UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Optimizing Semi-Supervised Models for Wildlife Classification in Tiputini Camera Trap Images

Proyecto de Titulación

William Felipe Toscano Acosta

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

Optimizing Semi-Supervised Models for Wildlife Classification in Tiputini Camera Trap Images

William Felipe Toscano Acosta

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:	William Felipe Toscano Acosta
Código de estudiante:	00338954
C.I.:	1725933673
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UNPUBLISHED DOCUMENT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 madre.

AGRADECIMIENTOS

Agradezco profundamente a mi hermana y sobrina por su compañía incondicional a lo largo de esta trayectoria, a mi padre por su confianza inquebrantable en mí, y especialmente a mi novia Madelyne, quien ha estado a mi lado en cada decisión, brindándome su apoyo constante y siendo mi mayor inspiración.

RESUMEN

El monitoreo de la vida silvestre presenta desafíos significativos en la clasificación de especies visualmente similares, especialmente cuando se dispone de datos etiquetados limitados. Este trabajo se centra en clasificar dos especies relacionadas, el pecarí de labios blancos (*Tayassu pecari*) y el pecarí de collar (*Pecari tajacu*), utilizando técnicas de aprendizaje semisupervisado. Aprovechando extractores de características de última generación—GoogLeNet, EfficientNet y Vision Transformer (ViT)—y explorando métodos como el autoentrenamiento y la propagación de etiquetas, esta investigación demuestra el potencial de los modelos semisupervisados para mejorar el desempeño de clasificación con conjuntos de datos etiquetados limitados. La validación cruzada Monte Carlo confirma la robustez de estos hallazgos, revelando que más allá del 70% de datos etiquetados, las mejoras en el desempeño disminuyen debido a los rendimientos sin mejoras representativas. El estudio también resalta el rendimiento superior de los modelos completamente supervisados con el 100% de datos etiquetados, subrayando la compensación entre el esfuerzo de etiquetado y la precisión del modelo.

Palabras clave: monitoreo de vida silvestre, aprendizaje semisupervisado, pecarí de labios blancos, pecarí de collar, Vision Transformer, propagación de etiquetas, autoentrenamiento, validación cruzada Monte Carlo, monitoreo de biodiversidad.

ABSTRACT

Wildlife monitoring poses significant challenges in the classification of visually similar species, especially with limited labeled data. This work focuses on classifying two related species, the white-lipped peccary (*Tayassu pecari*) and the collared peccary (*Pecari tajacu*), using semi-supervised learning techniques. By leveraging state-of-the-art feature extractors—GoogLeNet, EfficientNet, and Vision Transformer (ViT)—and exploring methods such as self-training and label propagation, this research demonstrates the potential of semi-supervised models in improving classification performance with limited labeled datasets. The Monte Carlo cross-validation validates the robustness of these findings, revealing that beyond 70% labeled data, performance gains diminish due to diminishing returns. The study also highlights the superior performance of fully supervised models with 100% labeled data, underscoring the trade-off between labeling effort and model accuracy.

Key words: wildlife monitoring, semi-supervised learning, white-lipped peccary, collared peccary, Vision Transformer, label propagation, self-training, Monte Carlo cross-validation, biodiversity monitoring.

TABLA DE CONTENIDO

I Introduction 12

II Prior works 12

III Materials and Methods 13
- III-A Dataset . . . 13
- III-B Feature extraction models . . . 14
- III-C Training Procedure . . . 14
- III-D Baseline Comparison . . . 14
- III-E Evaluation Metrics . . . 14
- III-F Implementation Details . . . 14
- III-G Validation and Testing . . . 14
- III-H GitHub . . . 14

IV Results and Discussion 15
- IV-A GoogLeNet . . . 15
  - IV-A1 SVM-RBF . . . 15
- IV-B EfficientNet . . . 15
  - IV-B1 SVM-RBF . . . 15
- IV-C ViT . . . 16
  - IV-C1 SVM - RBF . . . 16
  - IV-C2 Self-training . . . 16
  - IV-C3 Label Propagation . . . 16
  - IV-C4 Best Model . . . 17

V Conclusion 17

Appendix A: Confidence Interval Difference Test Calculations 18
- A-A Data and Parameters . . . 18
- A-B Standard Error Calculation . . . 18
- A-C Standard Error of the Difference . . . 19
- A-D Mean Difference and Confidence Interval . . . 19
- A-E Interpretation . . . 19

References 19

ÍNDICE DE TABLAS

I Training Set Distribution 13
II Validation Set Distribution 13
III Test Set Distribution 13
IV Classification Report for the Model on the Validation Set 15
V Classification Report for SVM-RBF Model on Validation Set 16
VI Best Configurations for Criterion = 'threshold' - ViT 17
VII Best Configurations for Criterion = 'k_best' - ViT 17
VIII Best Configurations for Label Propagation with RBF Kernel 17
IX Best Configurations for Label Propagation with KNN Kernel 18
X Accuracy Results 18
XI Precision Results 18
XII Recall Results 18
XIII F1 Score Results 19

ÍNDICE DE FIGURAS

1 White-Lipped Peccary 13
2 Collared Peccary 13
3 Comparison of White-Lipped Peccary (Taypec) and Collared Peccary (Taytaj). 13
4 GoogLeNet Feature Extraction Layers 15
5 ROC curves for Taypec and Taytaj - GoogLeNet 15
6 EfficientNet Feature Extraction Layers 15
7 ROC curves for Taypec and Taytaj - EfficientNet 15
8 Vision Transformer (ViT) Feature Extraction Layers. Each encoder block processes the input sequentially, repeating the self-attention and feed-forward layers 12 times. The CLS token captures the global representation of the image. 16
9 ROC curves for Taypec and Taytaj - ViT 16
10 Evolution of F1 Score for Self Training - ViT 16
11 Evolution of F1 Score for Label Propagation - ViT 17
12 F1 Score with Confidence Intervals Across Percentages. The shaded area represents the 95% confidence interval for the mean F1 score at different percentages of labeled data. . 17

Optimizing Semi-Supervised Models for Wildlife Classification in Tiputini Camera Trap Images

Felipe Toscano and Felipe Grijalva, *Senior Member, IEEE*

***Abstract*—Wildlife monitoring poses significant challenges in the classification of visually similar species, especially with limited labeled data. This work focuses on classifying two related species, the white-lipped peccary (*Tayassu pecari*) and the collared peccary (*Pecari tajacu*), using semi-supervised learning techniques. By leveraging state-of-the-art feature extractors—GoogLeNet, EfficientNet, and Vision Transformer (ViT)—and exploring methods such as self-training and label propagation, this research demonstrates the potential of semi-supervised models in improving classification performance with limited labeled datasets. The Monte Carlo cross-validation validates the robustness of these findings, revealing that beyond 70% labeled data, performance gains diminish due to diminishing returns. The study also highlights the superior performance of fully supervised models with 100% labeled data, underscoring the trade-off between labeling effort and model accuracy.**

***Index Terms*—wildlife monitoring, semi-supervised learning, white-lipped peccary, collared peccary, Vision Transformer, label propagation, self-training, Monte Carlo cross-validation, biodiversity monitoring.**

I. Introduction

In wildlife monitoring, the classification of animal species from camera trap images poses a significant challenge, especially when the species are visually similar and the available labeled data is limited [1]. This project focuses on classifying two closely related species: the white-lipped peccary (Taypec) and the collared peccary (Taytaj). These species are difficult to distinguish due to their similar physical characteristics, such as minor differences in snout length and neck color. The challenge is further compounded by the imbalanced nature of the dataset, which includes significantly fewer images of collared peccaries than white-lipped peccaries.

The dataset used for this project consists of three folders—Train, Validation (Val), and Test—each containing images of the animals annotated with bounding boxes. The distribution of bounding boxes across these folders is imbalanced:

- Train: 3396 white-lipped peccaries, 1816 collared peccaries
- Val: 875 white-lipped peccaries, 491 collared peccaries
- Test: 486 white-lipped peccaries, 269 collared peccaries

Felipe Toscano and Felipe Grijalva are with the Universidad San Francisco de Quito (USFQ), Ecuador. E-mails: ftoscanoa@estud.usfq.edu.ec, fgrijalva@usfq.edu.ec.

To address the data imbalance, data augmentation techniques were applied exclusively to the training set for the class with fewer images (collared peccary). These augmentations include horizontal flips and small rotations between -10 and 10 degrees, avoiding any transformations that could distort the animals' color or proportions. This decision ensures that the key distinguishing features between the two species, such as the length of the snout or neck color, are not altered, which could lead to misclassification.

Given that collecting and labeling images is resource-intensive and requires expert knowledge [2], the use of semi-supervised learning models is explored in this work. Semi-supervised learning allows the model to make use of both labeled and unlabeled data, which is crucial when labeled data is scarce. By leveraging the information from labeled images, the model can infer labels for the unlabeled images, thereby increasing the overall data available for training.

In this project, two different semi-supervised learning approaches are employed. Alongside this, three different pretrained models—GoogleNet [3], EfficientNet [4], and ViT [5]—are used as feature extractors before applying the semi-supervised models. These models were chosen for their varied architectures and strengths: GoogleNet's inception modules, EfficientNet's balance between accuracy and computational efficiency, and ViT's novel approach of processing images as sequences of patches. As Domingos highlights in "The Master Algorithm" [6], integrating diverse machine learning paradigms can effectively address complex, real-world problems such as wildlife monitoring.

The expected outcome of this research is to demonstrate that semi-supervised learning can significantly improve classification accuracy in wildlife monitoring projects, especially in cases where labeled data is limited. By automating parts of the labeling process, this approach could reduce the need for extensive manual labeling, making it easier to scale biodiversity monitoring efforts. Ultimately, this could aid conservation initiatives by providing more accurate data on species populations, particularly for species that are difficult to distinguish visually.

II. Prior works

Previous research has explored the use of supervised deep learning models for the classification of animal species in wildlife monitoring. One notable study, titled *Towards Automatic Animal Classification in Wildlife Environments for Native Species Monitoring in the Amazon* [7], focused

on using the YOLOv5 and Faster R-CNN architectures to classify white-lipped peccaries (Taypec) and collared peccaries (Taytaj) from camera trap images. This work demonstrated the effectiveness of these state-of-the-art object detection models in detecting and classifying these species with a high degree of accuracy. YOLOv5, in particular, was found to outperform Faster R-CNN in terms of mean Average Precision (mAP) and robustness, especially in challenging conditions where multiple animals appeared in the frame.

However, these approaches were fully supervised, relying entirely on labeled data for training, which limits their applicability when labeled data is scarce. Given the resource-intensive nature of manual image labeling, this study highlights a gap in existing methodologies, specifically in cases where data collection and annotation are constrained. Wildlife communities often exhibit a long-tailed distribution, with fewer images of rare species, making classification challenging due to the imbalance in data representation [8], [9]. Additionally, it is not just the rarity of some species that poses a challenge; in many wildlife datasets, a significant portion of the collected information remains unlabeled. For example, Liu *et al.* (2024) report that in their study of aquatic biodiversity monitoring, only 15% of their dataset—comprising 300,000 hours of video—was labeled [10]. Miao *et al.* (2021) addressed the labeling challenge by implementing an iterative human-in-the-loop framework, where human intervention is required only for low-confidence predictions. This approach reduced the need for human annotations by approximately 80% without sacrificing classification accuracy. By leveraging high-confidence predictions as pseudo-labels, their method efficiently updated the model while maintaining an accuracy rate of about 90% in high-confidence predictions, demonstrating the effectiveness of minimizing expert intervention.

FixMatch, a recent semi-supervised learning (SSL) method, combines pseudo-labeling (also referred to as self-training) with consistency regularization to enhance model performance with limited labeled data [11]. This approach uses pseudo-labeling by generating artificial labels from weakly-augmented unlabeled images, retaining these labels only if the model's confidence exceeds a certain threshold. These pseudo-labels are then used to train the model on strongly-augmented versions of the same images. The effectiveness of self-training in this context demonstrates that pseudo-labeling is a viable component for SSL models, particularly in scenarios with limited labeled data, as shown by FixMatch's state-of-the-art results across various benchmarks.

Building upon this foundation, our research proposes the use of semi-supervised learning techniques to overcome these limitations. By incorporating both labeled and unlabeled data, we aim to improve classification performance while reducing dependency on exhaustive manual labeling. Additionally, our work applies data augmentation techniques to address the class imbalance inherent in the dataset, focusing specifically on the collared peccary, which has fewer labeled images. This semi-supervised approach offers a scalable solution that could enhance biodiversity monitoring efforts, particularly for visually similar species where labeled data is scarce.

III. Materials and Methods

A. Dataset

The dataset used in this study was provided by the Tiputini Biodiversity Station USFQ, specifically curated for wildlife monitoring of peccary species. It includes images and bounding box annotations for two species: the white-lipped peccary (Tayassu pecari, referred to as "Taypec") and the collared peccary (Pecari tajacu, referred to as "Taytaj") [12].

Figure 1. White-Lipped Peccary

Figure 2. Collared Peccary

Figure 3. Comparison of White-Lipped Peccary (Taypec) and Collared Peccary (Taytaj).

Tables I, II, and III summarize the distribution of images and bounding boxes for each dataset subset.

Table I
Training Set Distribution

Species	Images	Bounding Boxes
Taypec	1,178	3,396
Taytaj	1,070	1,816
Total	2,248	5,212

Table II
Validation Set Distribution

Species	Images	Bounding Boxes
Taypec	336	875
Taytaj	306	491
Total	642	1,366

Table III
Test Set Distribution

Species	Images	Bounding Boxes
Taypec	175	486
Taytaj	163	269
Total	338	755

To address class imbalance, weak data augmentation techniques, including horizontal flips and small rotations

between -10 and 10 degrees, were applied exclusively to the training set for the class with fewer examples (Taytaj).

B. Feature extraction models

In this study, three feature extraction models were selected to leverage their unique architectures and the progression of innovations over time: GoogLeNet, EfficientNet, and Vision Transformer (ViT). GoogleNet, introduced in 2015 by Szegedy et al. [3], pioneered the use of inception modules, which allow the model to capture multi-scale information within the same layer. EfficientNet, developed by Tan and Le in 2019 [4], optimizes both model accuracy and computational efficiency by systematically scaling depth, width, and resolution, making it well-suited for computationally constrained environments. Finally, ViT, introduced by Dosovitskiy et al. in 2021 [5], represents a novel approach in image processing by treating images as sequences of patches, akin to words in natural language processing, which allows it to capture long-range dependencies within an image. These models were chosen for their varied architectures, time of introduction, and their suitability for handling complex image data.

C. Training Procedure

To evaluate the semi-supervised models, a baseline model was created using a Support Vector Machine with a radial basis function (RBF) kernel. The parameters for this model were optimized using a grid search across a range of values:

- **Regularization (C)**: 0.1, 1, 10, 100
- **Kernel Coefficient (gamma)**: 1, 0.1, 0.01, 0.001
- **Kernel Type**: RBF

In addition, two semi-supervised techniques were tested: **self-training** and **label propagation**.

For **self-training**, the following parameters were adjusted:

- **Labeled Data Percentages**: 10%, 30%, 50%, 70%, 90%, 100%
- **Criterion**: threshold-based or k-best selection
- **Confidence Thresholds**: 0.6, 0.75, 0.9 (relevant only if threshold criterion is selected)
- **k-best Values**: 5, 10, 20 (relevant only if k-best criterion is selected)
- **Maximum Iterations**: 20 (evaluated based on convergence warnings)

For **label propagation**, several configurations were explored based on varying the percentage of labeled data and kernel choice:

- **Labeled Data Percentages**: 10%, 30%, 50%, 70%, 90%, 100%
- **Kernel Types**: RBF and k-nearest neighbors (kNN)
- **Gamma**: 0.0001, 0.001, 0.01, 0.1, 1, 10 (relevant only if RBF Kernel is selected)
- **Neighbors**: 3, 5, 7, 9, 11, 13 (relevant only if kNN Kernel is selected)
- **Maximum Iterations**: 5000
- **Tolerance for Convergence**: 0.001

This structured parameter setup provided a robust basis for comparing the performance of semi-supervised models against the fully supervised SVM baseline.

D. Baseline Comparison

To evaluate the effectiveness of the semi-supervised learning models, a Support Vector Machine (SVM) with a radial basis function (RBF) kernel was used as the baseline. The SVM-RBF model served as a fully supervised comparison, allowing for a direct assessment of the performance improvements gained through semi-supervised techniques. This baseline model provided a controlled standard to gauge the advantages of leveraging unlabeled data in scenarios with limited labeled samples.

E. Evaluation Metrics

To assess model performance, multiple metrics were considered, with all metrics calculated using the macro averaging method. This approach gives equal weight to each class, making it particularly useful for datasets with class imbalances. The primary metrics used include:

- **Accuracy**: Measures the overall proportion of correct predictions.
- **Precision**: Assesses the proportion of true positives out of all positive predictions.
- **Recall**: Measures the proportion of true positives identified out of the actual positives.
- **F1 Score**: The harmonic mean of precision and recall, emphasizing balanced performance. The F1 score was used as the main criterion for model selection.

F. Implementation Details

All experiments were implemented using Python 3.10.12, with PyTorch 2.4.1 and PyTorch Lightning 2.3.3 as the primary deep learning frameworks. Training and evaluation were conducted on a system equipped with two NVIDIA A100 GPUs, each with 80GB of memory, running CUDA version 12.6.

G. Validation and Testing

The evaluation process began with validating all models, including the baseline SVM-RBF and the semi-supervised models, on the validation set. This initial validation step was used to fine-tune hyperparameters and assess preliminary model performance. Finally, the test set was used to evaluate the best-performing model, ensuring an unbiased assessment of their final performance. To ensure robustness in the results, a Monte Carlo cross-validation with 10 iterations and a confidence level of 95% was conducted, providing a statistical basis for comparison.

H. GitHub

https://github.com/FelipeT03/MMIA

IV. Results and Discussion

A. GoogLeNet

GoogLeNet, a convolutional neural network architecture known for its innovative inception modules, has been used for its ability to perform multi-scale feature extraction [3]. Specifically, the network was truncated before its classification layers, retaining only the feature extraction components. The output of the feature extractor was a feature vector of size 50 176, representing rich spatial and semantic features extracted from the input images. These feature vectors served as inputs for the subsequent classification models.

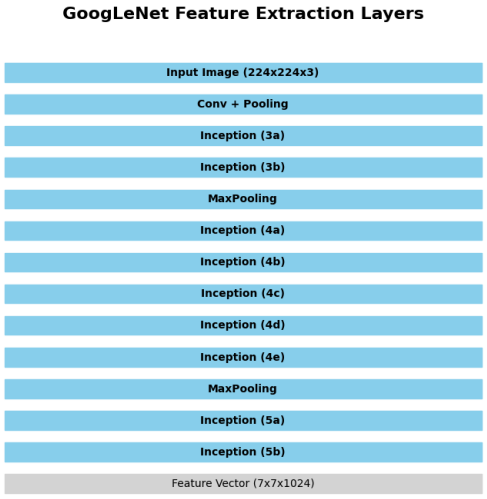

Figure 4. GoogLeNet Feature Extraction Layers

1) SVM-RBF: After performing a grid search, the optimal parameters for the SVM model with an RBF kernel were found to be `C = 10` and `gamma = 0.001`. Using these parameters, The model was evaluated on the validation set, yielding the results shown in Figure 5.

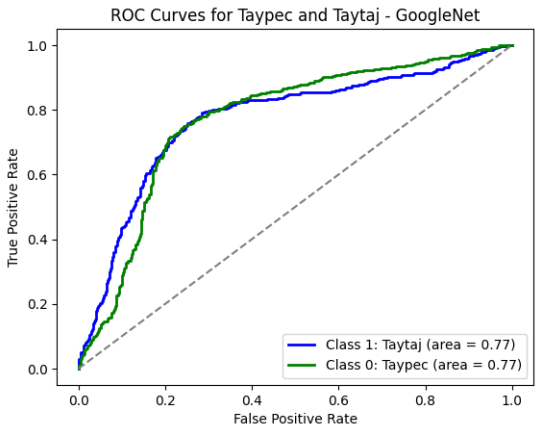

Figure 5. ROC curves for Taypec and Taytaj - GoogLeNet

B. EfficientNet

EfficientNet, a convolutional neural network architecture known for its compound scaling approach, has been utilized for its ability to balance depth, width, and resolution effectively [4]. Specifically, the network was truncated before its classification layers, retaining only the feature extraction components. The output of the feature extractor was a feature vector of size 62 720, representing rich spatial and semantic features extracted from the input images. These feature vectors served as inputs for the subsequent classification models.

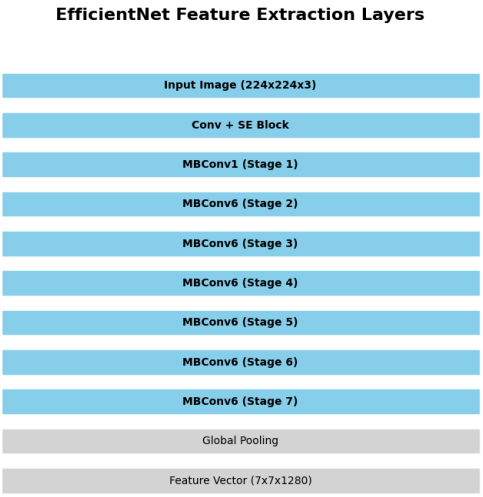

Figure 6. EfficientNet Feature Extraction Layers

1) SVM-RBF: After performing a grid search, the optimal parameters for the SVM model with an RBF kernel were found to be `C = 0.1` and `gamma = 1`. Using these parameters, the model was evaluated on the validation set, yielding the results shown in Figure 7 and Table IV.

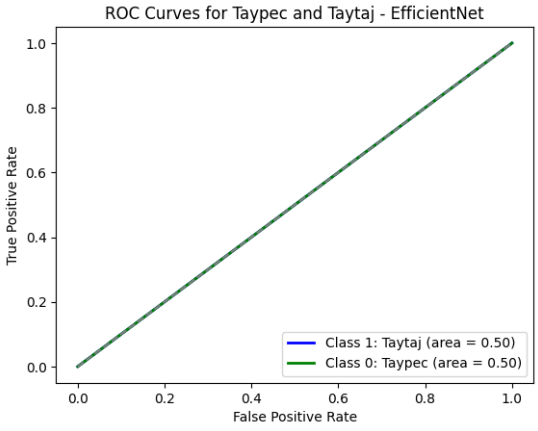

Figure 7. ROC curves for Taypec and Taytaj - EfficientNet

Table IV
Classification Report for the Model on the Validation Set

Class	Precision	Recall	F1-Score	Support
0	0.64	1.00	0.78	875
1	0.00	0.00	0.00	491
Accuracy	0.64			
Macro Avg	0.32	0.50	0.39	1366
Weighted Avg	0.41	0.64	0.50	1366

The high-dimensional output of the feature extractor (62,720 dimensions per image) likely contributes to severe overfitting, as evidenced by the F1 score of 1 achieved on the training set. This behavior results from the curse of dimensionality, where the sparsity and complexity of the feature space hinder the model's ability to generalize effectively. In such spaces, data becomes sparse, and distance metrics lose effectiveness, which poses a significant challenge for methods like SVM with an RBF kernel [13], [14]. This is compounded by the relatively small dataset, with only 6 792 training samples. The imbalance between the feature dimensions and the dataset size makes it difficult for the model to generalize effectively.

To address this, dimensionality reduction techniques like Principal Component Analysis (PCA) can be applied. PCA projects the high-dimensional features into a lower-dimensional space, retaining most of the variance while discarding noise and redundant information. For instance, applying PCA to retain 95% of the variance could reduce the 62,720-dimensional feature vectors to a manageable size, improving computational efficiency and reducing overfitting. Incorporating PCA into the preprocessing pipeline could enhance the generalization of the SVM model and improve validation performance.

C. ViT

The Vision Transformer (ViT) feature extractor utilizes the embeddings from each patch and the embedding of the class token. For each image, the ViT model outputs a 768-dimensional vector, which was used as the input for the subsequent classification models.

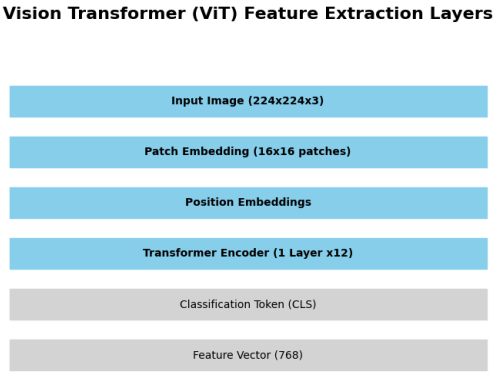

Figure 8. Vision Transformer (ViT) Feature Extraction Layers. Each encoder block processes the input sequentially, repeating the self-attention and feed-forward layers 12 times. The CLS token captures the global representation of the image.

1) SVM - RBF: After performing a grid search, the optimal parameters for the SVM model with an RBF kernel were found to be `C = 1` and `gamma = 0.001`. Using these parameters, the model was evaluated on the validation set, yielding the results shown in Figure 9.

The classification report for the validation set is shown in Table V. The model achieved an accuracy of 88%, with precision, recall, and F1-scores for each class detailed below.

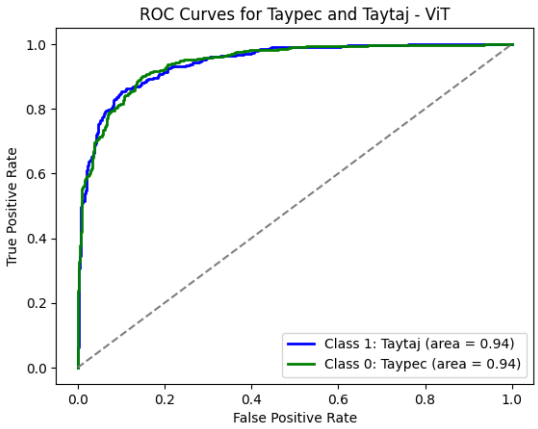

Figure 9. ROC curves for Taypec and Taytaj - ViT

Table V
Classification Report for SVM-RBF Model on Validation Set

Class	Precision	Recall	F1-Score	Support
0 (Taypec)	0.90	0.91	0.91	875
1 (Taytaj)	0.84	0.83	0.84	491
Accuracy		0.88		
Macro Avg	0.87	0.87	0.87	1366
Weighted Avg	0.88	0.88	0.88	1366

2) Self-training: After conducting a grid search, as illustrated in Figure 10, the optimal parameters for the self-training model were determined to be `criterion = threshold` and `threshold = 0.9`. Although the performance of both criteria was comparable, the threshold criterion was chosen to maintain greater control over noisy pseudo-labels. Additionally, a threshold value of 0.9 was selected as it yielded better results for lower percentages of labeled data.

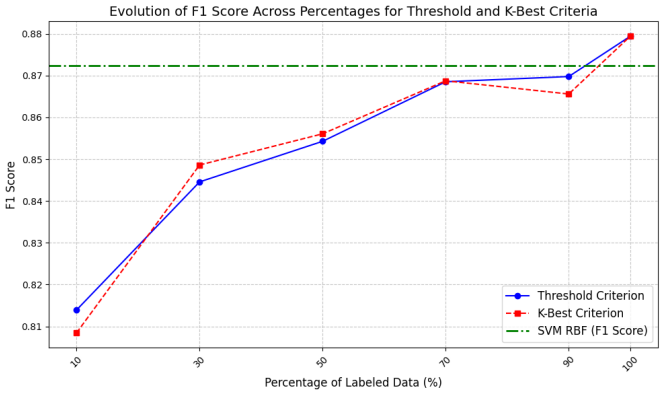

Figure 10. Evolution of F1 Score for Self Training - ViT

3) Label Propagation: A grid search was conducted to optimize the parameters for the label propagation model. The results identified `kernel = rbf` as the most suitable choice due to its superior performance at lower percentages of labeled data, with the optimal value for `gamma` determined to be 0.001.

Table VI
BEST CONFIGURATIONS FOR CRITERION = 'THRESHOLD' - VIT

Threshold	Labeled Percentage (%)	F1-Score
0.9	10.0	0.8139
0.9	30.0	0.8446
0.9	50.0	0.8543
0.9	70.0	0.8685
0.6	90.0	0.8698
0.6	100.0	0.8799

Table VII
BEST CONFIGURATIONS FOR CRITERION = 'K_BEST' - VIT

K-Best	Labeled Percentage (%)	F1-Score
5.0	10.0	0.8085
10.0	30.0	0.8486
5.0	50.0	0.8566
5.0	70.0	0.8656
5.0	90.0	0.8659
5.0	100.0	0.8799

4) Best Model: Before performing the Monte Carlo cross-validation, the Vision Transformer (ViT) was identified as the best-performing feature extractor, achieving the highest results when combined with the self-training technique. This combination is used as the basis for the Monte Carlo cross-validation, and the results for F1 score are presented in Table X and Figure 12.

The Monte Carlo cross-validation was conducted using the test validation set, performing 10 iterations for each percentage of labeled data with a confidence level of 0.95. The results demonstrated an upward trend as the percentage of labeled data used for the supervised section increased. However, beyond approximately 70% labeled data, the improvements became marginal, indicating diminishing returns. Specifically, a statistical comparison of the F1 scores for 70% and 90% labeled data using the *Confidence Interval Difference Test* showed that the difference in their means (0.001729) was not statistically significant, as the 95% confidence interval for the difference $[-0.00178, 0.00524]$ includes zero. This suggests that increasing the labeled data beyond 70% does not provide a meaningful improvement in performance.

Additionally, the model with 100% labeled data represents a fully supervised learning approach and achieved the best results among all models tested. This is expected, as supervised learning has access to all labeled data, unlike the semi-supervised models, which rely on smaller labeled datasets combined with unlabeled data (10% to 90% labeled). While the semi-supervised models demonstrate strong performance with limited labeled data, they do not surpass the fully supervised model's results, highlighting the advantage of having a completely labeled dataset in terms of classification accuracy.

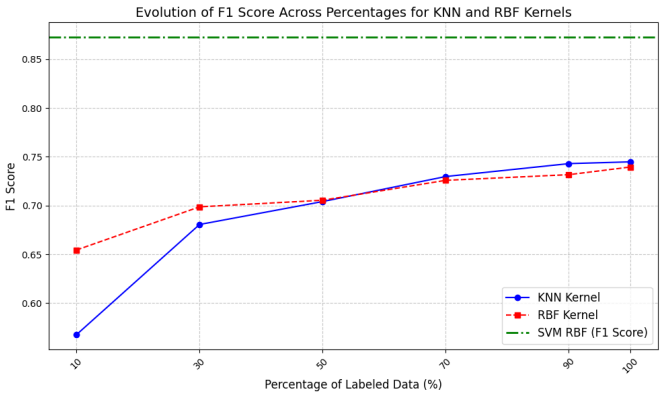

Figure 11. Evolution of F1 Score for Label Propagation - ViT

Table VIII
BEST CONFIGURATIONS FOR LABEL PROPAGATION WITH RBF KERNEL

Percentage (%)	Best Gamma	F1-Score
10%	0.1	0.6545
30%	0.1	0.6987
50%	0.1	0.7055
70%	0.1	0.7259
90%	0.1	0.7317
100%	0.1	0.7395

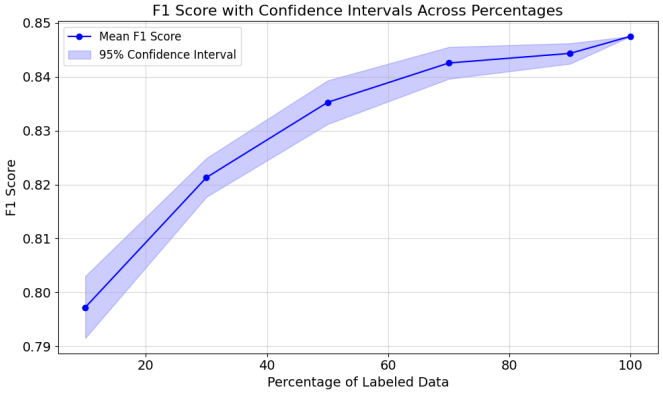

Figure 12. F1 Score with Confidence Intervals Across Percentages. The shaded area represents the 95% confidence interval for the mean F1 score at different percentages of labeled data.

V. CONCLUSION

This study presents a novel approach to wildlife monitoring by leveraging semi-supervised learning techniques to classify visually similar species with limited labeled data. Focusing on the white-lipped peccary and collared peccary, the work addresses critical challenges, including class imbalance and the resource-intensive nature of manual labeling. By incorporating state-of-the-art feature extraction models—GoogLeNet, EfficientNet, and Vision Transformer (ViT)—and exploring semi-supervised methods such as self-training and label propagation, this research provides scalable solutions for improving biodiversity monitoring efforts.

Table IX
Best Configurations for Label Propagation with KNN Kernel

Percentage (%)	Best *n_neighbors*	F1-Score
10%	3	0.5678
30%	5	0.6806
50%	5	0.7041
70%	11	0.7298
90%	13	0.7430
100%	11	0.7448

Table X
Accuracy Results

Percentage (%)	Accuracy Mean	Accuracy CI
10	0.812318	[0.8066, 0.8188]
30	0.835497	[0.8322, 0.8388]
50	0.848784	[0.8454, 0.8523]
70	0.855497	[0.8528, 0.8582]
90	0.858146	[0.8565, 0.8598]
100	0.860927	[0.8609, 0.8609]

Table XI
Precision Results

Percentage (%)	Precision Mean	Precision CI
10	0.795358	[0.7891, 0.8016]
30	0.823383	[0.8168, 0.8293]
50	0.835326	[0.8318, 0.8388]
70	0.842443	[0.8394, 0.8455]
90	0.846827	[0.8450, 0.8468]
100	0.849693	[0.8497, 0.8497]

Table XII
Recall Results

Percentage (%)	Recall Mean	Recall CI
10	0.800107	[0.7944, 0.8058]
30	0.822589	[0.8185, 0.8265]
50	0.835307	[0.8306, 0.8400]
70	0.842692	[0.8397, 0.8457]
90	0.844329	[0.8442, 0.8443]
100	0.845499	[0.8455, 0.8455]

The experiments demonstrated that semi-supervised learning techniques can significantly enhance classification performance with limited labeled data. Among the methods evaluated, the ViT feature extractor combined with self-training achieved the best results in most scenarios, illustrating the power of transformer-based models for this domain. The Monte Carlo cross-validation further validated the robustness of these findings, highlighting an upward trend in accuracy and F1 score as the percentage of labeled data increased. However, statistical analysis using the *Confidence Interval Difference Test* revealed that beyond 70% labeled data, the performance gains diminished, suggesting a point of diminishing returns for adding more labeled data.

The fully supervised model with 100% labeled data achieved the highest performance, as expected, underscoring the inherent advantage of complete labeling. Nevertheless, this study demonstrates the potential of semi-supervised learning to approximate such performance while significantly reducing the need for exhaustive manual annotations. Similar findings have been reported in aquatic biodiversity monitoring, where semi-supervised learning advanced species recognition efforts [2]. Additionally, the application of deep learning models, including semi-supervised approaches, has shown promise in detecting and classifying wildlife in aerial images, further supporting the effectiveness of these methods in biodiversity monitoring [15].

This research contributes to the field of wildlife monitoring by offering a practical framework for handling imbalanced and limited datasets. By automating parts of the labeling process and incorporating innovative models, it paves the way for more efficient and accurate species classification. Future work could focus on extending this approach to include additional species, incorporating more complex augmentation techniques, and exploring advanced semi-supervised frameworks like consistency-based regularization to further enhance performance. Additionally, dimensionality reduction techniques could be investigated to address the high dimensionality of feature vectors generated by certain extractors, such as GoogLeNet and EfficientNet, which may help mitigate overfitting and improve computational efficiency. Another promising direction involves fine-tuning the feature extractors specifically for this task, training them with a focus on distinguishing the white-lipped peccary and collared peccary to improve their ability to capture subtle species-specific features.

Appendix A
Confidence Interval Difference Test Calculations

The statistical analysis using the *Confidence Interval Difference Test* demonstrated that beyond 70% labeled data, the performance gains diminished, suggesting a point of diminishing returns. This appendix provides the detailed calculations.

A. Data and Parameters

The following data was used for the calculations:

- **Mean F1 Score for 70% labeled data:** 0.842547
- **Mean F1 Score for 90% labeled data:** 0.844276
- **Confidence Interval for 70%:** [0.8396, 0.8455]
- **Confidence Interval for 90%:** [0.8424, 0.8462]

B. Standard Error Calculation

The standard error (SE) for each group is calculated as:

$$SE = \frac{\text{Upper Bound} - \text{Lower Bound}}{2 \times 1.96}$$

- For 70%:

$$SE_{70} = \frac{0.8455 - 0.8396}{2 \times 1.96} = 0.001504$$

- For 90%:

$$SE_{90} = \frac{0.8462 - 0.8424}{2 \times 1.96} = 0.000969$$

Table XIII
F1 Score Results

Percentage (%)	F1 Mean	F1 CI
10	0.797219	[0.7915, 0.8030]
30	0.823110	[0.8177, 0.8249]
50	0.835251	[0.8312, 0.8393]
70	0.842547	[0.8396, 0.8455]
90	0.844276	[0.8424, 0.8462]
100	0.847504	[0.8475, 0.8475]

C. Standard Error of the Difference

The standard error of the difference (SE_{diff}) is calculated as:

$$SE_{\text{diff}} = \sqrt{SE_{70}^2 + SE_{90}^2}$$

$$SE_{\text{diff}} = \sqrt{(0.001504)^2 + (0.000969)^2} = 0.00178$$

D. Mean Difference and Confidence Interval

The mean difference ($\Delta\mu$) is:

$$\Delta\mu = \mu_{90} - \mu_{70} = 0.844276 - 0.842547 = 0.001729$$

The 95% confidence interval (CI_{diff}) for the difference is calculated as:

$$CI_{\text{diff}} = [\Delta\mu - 1.96 \cdot SE_{\text{diff}}, \Delta\mu + 1.96 \cdot SE_{\text{diff}}]$$

$$CI_{\text{diff}} = [0.001729 - 1.96 \cdot 0.00178, 0.001729 + 1.96 \cdot 0.00178]$$

$$CI_{\text{diff}} = [-0.00178, 0.00524]$$

E. Interpretation

Since the confidence interval for the difference includes zero ($[-0.00178, 0.00524]$), the difference in F1 scores between 70% and 90% labeled data is not statistically significant at the 95% confidence level. This suggests that increasing the labeled data beyond 70% does not result in meaningful performance improvements.

Acknowledgment

We thank the Tiputini Biodiversity Station - USFQ for providing the dataset that enabled this research.

References

[1] X. Zheng, R. J. Smith, and L. Zhao, "Self-supervised pretraining and controlled augmentation improve rare wildlife recognition in uav images," *arXiv preprint arXiv:2108.07582*, 2021.

[2] X. Ma, S. Johnson, and C. Lee, "Semi-supervised learning advances species recognition for aquatic biodiversity monitoring," *Frontiers in Marine Science*, vol. 11, p. 1373755, 2024.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[4] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2021.

[6] P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books, 2015.

[7] M.-J. Zurita, D. Riofrío *et al.*, "Towards automatic animal classification in wildlife environments for native species monitoring in the amazon," in *Proceedings of IEEE Conference*. IEEE, 2023.

[8] Z. Miao, Z. Liu, K. M. Gaynor, M. S. Palmer, S. X. Yu, and W. M. Getz, "Iterative human and automated identification of wildlife images," *arXiv preprint arXiv:2105.02320*, 2021.

[9] E. Lanzini and S. Beery, "Image-to-image translation of synthetic samples for rare classes," *arXiv preprint arXiv:2106.12212*, 2021.

[10] J. Liu, R. Smith, and L. Chen, "Semi-supervised learning advances species recognition for aquatic biodiversity monitoring," *Frontiers in Marine Science*, vol. 11, p. 1373755, 2024. [Online]. Available: https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2024.1373755/full

[11] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems*, 2020.

[12] V. Pacheco, F. A. S. Fernandez *et al.*, "Molecular ecology and conservation genetics of neotropical mammals," in *Neotropical Diversification: Patterns and Processes*, V. Rull and A. C. Carnaval, Eds. Springer, 2019, pp. 415–438. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-28868-6_17

[13] Y. Bengio, O. Delalleau, and N. Le Roux, "The curse of dimensionality for local kernel machines," *Technical Report 1258, Université de Montréal*, 2005.

[14] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *International conference on database theory*. Springer, 1999, pp. 217–235.

[15] Authors, "Deep learning models for waterfowl detection and classification in aerial images," *Information*, vol. 15, no. 3, p. 157, 2024. [Online]. Available: https://www.mdpi.com/2078-2489/15/3/157