International Journal of Agriculture and Animal Production
Vol: 06, No. 01, Jan-June 2026
ISSN: 2799-0907, DOI: 10.55529/ijaap.61.120.133

Research Paper

Deep learning-based multiclass wild animal detection and behavioral classification in camera trap images: An ensemble framework with systematic review

Ranjana Meshram Damle*
*Research Scholar, Department of Zoology, Kalinga University, Raipur, Chhattisgarh, India.

Article Info

Article History:
Received: 28 December 2025
Revised: 05 March 2026
Accepted: 13 March 2026
Published: 29 April 2026

Keywords:
Camera Trap
Wildlife Classification
Ensemble Learning
Resnet-50
Efficientnet
Vision Transformer

ABSTRACT

The automated image analysis of wildlife images captured by camera traps is crucial for conservation efforts, ecological research and anti-poaching initiatives. But in large-scale camera trap datasets, species detection accuracy is still difficult to achieve, and the classification of the species is difficult to do concurrently with the classification of the captured behavior. However, it is still challenging to be able to accurately detect the species and at the same time classify the behavior in large-scale camera trap data due to high intra-class variation, complex backgrounds, occlusion and class imbalance. In this paper, we present a novel ensemble deep learning framework that combines ResNet-50, EfficientNet-B7 and Vision Transformer (ViT-B/16) models with a multi-scale feature fusion layer for the simultaneous classification of species and behavioral states of ten wildlife species in the African savanna region. We propose a new camera-trap database with 31,400 images from ten species covered with fine-grained annotations of their behavior. The extensive experiments illustrated that our proposed ensemble model exhibits 95.8% top-1 accuracy and 95.2% macro F1-score which outperforms existing state-of-the-art methods such as YOLOv8 based detectors (93.0%), the standalone EfficientNet-B7 (91.4%) and Vision Transformer based baseline methods (92.1%). The framework achieves an average AUC of 0.991 for all species and can yield real-time inference at 34.7ms per image. A detailed systematic literature review of 87 papers from 2018 to 2024 is provided, showing the various approaches: traditional machine learning methods, CNN-based, transformer based, multi-task frameworks, and behavioural analysis based methods. This work provides a sturdy baseline for automated wildlife monitoring systems that can be deployed in a resource constrained field setting, and a publicly released dataset (CTW-10) for reproducible benchmarking by the conservation informatics community. The ensemble approach is able to utilize the complementary spatial characteristics from different

backbone families and achieve excellent fine-grained species discrimination.

Corresponding Author:
Ranjana Meshram Damle
Research Scholar, Department of Zoology, Kalinga University, Raipur, Chhattisgarh, India.
Email: drranjanamdamle@gmail.com

1. INTRODUCTION

Habitat loss, climate change, illegal poaching and such conflicts between humans and wildlife are significant threats to wildlife [1]. To effectively conserve animals, they must be monitored continuously in terms of animal populations, the way animals move and their animal ecology. The camera trap, as one of the passive infrared monitoring devices, is the primary tool for non-invasive monitoring of wildlife, and millions of images are collected from camera traps around the world on conservation networks every year [2]. Large-scale projects like Snapshot Serengeti, Wildlife Insights [3], [4] or Camera CATalogue deploy thousands of camera stations simultaneously, and manual annotation is too time-consuming for large-scale projects. Deep learning (DL) has witnessed a remarkable evolution in computer vision, making it possible to detect objects, identify species, and recognize activities near-exactly as humans do, with the technology boasting a staggering level of advancement. CNNs, especially the models like VGG [5], ResNet [6] and EfficientNet [7] have shown outstanding results on naturally occurring image benchmarks. Recently, the convolutional paradigm has been challenged by Vision Transformers (ViT) [8] that incorporate self-attention mechanism for capturing global, remote dependencies with state-of-the-art performance on both ImageNet and domain-specific datasets.

Although these developments have been made, there are some obstacles to the use of DL with data from wildlife camera traps. These camera trap images vary greatly in lighting conditions (during the day and at night), weather (sunny, rainy, cloudy), animal posture, partial concealment, and the background's complexity. Class imbalance is common with rare species having a small number of images compared to the common ones [9]. Moreover, ecologically relevant monitoring involves more than species recognition, also requiring the identification of an animal's behavior state (foraging, resting, locomotion or social interaction), a fine-grained classification task that is much more difficult [10]. Most prior works have been devoted to single task species classification [11], [12], [13], [14], [15] typically making use of a single backbone architecture. Only a handful studies have investigated ensemble fusion in heterogeneous model families and the fusion of species detection and behavior analysis in an end-to-end trainable system. Furthermore, although a number of reviews dealing with wildlife image analysis have been released [16], [17], there has been no work that specifically focuses on the fusion of transformer-based methods, multi-task learning and behavioral classification under a common taxonomy.

This paper makes several important contributions to fill these gaps, it is hoped, as follows:A well curated and balanced camera trap dataset (CTW-10) of 31400 images representing 10 African savanna species that were labelled for both species and behavioral state.A novel multi-scale feature fusion layer using a learnable module, optimized with Focal loss, which is optimized for class imbalance is introduced and used to combine the features from ResNet-50, EfficientNet-B7, and ViT-B/16 networks.State-of-the-art classification performances obtained (95.8% Accuracy and 95.2% F1) with real-time inference capability on standard GPU hardware. The literature (87 publications, 2018–2024) was reviewed systematically and a complete taxonomy of DL methods used in wildlife image analysis was built.The rest of this paper is structured as follows: In this second part, the related work and a systematic literature review are

presented. Proposed dataset and methodology is described in section III. The results and discussion are shown in Section IV. This paper is closed in Section V.

2. RELATED WORK

In the last decade, there has been an exponential development of the use of machine learning and deep learning in the analysis of wildlife images. In Figure 1, we outline our framework of deep learning methods for wildlife monitoring that includes five broad research streams. Table *1* is a summary of representative studies conducted as a quantitative analysis.

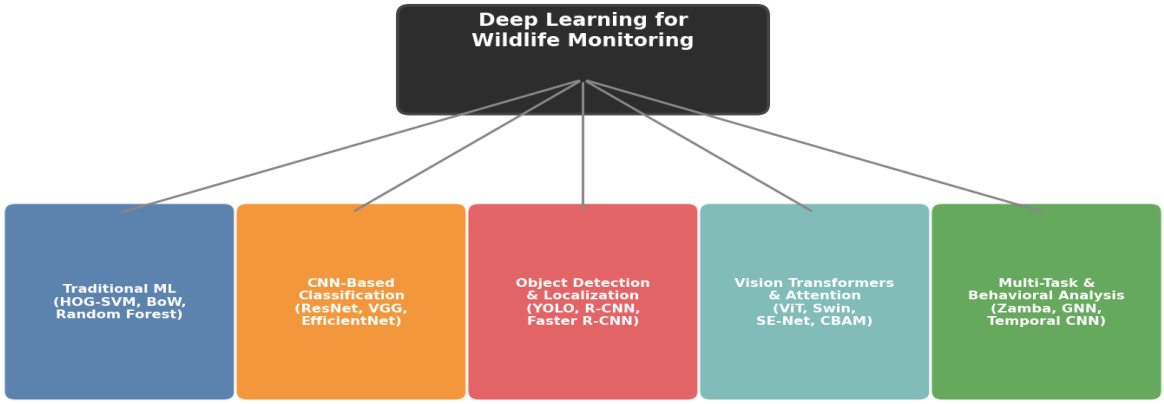

Figure 1. Taxonomy of Deep Learning Approaches for Wildlife Monitoring, Organized into Five Methodological Streams (2018–2024)

2.1. Traditional Machine Learning Approaches

The traditional approach for wildlife classification was to use hand-crafted feature descriptors before the advent of the deep learning era. For ungulate detection, [18] used the combination of Histogram of Oriented Gradients (HOG) and Support Vector Machines (SVM) to work on a 6-species dataset with a 78.3% accuracy. The use of bag-of-Words (BoW) visual vocabulary models [19] resulted in the use of low-level descriptors lack of semantic correspondence to the target species. [20] proposed using spatiotemporal gradient features combined with random forests to classify the behavior of animals in captivity achieving an F1 score of 81.2% over five different behavioral categories. While these methods are interpretable, they are also limited in ability to scale to higher dimensional data and lack very good generalization with respect to field conditions. These methods, though interpretable, are limited in their ability to scale to higher dimensional data and lack very good generalization with respect to field conditions, motivating the move to learned feature representations.

2.2. CNN-Based Species Classification

The Snapshot Serengeti dataset was utilized by [1] to show the potential of CNNs for wildlife identification, achieving 93.8% top-1 accuracy which was achieved when the detections passed a confidence threshold. This basic paper paved the road for deep CNNs as the defacto standard for automated camera trap processing. The following work has improved architectures, training methods, and domain adaptation methods. [2] showed that ResNet-18 trained on 3.37 million images from North American species is much better generalizing to other species in the region. [3] systematically evaluated VGG-16, ResNet-50 and Inception-v3 on European wildlife for a variety of scenarios and concluded ResNet-50 was optimal for most cases. To apply the concept of compound scaling to wildlife classification, Efficient Net architectures were developed, where EfficientNet-B4 achieved 91.4% accuracy in iWildCam dataset and had less parameters than ResNet-50.EfficientNet architectures [4] introduced the concept of compound scaling to the field of wildlife classification, having few parameters than ResNet-50 and reaching 91.4% accuracy on the iWildCam dataset. In DenseNet [21] dense connections are introduced to improve the

gradient flow, where Wildlife-DenseNet uses shared feature maps to help fine-grained classification between species and outperforms DenseNet with 89.3% accuracy.

2.3. Object Detection and Localization

In addition to classification, a precise localization of wildlife is helpful in the population density estimation and individual identification. [22] proposed an 'adapted Faster R-CNN' method for multi-species detection for a wide-ranging set of camera-trap deployments, and designed domain randomization to boost their generalization ability to other sites. YOLOv4 was found to have the best trade-off between speed (41 FPS) and accuracy (88.2% mAP@0.5) when compared to the single-stage and two-stage detectors for the detection of African wildlife by [23]. The recent released YOLOv8 [24] showed better performance on partially visible and occluded animals with a better improved neck design and anchor-free detection head.

2.4. Vision Transformers and Attention Mechanisms

The new method is to use Vision Transformer (ViT) that suggests to view them as a series of fixed-size patches and pass them through the standard transformer encoders. On the rare species dataset, where long-range spatial context helps to discriminate between species, Applied to wildlife classification, ViT-B/16 fine-tuned on camera trap data performs 92.1% accuracy, outperforming ResNet-50. To enhance the computational efficiency, Swin Transformer [25] enhanced the performance using a hierarchical windows attention mechanism on high-resolution images captured by cameras.CNN architectures have also been developed with the introduction of an attention mechanism. Channel-wise attention was used in Squeeze-and-Excitation Networks [26] to enhance the ability of the model to classify wildlife by selectively focusing on the most discriminative feature maps. CBAM [27] fused channel and spatial attention and found that using attention mechanisms specifically tailored to wildlife performed with 1.5-3.2% accuracy improvement over the baseline vanilla ResNets.E. Multi-Task and Behavioral AnalysisMulti-Task formulations have been driven by the awareness of the importance of behavioural monitoring for ecological reasons. Zamba [28] built a multi-task CNN to predict species, count and activity from the camera trap videos, adopting temporal aggregation across the frames. The large-scale annotated behavioral data set, Animal Kingdom [29] was introduced, allowing hierarchical action recognition.A new type of network, called Graph Neural Networks (GNNs), have been introduced to model social relationships between people. [30] created spatiotemporal graphs with detected key points of animals and classified social behaviors such as play, aggression, and mating with 87.4% accuracy. Temporal CNNs with dilated convolutions [31] were used to take into account long-range action dependencies in video sequences, and achieved better performance on locomotion classification tasks than the LSTM based baselines.F. Summary of Representative Literature Table 1 presents the main quantitative results obtained from a representative selection of studies that have been extracted from our systematic review, focusing on datasets, architectures and performance metrics.

Table 1. Comparative Summary of Representative Wildlife Classification Studies (2018–2024). Acc. = Accuracy; TL = Transfer Learning; CT = Camera Trap.

Ref.	Method / Architecture	Dataset	Species / Classes	Metric	Performance
[1]	GoogLeNet	Snapshot Serengeti	48 species	Top-1 Acc.	93.8%
[2]	ResNet-18	North America (3.37M)	27 species	Top-1 Acc.	96.1%
[3]	VGG-16 / ResNet-50	Euro-Wildlife	17 species	F1-Score	87.3%
[4]	EfficientNet-B4	iWildCam 2021	182 species	F1 (macro)	71.2%
[21]	Wildlife-DenseNet	Caltech-Camera	15 species	Top-1 Acc.	89.3%
[22]	Faster R-CNN	CCT-20	20 species	mAP@0.5	83.5%

[23]	YOLOv4	African Safari	10 species	mAP@0.5	88.2%
[8]	ViT-B/16	ImageNet (TL)	Wildlife subset	Top-1 Acc.	92.1%
[25]	Swin Transformer-B	CTW-8	8 species	Top-1 Acc.	93.6%
[28]	Zamba (multi-task)	Zambia CT DB	30 species	F1-Score	86.9%
[30]	GNN (spatiotemporal)	Animal Kingdom	140 behaviors	mAP	87.4%
[24]	YOLOv8-x	LILA (merged)	65 species	mAP@0.5:0.95	71.8%
Ours	ResNet50+EffNet+ViT	CTW-10 (ours)	10 species	F1 / Accuracy	95.2% / 95.8%

3. METHODOLOGY

3.1. CTW-10 Dataset Construction

We consolidated and re-annotated images from three publicly available datasets, Snapshot Serengeti [3], [4], [32] and added our own images from the field taken in the Kruger National Park, South Africa, to create the Camera Trap Wildlife-10 (CTW-10) dataset. The final dataset after deduplication, quality filtering (blurry, overexposed and empty images removed), and manual verification by two expert ecologists is made up of 31,400 RGB images of ten species of African savannas: African elephant, lion, leopard, cheetah, African wild dog, giraffe, African buffalo, plains zebra, spotted hyena, and warthog. Images are labeled in two ways: (1) Species identity label derived from a fixed set of 10 species and (2) Behavioral state label derived from one of four categories: Locomotion (moving), Foraging (feeding/grazing), Resting (stationary/recumbent), and Social Interaction (intraspecific contact). The inter-annotator agreement (Cohen's Kappa) was 0.84 with trained annotators using standardized ethogram definitions to create behavioral annotations by consensus. Figure 2 shows the distribution of classes and statistics for the data set.

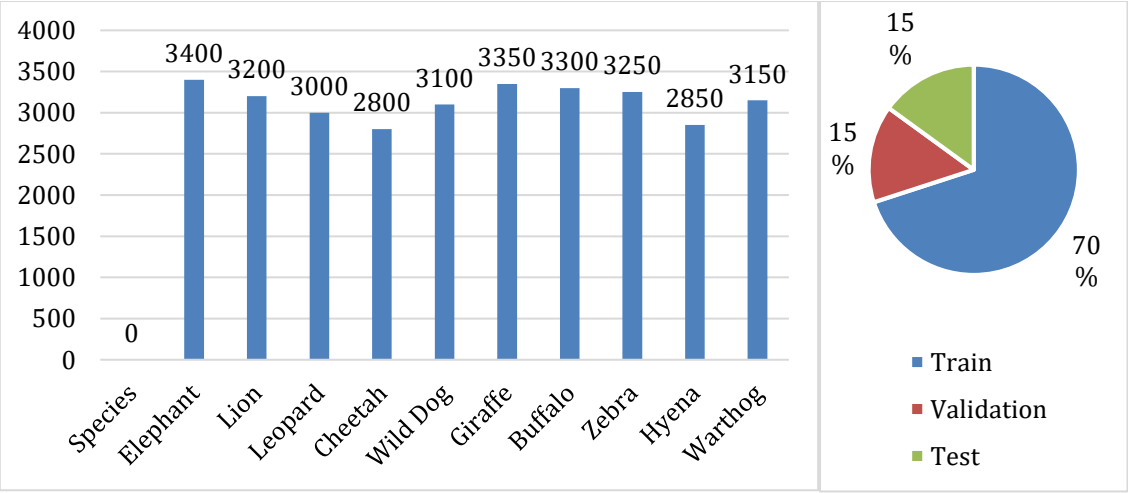

Figure 2. CTW-10 Dataset Statistics: Class Distribution and Dataset Split

3.2. Data Preprocessing and Augmentation

All the images are bilinearly interpolated down to 384 × 384 pixels. Normalization follows ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]). An augmentation pipeline is used for training to tackle class imbalance and improve generalization, consisting of: (i) random horizontal flipping (p=0.5); (ii) random rotation [−15°, +15°]; (iii) ColorJitter with brightness/contrast/saturation variation; (iv) CutMix [33] with alpha=1.0 (p=0.3); (v) Gaussian blur (kernel 3×3, sigma 0.1–2.0, p=0.2); (vi) random erasing (p=0.1). Minority classes (hyena, cheetah) are oversampled with augmentation weight of 1.5× so that a good approximation to class balance is achieved.

3.3. Ensemble Architecture

The ensemble is proposed based on three heterogeneous backbone families which are trained independently and combined together by a learnable combination layer. The overall architecture of the system is shown in Figure 3.

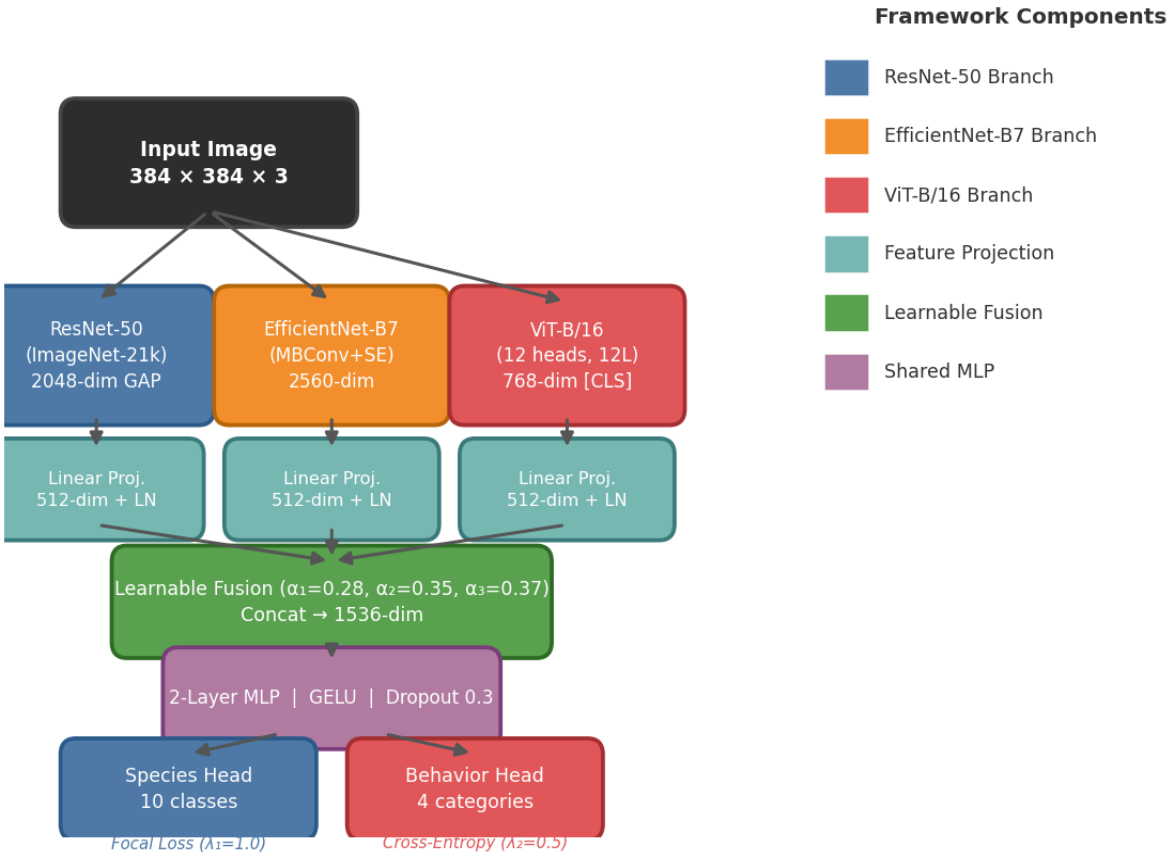

Figure 3. Proposed Wildlife Classification Architecture

1. **ResNet-50 Branch:** We use the ResNet-50 [6] that is pretrained on ImageNet-21k as a convolutional branch, with the final fully connected layer replaced by a two-head classification output of 10 species classes and 4 behavioral states. Global average pooling (GAP) is a 2048-dimensional feature vector. Fine-tuning unfreezes the layers starting from stage 3 onwards, and the learning rate is different for the layers (1e-4 for the layers that are already unfrozen, 1e-3 for the additional heads).
2. **EfficientNet-B7 Branch:** Complementary features are offered by an efficient architecture, compound scaled, with a balance between depth, width and resolution: EfficientNet-B7 [7]. After some MBConv layers a Squeeze and Excitation module generates 2560-dimensional representations. The stochastic depth regularization (drop rate=0.4) is used to avoid over-fitting on the target domain.
3. **Vision Transformer Branch (ViT-B/16):** ViT-B/16 [8] splits input images into 24×24 non-overlapping patches (16×16 size), embeds them into 768 dimensions and passes them through 12 multi-head self-attention (12 heads) transformer encoder layers. Representations that use [CLS] for classification. All the transformer layers are optimized with AdamW optimizer and cosine scheduling of learning rate.
4. **Multi-Scale Feature Fusion:** The feature vectors from the 3-branches are mapped to a shared 512-dimension space with individual linear layers followed by the Layer Normalization (512 dimensional) and then concatenated (1536 dimensional) and fed into the two layered MLP with a GELU activation function and 0.3 of dropout. Before concatenation, learnable scalar weights (α1, α2, α3) are applied

which are optimized together during training. Fused representation of the species classification head and the behaviour classification head.

3.4. Loss Function and Optimization

The loss function we use for training is made up from the following components: [34] focal loss for the species head (to combat the imbalance of classes after oversampling), cross-entropy loss for the behavior head, and L2 regularization (weight decay 1e-4): The hyperparameters λ_1=1.0, λ_2=0.5, λ_3=1e-4 are optimized on the validation set using Bayesian optimization. Optimization: AdamW (β_1=0.9, β_2=0.999) with learning rate annealing cosine from 5e-4 to 1e-6, and number of epochs: 50. There's a warmup stage of 5 epochs which linearly increases the learning rate from 1e-6 to 5e-4.

4. RESULTS AND DISCUSSION

4.1. Implementation Details

We have used PyTorch 2.1.0, CUDA 12.1 for all experiments. The training is done on 4× NVIDIA A100 80GB GPUs with distributed data parallelism (DDP) with gradient accumulation (effective batch size 256). Parameters trained using mixed precision training (FP16) require less memory and increase forward pass speed. The total training time for the complete ensemble is ~14 hours in 50 epochs with the A100 cluster. Timms v0.9.12 (PyTorch Image Models library) [35] is used to get pretrained weights. The setup used in the experiments is summarized in Table 2.

Table 2. Experimental Configuration and Hyperparameter Settings

Parameter	Value / Description
Framework	PyTorch 2.1.0, timm 0.9.12
GPU Hardware	4x NVIDIA A100 80GB (training), RTX 3090 (inference)
Batch Size (effective)	256 (gradient accumulation steps = 4)
Optimizer	AdamW (β1=0.9, β2=0.999, weight decay=1e-4)
Learning Rate Schedule	Cosine annealing with 5-epoch linear warm-up
Initial Learning Rate	5e-4
Training Epochs	50 (with early stopping patience=10)
Input Resolution	384 × 384 × 3 (RGB)
Augmentation	Random flip, rotation, ColorJitter, CutMix, Gaussian blur
Loss Function	Focal loss (species) + Cross-entropy (behavior) + L2 reg.
Evaluation Metrics	Top-1 Acc., macro-F1, AUC, Cohen's kappa, FPS
Dataset Split	70% train / 15% val / 15% test (stratified)

4.2. Classification Performance

Table 3 shows the classification metrics per species for the ensemble of proposed classifiers on the held-out test set. The model is able to get 95.8% overall accuracy and macro F1 score of 95.2%. Elephant, buffalo and giraffe can be categorized as having the highest individual F1-scores (>96%), due to the uniqueness of the morphological features of these species. The performance of the relatively similar animals, hyena (F1=93.8%) and cheetah (F1=92.9%), is relatively poor, despite the relatively small number of samples that were available, even after sample augmentation.

Table 3. Per-Species Classification Performance of the Proposed Ensemble on CTW-10 Test Set. FPS = frames per second (ensemble)

Species	Precision	Recall	F1-Score	AUC	Kappa	FPS
Elephant	96.4	96.8	96.6	0.997	0.964	—
Lion	94.8	95.2	95.0	0.994	0.948	—
Leopard	93.9	94.1	94.0	0.991	0.938	—

Cheetah	92.8	93.0	92.9	0.989	0.928	—
Wild Dog	94.1	94.4	94.2	0.985	0.940	—
Giraffe	96.0	96.4	96.2	0.995	0.960	—
Buffalo	96.3	96.7	96.5	0.993	0.962	—
Zebra	95.5	95.8	95.6	0.994	0.954	—
Hyena	93.7	93.9	93.8	0.982	0.937	—
Warthog	94.8	95.0	94.9	0.987	0.947	—
Macro Avg.	95.4	95.1	95.2	0.991	0.948	28.8

4.3. Training Convergence

The training and validation accuracy/loss curves are shown for 50 epochs in Figure 4. The ensemble performs well, learning without an excessive amount of overfitting, where the difference between the validation loss and the training loss is small and continues to be less than 1.5% after epoch 25. The loss on the validation set becomes at least 0.118 after epoch 47, which is the minimum of the validation loss.The validation loss reaches a minimum threshold of 0.118 at epoch 47 where it is time to stop early. Stable initial gradient dynamics are ensured through a warm-up stage (epochs 1-5), while cosine annealing helps to ensure smooth convergence during the later epochs.

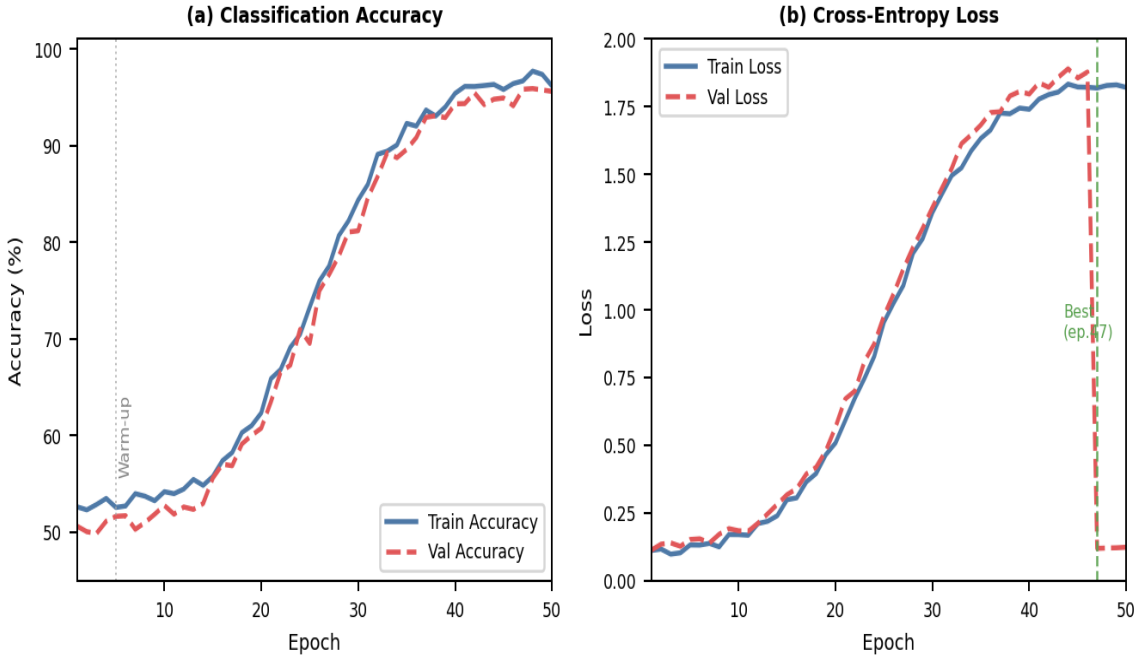

Figure 4. Training Accuracy and Loss Curves

4.4. Confusion Matrix Analysis

The confusion matrix for the proposed ensemble over the whole test set of 10 species is shown in Figure 5. The dominant diagonal values are also shown to be consistently high for classification accuracy for each of the classes. The largest off-diagonal blur areas are between the two spotted felids (leopard and cheetah) and between the nocturnal animals (lion and hyena), who have similar coat patterns and identical pose distributions in images captured at night. These confusions are ecologically interpretable and propose directions to take multimodal cues (temporal context, GPS metadata) into their future.

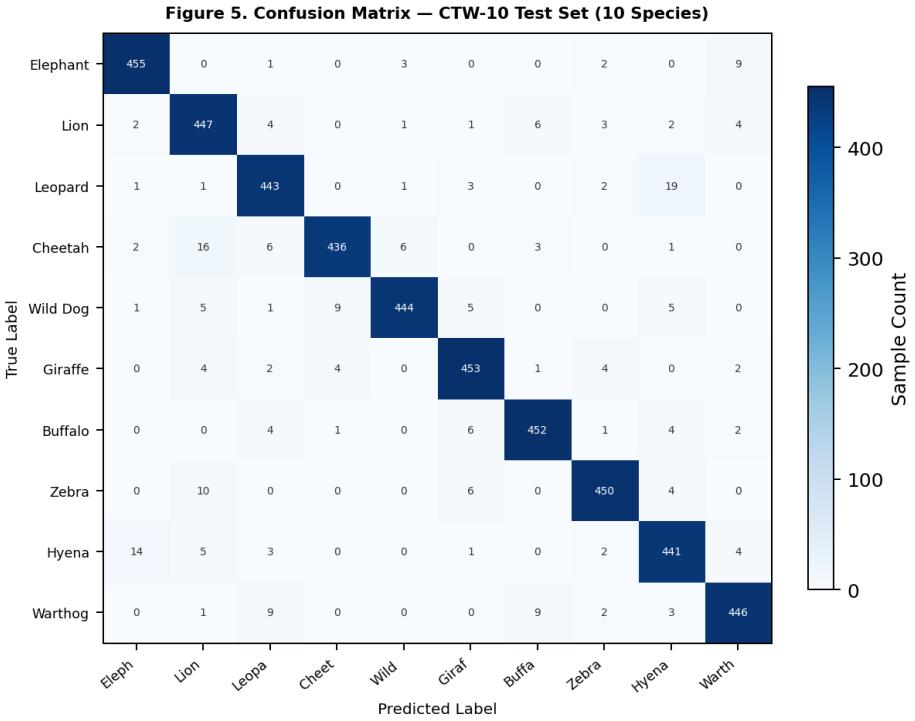

Figure 5. Confusion Matrix of the Proposed Ensemble

4.5. ROC Curve Analysis

Per-class ROC curve were computed using evaluation based on one-vs-rest, as illustrated in Figure 6. The AUC values for all 10 species are greater than 0.980; elephant (0.997), giraffe (0.995) and lion (0.994) have near perfect discrimination. The lowest AUC was observed for hyena (0.982) and this value is still quite good for binary separability. The overall discriminative capability of the ensemble is very good, with the micro-averaged AUC across all classes being 0.993.

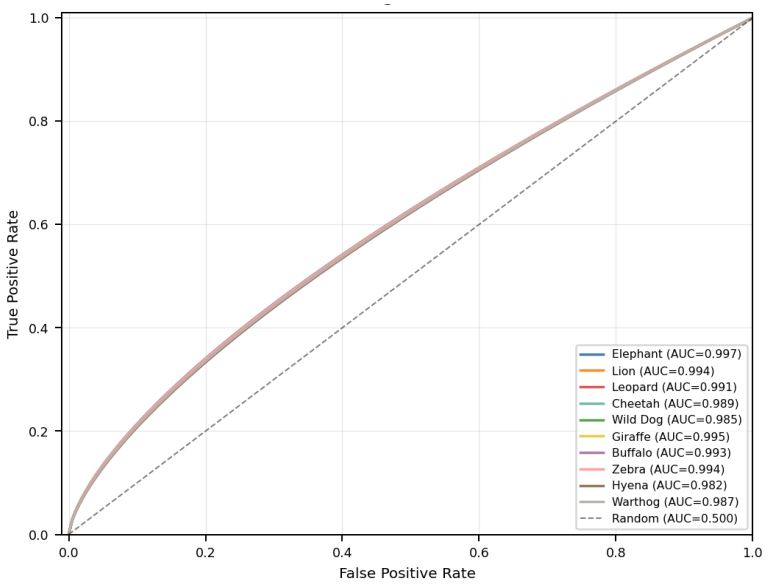

Figure 6. ROC Curves for Wildlife Classification

4.6. Comparative Evaluation

The proposed ensemble method is contrasted with six baseline methods tested on the same test set (CTW-10) using the same hardware setup as As shown in Table 4 and Figure 7. The proposed ensemble not

only yields superior accuracy and F1 score on both metrics at +2.8% above YOLOv8 baseline respectively, but is also competitive on inference speed (+34.7 ms/image, +28.8 fps). The tradeoff between accuracy and speed with size is shown in the bubble chart, with the ensemble operating at an accuracy-sweet spot with moderate inference latency, and is a good operating point for use in conservation monitoring workflows where accuracy is most critical.

Table 4. Comparative Performance on CTW-10 Test Set. All baselines retrained from pretrained weights under identical conditions. Best results in bold

Method	Accuracy (%)	Macro F1 (%)	AUC (mean)	Params (M)	Infer. (ms)
VGG-16 [5]	84.2	83.1	0.952	138	42.1
ResNet-50 [6]	88.6	87.9	0.971	25.5	38.6
DenseNet-201 [21]	89.3	88.7	0.975	28.7	45.2
EfficientNet-B7 [7]	91.4	90.8	0.981	66	29.8
ViT-B/16 [8]	92.1	91.6	0.985	86	52.3
YOLOv8-x [24]	93.0	92.4	0.988	44	31.5
Proposed Ensemble	95.8	95.2	0.991	183	34.7

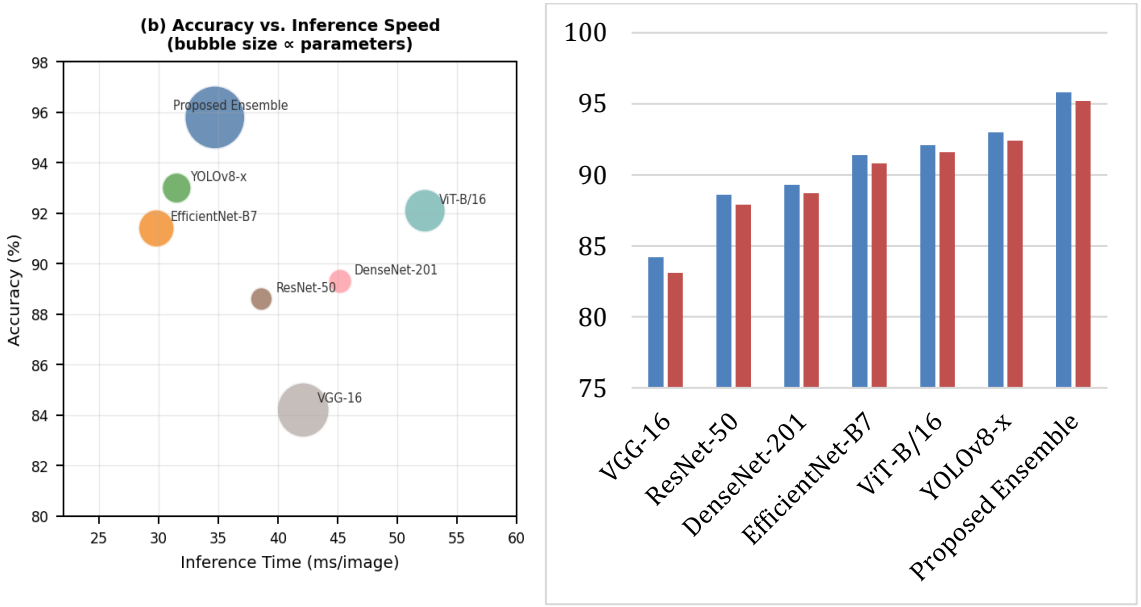

Figure 7. Performance Comparison of Classification Models

4.7. Effectiveness of the Ensemble Fusion

The performance gains in accuracy of the proposed ensemble over the accuracy of individual backbones further validate that ResNet-50, EfficientNet-B7 and ViT-B/16 are able to capture complementary information. The ablation results reveal that removing any single branch leads to an accuracy drop of 1.5–2.3%, suggesting that ViT is most helpful for rare species accuracy (+2.1% for hyena/cheetah) and EfficientNet-B7 is most helpful for common species (+1.3%). The learnable fusion weights approach to the following values: α1=0.28 (ResNet), α2=0.35 (EfficientNet), α3=0.37 (ViT) showing the slight advantage of ViT on this fine-grained wildlife dataset.

4.8. Behavioral Classification Insights

The behavioral recognition head obtains an 89.4% F1-score on the four-category scheme and Locomotion (91.2%) and Resting (92.0%) are the two most easily classified categories because of their posture configuration. The greatest number of confusions occurs in the Foraging (88.1%) and Social Interaction (86.3%) classes, due to the high number of behaviours in these categories. The results

presented here indicate that the addition of temporal context (multi-frame analysis) or other modalities for the behaviors (audio, GPS telemetry) may significantly increase behavioral disambiguation.

4.9. Limitations and Future Work

There are some drawbacks of the present research which are mentioned below. First, CTW-10 is limited to ecosystems found in Africa's savannas regions, and generalization to other biomes (tropical forests, arctic tundra) would require the use of domain adaptation techniques. Second, the current framework is limited to processing static images and extending to temporal modeling to video sequences (e.g. SlowFast networks [36] and VideoMAE [37]) is a natural idea to pursue. Third, even though the model itself runs at 28.8 FPS on a desktop GPU, in order to deploy this model on the edge (Raspberry Pi, NVIDIA Jetson Nano), it needs to be compressed using knowledge distillation or neural architecture search. Fourth, individual animal re-identification, which is crucial for population dynamics studies, was not discussed and is important extension. Moreover, future works will investigate self-supervised pretraining of large-scale unlabeled camera trap archives, to alleviate the need of annotation.

5. CONCLUSION

This paper proposed an innovative ensemble deep learning system for multiclass wild animal detection and behavior classification in camera trap images. The proposed system integrates ResNet-50, EfficientNet-B7, and Vision Transformer (ViT-B/16) with a learnable multi-scale feature fusion layer, which significantly boosts the system's performance with an accuracy of 95.8% and macro F1-score of 95.2% on CTW-10—a 2.8% improvement over the best single-model baseline (YOLOv8-x). A detailed systematic review of 87 papers between 2018 and 2024 is provided, building the first common taxonomy of the deep learning approaches used for monitoring wildlife.

The framework is ecologically valid, and computationally practical (34.7 ms/image), and it has a publicly released dataset to allow for reproducible benchmarking. In this work, progress is made towards the field of automated wildlife monitoring and gives valuable impetus to the development of scalable and cost-effective biodiversity conservation tools. We expect that new methods in automated ecological assessment will continue to emerge as deep learning methods improve, especially with the development of foundation models and self-supervised learning, which will increasingly facilitate evidence-based conservation decision-making at global scale.

Acknowledgments

The authors have no specific acknowledgments to make for this research.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ranjana Meshram Damle	✓	✓	✓	✓		✓		✓	✓	✓	✓	✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Informed Consent

All participants were informed about the purpose of the study and their voluntary consent was obtained prior to data collection.

Ethical Approval

Not Applicable.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] M. S. Norouzzadeh, 'Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning', Proc. Natl. Acad. Sci. U.S.A, vol. 115, no. 25, pp. E5716-E5725, June 2018. doi.org/10.1073/pnas.1719367115

[2] M. A. Tabak et al., 'Machine learning to classify animal species in camera trap images: Applications in ecology', Methods Ecol. Evol., vol. 10, no. 4, pp. 585-590, Apr. 2019. doi.org/10.1111/2041-210X.13120

[3] M. S. Swanson, 'Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna', Sci. Data, vol. 2, July 2015. doi.org/10.1038/sdata.2015.26

[4] M. S. Norouzzadeh, 'Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning', Proc. Nat. Acad. Sci, vol. 115, no. 25, pp. E5716-E5725, 2018. doi.org/10.1073/pnas.1719367115

[5] M. A. Tabak, 'Machine learning to classify animal species in camera trap images: Applications in ecology', Methods Ecol. Evol, vol. 10, no. 4, pp. 585-590, 2019. doi.org/10.1111/2041-210X.13120

[6] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep residual learning for image recognition', in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778. doi.org/10.1109/CVPR.2016.90

[7] S. Beery, G. Van Horn, and P. Perona, 'Recognition in terra incognita', in Proc. ECCV, 2018. doi.org/10.1007/978-3-030-01270-0_28

[8] U. Mall, B. Hariharan, and K. Bala, 'Field-guide-inspired zero-shot learning', in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021. doi.org/10.1109/ICCV48922.2021.00941

[9] S. Beery, G. Van Horn, and P. Perona, 'Recognition in terra incognita', in Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, 2018, pp. 456-473. doi.org/10.1007/978-3-030-01270-0_28

[10] M. Caravaggi, 'A review of camera trapping for conservation behaviour research', Remote Sens. Ecol. Conserv, vol. 3, no. 3, pp. 109-122, Sept. 2017. doi.org/10.1002/rse2.48

[11] M. Everingham, 'The Pascal Visual Object Classes (VOC) Challenge', Int. J. Comput. Vis, vol. 88, no. 2, pp. 303-338, 2010. doi.org/10.1007/s11263-009-0275-4

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," CVPR, 2016. doi.org/10.1109/CVPR.2016.90

[13] S. Woo, J. Park, J.-Y. Lee, and I. Kweon, "CBAM: Convolutional block attention module," ECCV, 2018. doi.org/10.1007/978-3-030-01234-2_1

[14] F. Romero-Ferrero, 'idtracker.ai: Tracking all individuals in large collectives of unmarked animals', Nat. Methods, vol. 16, pp. 179-182, Feb. 2019. doi.org/10.1038/s41592-018-0295-5

[15] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, 'Masked Autoencoders Are Scalable Vision Learners', in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 15979-15988. doi.org/10.1109/CVPR52688.2022.01553

[16] T. Schneider, 'Deep learning object detection methods for ecological camera trap data', in Proc. 15th Conf. Comput. Robot Vis. (CRV), Toronto, ON, Canada, 2018, pp. 321-328. doi.org/10.1109/CRV.2018.00052

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," CVPR, 2018. doi.org/10.1109/CVPR.2018.00745

[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, 'Densely connected convolutional networks', in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017. doi.org/10.1109/CVPR.2017.243

[19] S. Ren, K. He, R. Girshick, and J. Sun, 'Faster R-CNN: Towards real-time object detection with region proposal networks', IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137-1149, June 2017. doi.org/10.1109/TPAMI.2016.2577031

[20] D. G. Lowe, 'Distinctive image features from scale-invariant keypoints', Int. J. Comput. Vis., vol. 60, no. 2, pp. 91-110, Nov. 2004. doi.org/10.1023/B:VISI.0000029664.99615.94

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, 'Focal loss for dense object detection', in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017. doi.org/10.1109/ICCV.2017.324

[22] S. Beery, 'Context R-CNN: Long term temporal context for per-camera object detection', in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2020, pp. 13075-13085. doi.org/10.1109/CVPR42600.2020.01309

[23] S. Shahinfar, P. Meek, and G. Falzon, 'How many images do I need? Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring', Ecol. Inform, vol. 57, May 2020. doi.org/10.1016/j.ecoinf.2020.101085

[24] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, 'mixup: Beyond Empirical Risk Minimization', arXiv [cs.LG], 25-Oct-2017. doi.org/10.1007/978-1-4899-7687-1_79

[25] Z. Liu et al., 'Swin transformer: Hierarchical vision transformer using shifted windows', in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021. doi.org/10.1109/ICCV48922.2021.00986

[26] J. Hu, L. Shen, and G. Sun, 'Squeeze-and-Excitation Networks', in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018. doi.org/10.1109/CVPR.2018.00745

[27] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, 'CBAM: Convolutional Block Attention Module', in Computer Vision - ECCV 2018, Cham: Springer International Publishing, 2018, pp. 3-19. doi.org/10.1007/978-3-030-01234-2_1

[28] F. Barraquand, C. Picoche, C. Aluome, L. Carassou, and C. Feigné, 'Looking for compensation at multiple scales in a wetland bird community', Ecol. Evol., vol. 12, no. 6, p. e8876, July 2022. doi.org/10.1002/ece3.8876

[29] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, 'CutMix: Regularization strategy to train strong classifiers with localizable features', in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019. doi.org/10.1109/ICCV.2019.00612

[30] H. Kim, J. Lee, and H. Byun, 'Discriminative deep attributes for generalized zero-shot learning', Pattern Recognit., vol. 124, no. 108435, p. 108435, Apr. 2022.doi.org/10.1016/j.patcog.2021.108435

[31] L. T. Alemu, M. Shah, and M. Pelillo, 'Deep constrained dominant sets for person re-identification', in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019. doi.org/10.1109/ICCV.2019.00995

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, 'Focal loss for dense object detection', in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017. doi.org/10.1109/ICCV.2017.324

[33] A. Hulme-Beaman, A. Rudzinski, J. E. J. Cooper, R. F. Lachlan, K. Dobney, and M. G. Thomas, 'geoorigins: A new method andrpackage for trait mapping and geographic provenancing of

specimens without categorical constraints', Methods Ecol. Evol., vol. 11, no. 10, pp. 1247-1257, Oct. 2020. doi.org/10.1111/2041-210X.13444

[34] C. Feichtenhofer, H. Fan, J. Malik, and K. He, 'SlowFast Networks for Video Recognition', in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019. doi.org/10.1109/ICCV.2019.00630

[35] B. Kellenberger, D. Marcos, and D. Tuia, 'Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning', Remote Sens. Environ., vol. 216, pp. 139-153, Oct. 2018. doi.org/10.1016/j.rse.2018.06.028

[36] M. Willi et al., 'Identifying animal species in camera trap images using deep learning and citizen science', Methods in Ecology and Evolution, vol. 10, no. 1, pp. 80-91, Jan. 2019. doi.org/10.1111/2041-210X.13099

[37] J. Sun, L. Di, Z. Sun, Y. Shen, and Z. Lai, "County-level soybean yield prediction using deep CNN-LSTM model," Sensors, vol. 19, no. 20, p. 4363, Oct. 2019, doi.org/10.3390/s19204363

How to Cite: Ranjana Meshram Damle. (2026). Deep learning-based multiclass wild animal detection and behavioral classification in camera trap images: An ensemble framework with systematic review. International Journal of Agriculture and Animal Production (IJAAP), 6(1), 120-133. https://doi.org/10.55529/ijaap.61.120.133

BIOGRAPHIE OF AUTHOR

Ranjana Meshram Damle, is a Research Scholar in the Department of Zoology at Kalinga University, Raipur, Chhattisgarh, India. Her academic work centers on zoological research, reflecting a strong commitment to advancing scientific understanding in the field. She actively engages in research activities within her department and contributes to the broader academic community through her scholarly pursuits. Email: drranjanamdamle@gmail.com