
		accuracy	-	-	0.940	3373.0
		macro avg	0.791	0.816	0.751	3373.0
		weighted avg	0.937	0.940	0.925	3373.0

Table II: Detailed Comparison of ResNet50 Model by Class

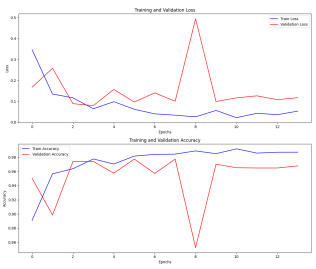
Modelo	Conf.	Clase	Precision	Recall	F1-score	Support
ResNet50	60%	0	0.989	0.990	0.989	2012.0
		1	0.667	0.818	0.735	44.0
		2	0.695	0.975	0.812	316.0
		3	0.929	0.860	0.893	121.0
		4	0.899	0.704	0.789	378.0
		5	0.989	0.963	0.976	1834.0
		accuracy	-	-	0.950	4705.0
		macro avg	0.861	0.885	0.866	4705.0
		weighted avg	0.957	0.950	0.951	4705.0
ResNet50	70%	0	0.988	0.976	0.982	1555.0
		1	0.647	1.00	0.786	22.0
		2	0.375	0.980	0.543	100.0
		3	0.438	0.789	0.563	71.0
		4	0.636	0.042	0.079	166.0
		5	0.975	0.938	0.956	1459.0
		accuracy	-	-	0.910	3373.0
		macro avg	0.677	0.787	0.651	3373.0
		weighted avg	0.923	0.910	0.903	3373.0

The comparative analysis of MobileNetV3 and ResNet50 classification models was conducted under two scenarios: one where frames with confidence above 60% were processed, and another more restrictive scenario with frames exceeding 70% confidence. ResNet50 demonstrated marginal but statistically significant superiority, achieving a weighted F1-score of 0.951 with frames above 60% confidence, compared to MobileNetV3's 0.946. The implementation of weighted cross-entropy loss to address class imbalance showed differential effectiveness: ResNet50 maintained remarkable stability in its performance across the class spectrum (macro avg: 0.866), while MobileNetV3 exhibited a more pronounced degradation in minority classes (macro avg: 0.851), suggesting a more limited generalization capacity in its shallower architecture.

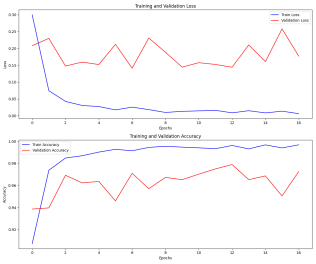
Processing frames with confidence above 70% revealed a significant divergence in model robustness: ResNet50 experienced a sharper drop in its macro average F1-score (from 0.866 to 0.651) compared to MobileNetV3 (from 0.851 to 0.751), although ResNet50 retained superior accuracy in



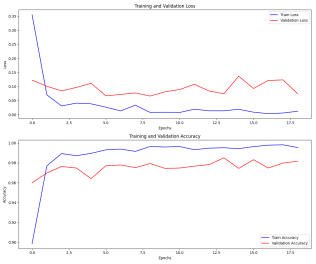
(a) ResNet50 learning curve with 60% confidence threshold



(b) ResNet50 learning curve with 70% confidence threshold



(c) MobileNetV3 learning curve with 60% confidence threshold



(d) MobileNetV3 learning curve with 70% confidence threshold

Figure 5: Comparative learning curves: ResNet50 vs MobileNetV3

majority classes. This behavior suggests that ResNet50's deeper architecture, with its residual connections, provides better capacity for extracting discriminative features but is more sensitive to frame selection constraints. The support metric evidenced a substantial reduction in the set of considered frames when increasing the confidence threshold, with direct implications for the practical applicability of the models in continuous monitoring scenarios.

The learning curve analysis revealed distinctive behaviors in convergence and stability patterns. ResNet50 exhibited higher volatility in its validation loss, particularly when processing frames with confidence above 60%, with pronounced peaks suggesting greater sensitivity to variability in validation data and batch composition. This can be attributed to its deeper architecture and residual connections. In contrast, MobileNetV3 demonstrated more stable and consistent convergence in both confidence scenarios, with a more uniform separation between training and validation curves. The implementation of early stopping at 10 epochs proved effective for both models, preventing significant overfitting.

The evaluation through ROC-AUC curves provided detailed insights into the discriminative behavior of each model. Both architectures exhibited excellent areas under the curve for the most represented species in the dataset, such as the Central American Agouti and the Great Tinamou, demonstrating a high ability to distinguish these species under various conditions. However, less-represented species, particularly Rodents, displayed significantly lower AUC values, highlighting the influence of class imbalance on

model performance. This disparity can be attributed to the limited availability of samples for certain species, which affects the model's ability to learn robust discriminative features. Additionally, less-represented species tend to appear under more challenging conditions, such as nighttime scenes or partial occlusion, increasing classification complexity. Given these limitations, analyzing the impact of the 70% confidence threshold revealed contrasting behaviors between the models: ResNet50 showed greater sensitivity to data restriction, while MobileNetV3 managed to maintain more stable performance across the different classes.

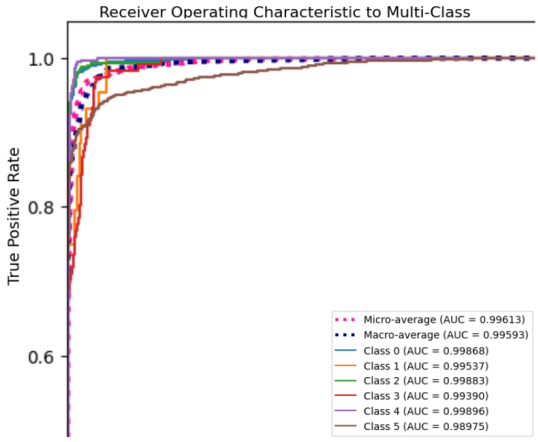


Figure 6: ROC curves for MobileNetV3 with a 60% confidence threshold.

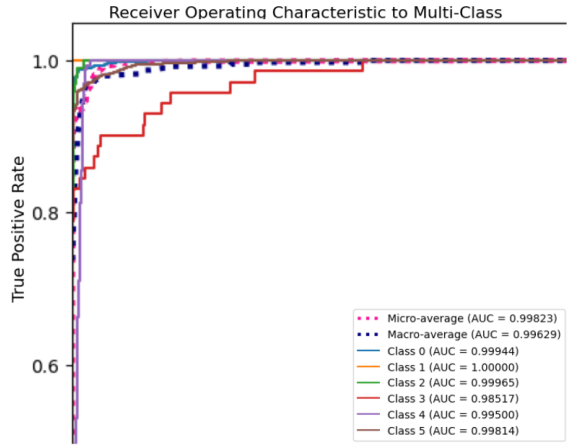


Figure 7: ROC curves for MobileNetV3 with a 70% confidence threshold.

The evaluation of confusion matrices corroborated the observed trends, where ResNet50 demonstrated superior overall discriminative capability, particularly notable in the Central American Agouti and Great Tinamou (classes 0 and 5). However, it showed more complex patterns of

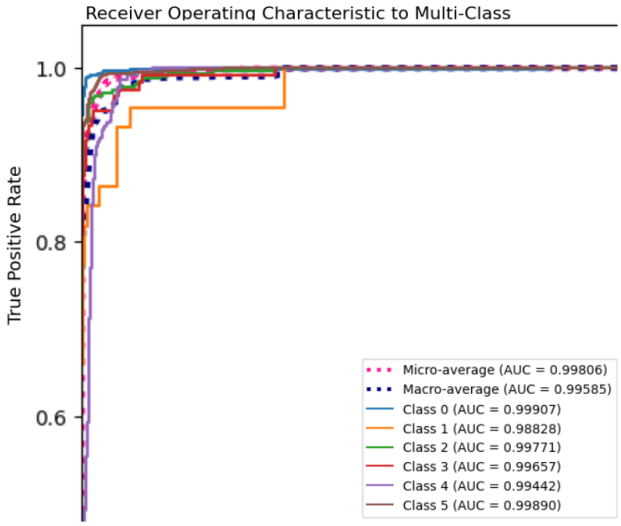


Figure 8: ROC curves for ResNet50 with a 60% confidence threshold.

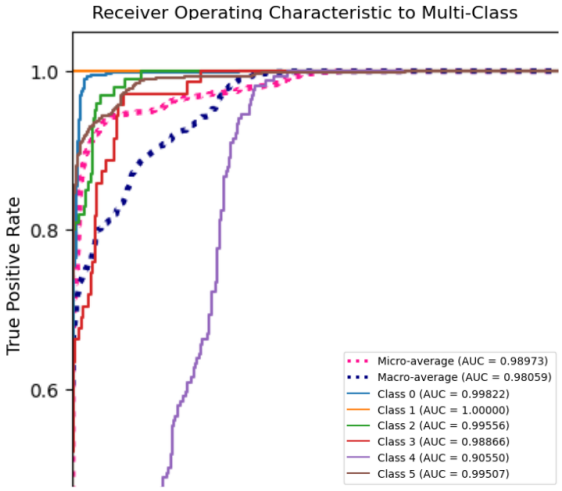


Figure 9: ROC curves for ResNet50 with a 70% confidence threshold.

confusion in intermediate species, especially between the Armadillo and Rodents (106 cases with frames above 60% confidence). When increasing the restriction to frames above 70% confidence, ResNet50 showed a significant reduction in interspecies confusion (81 cases between Armadillo and Rodents), but also a notable decrease in the total number of valid predictions. For its part, MobileNetV3, although exhibiting slightly inferior overall performance, demonstrated a more consistent error distribution and better adaptation to the increased frame confidence restriction. This was evidenced by a more uniform reduction in confusion between adjacent species and a more stable number of valid predictions across both confidence thresholds.

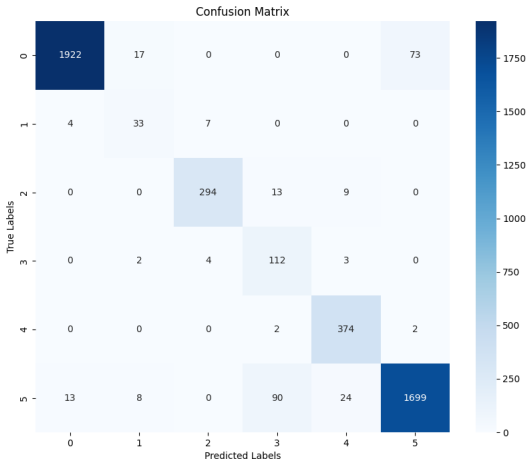


Figure 10: Confusion matrix: MobileNetV3 with a 60% confidence threshold.

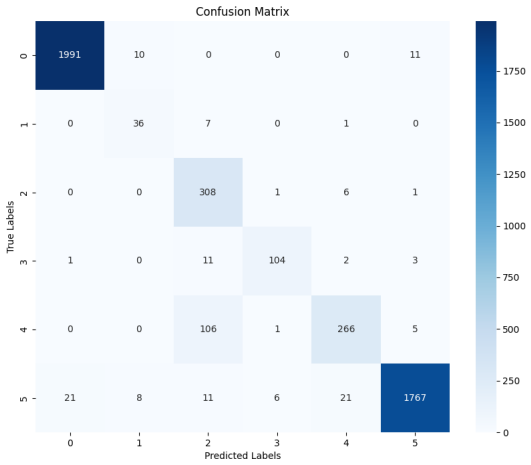


Figure 12: Confusion matrix: ResNet50 with a 60% confidence threshold.

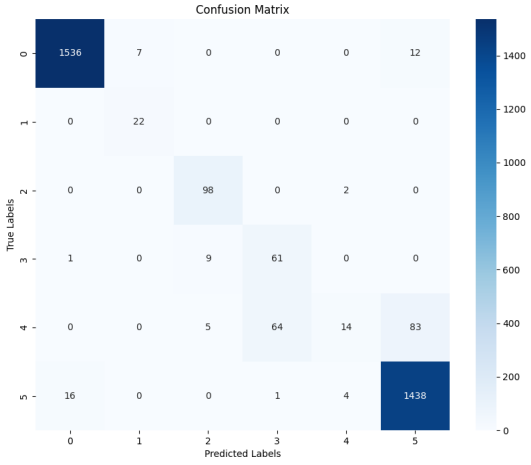


Figure 11: Confusion matrix: MobileNetV3 with a 70% confidence threshold.

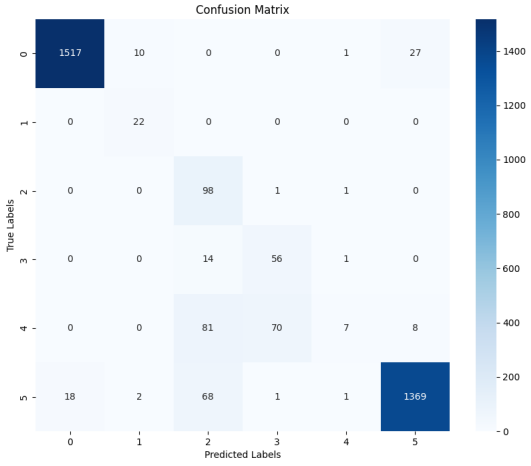


Figure 13: Confusion matrix: ResNet50 with a 70% confidence threshold

IV. Conclusion

The results demonstrate the effectiveness of the two-stage approach implemented for automated wildlife monitoring. The system efficiently processed 780 trap camera videos, generating approximately 33,000 frames with good accuracy. The YOLOv5l and YOLOv8x detectors showed complementary effectiveness, processing videos under various environmental conditions. However, the unique characteristics of images from the Chocó forest, such as extreme environmental variability and changing lighting conditions, suggest the need for specialized detectors like Megadetector, specifically optimized for wildlife scenarios. This could significantly improve detection accuracy for elusive species, particularly under nocturnal and low-visibility conditions.

The performance of the classification models revealed significant patterns: ResNet50 emerged as the superior option for applications prioritizing precision. On the other hand, MobileNetV3 demonstrated greater stability when varying the confidence threshold, making it a viable alternative for implementations with limited computational resources.

Expanding the dataset emerges as a crucial factor for future development. A larger and more diverse dataset would allow for more robust metrics, especially for underrepresented species such as rodents (586 frames) compared to the Central American agouti (10,227 frames). Optimizing the confidence threshold is presented as a critical parameter that should be adjusted according to the specific requirements of each application, considering the balance between accuracy and monitoring coverage.

The developed system not only meets the established

technical objectives but also offers a practical solution to the challenge of manual monitoring, significantly reducing video processing time from weeks to hours. This contribution sets a methodological precedent for the development of similar systems in other natural reserves in the Andean and Amazon regions, where similar challenges in wildlife monitoring are faced. The implementation of these technologies on public-use platforms has the potential to democratize biodiversity monitoring, directly benefiting biologists, researchers, and institutions dedicated to the conservation of the Chocó forest, and establishing a new standard in data-driven conservation practices.

Appendix A
Examples of how frame detector works with YOLO

Figure 14: YOLOV5 Detector for Class 0(Central American Agouti)

Figure 15: YOLOV5 Detector for Class 1 (Squirrels)

Figure 16: YOLOV8 Detector for Class 2 (Nine-banded Armadillo)

Figure 17: YOLOV5 Detector for Class 3 (Lowland Paca)

Figure 18: YOLOV8 Detector for Class 4 (Rodents)

Figure 19: YOLOV5 Detector for Class 5 (Great Tinamou)

Acknowledgment

I am deeply grateful to the Jocotoco Foundation for facilitating legal access to the data and providing access to the natural spaces of the Canandé Reserve, significantly contributing to the realization of this project. Special thanks to biologist Eliana Montenegro for her invaluable work in the systematic collection of data during the 2020-2021 period, whose dedication and field expertise were fundamental in ensuring the scientific quality and validity of this study.

References

[1] P. D. Meek, G.-A. Ballard, P. J. S. Fleming, M. Schaefer, W. Williams, and G. Falzon, "Camera traps can be heard and seen by animals," *PloS one*, vol. 9, no. 10, p. e110832, 2014.

[2] F. Trolliet, M.-C. Huynen, C. Vermeulen, and A. Hambuckers, "Use of camera traps for wildlife studies: A review," *Biotechnol. Agron. Soc. Environ.*, vol. 18, no. 3, pp. 446–454, 2014.

[3] P. Lozano, L. Roa, D. A. Neill, R. N. F. Simpson, and B. B. Klitgaard, "Flora, ecology and phytogeography of the canadé reserve, equatorial chocó," *Universidad Estatal Amazónica, Puyo, Ecuador, y Royal Botanic Gardens, Kew, Reino Unido*, 2024.

[4] V. M. N. Adriana Rodríguez Caguana1, "The protection of the chocó andino in the light of the rights of nature and the draft statute of autonomy of the metropolitan district of quito," *Unpublished Manuscript*, 2024.

[5] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. Palmer, C. Packer, and J. Clune, "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," 2017. [Online]. Available: https://arxiv.org/abs/1703.05830

[6] M. Willi, R. Pitman, A. Cardoso, C. Locke, A. Swanson, A. Boyer, M. Veldthuis, and L. Fortson, "Identifying animal species in camera trap images using deep learning and citizen science," *Methods in Ecology and Evolution*, vol. 10, 10 2018.

[7] A. B. P. G. Brahm Dave, Meet Mori, "Wild animal detection using yolov8," *Procedia Computer Science*, vol. 230, pp. 100–111, 2023.

[8] A. T. S. T. K. K. K. V. Nagagopiraju, Suvarna Pinninti, "Advanced wild animal detection and alert system using the yolo v5 model powered by ai," *Turkish Journal of Computer and Mathematics Education*, vol. 15, no. 1, pp. 142–145, 2024.

[9] S. M.-C. Andrés Felipe Rivera-Carrillo, Darwin Orlando Cardozo-Sarmiento, "Automatic classification for endangered animals in colombia using convolutional neural networks," *Mundo Fesc*, vol. 11, no. 22, pp. 95–105, 2021.

[10] N. P.-D. R. D. S. B. R. F. M. F. G. María-José Zurita, Daniel Riofrío and M. Baldeon-Calisto, "Towards automatic animal classification in wildlife environments for native species monitoring in the amazon," in *Proceedings of the IEEE*. IEEE, 2023, pp. 1–10.

[11] M. A. Tabak, M. S. Norouzzadeh, D. W. Wolfson, S. J. Sweeney, K. C. Vercauteren, N. P. Snow, J. M. Halseth, P. A. Di Salvo, J. S. Lewis, M. D. White *et al.*, "Machine learning to classify animal species in camera trap images: Applications in ecology," *Methods in Ecology and Evolution*, vol. 10, no. 4, pp. 585–590, 2019.

[12] G. Jocher, "YOLOv5 by Ultralytics," 2020, accessed: 2024-11-01. [Online]. Available: https://github.com/ultralytics/yolov5

[13] ——, "YOLOv8 by Ultralytics," 2023, accessed: 2024-11-01. [Online]. Available: https://github.com/ultralytics/ultralytics

[14] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Mobilenetv3: Next-generation mobile neural networks for computer vision," *arXiv preprint arXiv:1905.02244*, 2019.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[17] K. Kumar, R. V. Asundi, and R. Prakash, "Class weight technique for handling class imbalance," BMS Institute of Technology and Management, Technical Report, July 2022. [Online]. Available: https://www.researchgate.net/publication/362066936_Class_Weight_technique_for_Handling_Class_Imbalance

[18] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.