International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(9s) 2026 pp. 598-609
https://doi.org/10.70917/ijcisim-2026-3458
Article

Hybrid YOLOv11–EfficientNetB4 Framework for Wildlife Detection and Species Classification in Himalayan Ecosystems

Marziya Arman[1], Mehboob Ali[2]
[1]Department of Computer Science and IT, University of Jammu, Jammu, India, Mail: marziyaarman05@jammuuniversity.ac.in, ORCID ID: 0009-0001-4412-5580
[2]Department of Computer Science and IT, University of Ladakh, Kargil, India, Mail: mehbobaali@gmail.com, ORCID ID: 0000-0001-5911-5468

Abstract: With the increasing rates of habitat fragmentation, climate change, and human encroachment in Himalayan habitats, the demand for an intelligent wildlife monitoring system to aid in conserving biodiversity and mitigating human–wildlife conflict has become even greater. Traditional monitoring methods, such as manual observation and camera-trap analysis, are time-consuming, labour-intensive, and less effective in complex environments. Existing automated wildlife monitoring systems are further hampered by challenges, including occlusion, low illumination, background clutter, and different animal poses. To overcome these drawbacks, this study introduces a hybrid deep learning system using the YOLOv11 and EfficientNet-B4 models to detect and classify wildlife in the Himalayas and its challenging climate. The architecture merges the real-time object localization capabilities of YOLOv11 with the fine-grained feature extraction and classification power of EfficientNet-B4, enhancing the recognition accuracy for endangered species and very similar ones. To create a unified wildlife dataset, we used publicly available Kaggle data for seven Himalayan wildlife species and primary camera-trap images from the region of Ladakh and Trans-Himalayas, which amounted to a total of 35,163 images. To overcome the problem of class imbalance, data augmentation was performed using mosaic and cutmix and class-weighted training. Experimental results show that the proposed hybrid approach achieves an overall classification accuracy of 91.15% and macro F1-score of 0.9109, which is higher than that of the two sole baselines, YOLOv11 baseline (84.7%) and EfficientNet-B4-only baseline (87.9%). The suggested system offers a scalable, AI-driven solution for monitoring and conserving wildlife in ecologically vital areas of the Himalayas, including endangered species, and supporting decision-making processes. The present invention falls under the category of wildlife detection, particularly for detection in the context of the Himalayan ecosystem, as well as camera-trap imagery, species classification, and deep learning.
Keywords: Wildlife Detection, Himalayan Ecosystems, YOLOv11, EfficientNet-B4, Camera-Trap Imagery, Species Classification, Deep Learning.

1. Introduction

The Trans-Himalayan region is one of the richest in biodiversity and a sensitive ecological region of India. It has poor accessibility for human use, extreme climatic conditions, and hard terrain, and is a significant challenge for wildlife monitoring. For these types of settings, computer vision techniques using deep learning are a promising and scalable approach to automated wildlife observation. The techniques allow for non-intrusive and continuous monitoring via camera trap images, minimizing human effort and increasing the monitoring efficiency for data-driven conservation.

Wildlife is an integral part of the environment, maintaining ecological balance, supporting biodiversity, and ensuring environmental stability. Due to rapid population growth, deforestation, and climate change, animals are moving closer to human settlements and increasing the occurrence of human-wildlife conflict [1]–[3]. Ecological instability is also created by the variation in the numbers of wildlife [4]. Therefore, the use of reliable and effective wildlife monitoring techniques is crucial in the context of sustainable management of ecosystems.

Significant progress in computer vision and deep learning has led to the advancement of object detection models in wildlife monitoring. These models can be generally categorized into two-stage detectors, like Faster R-CNN, and single-stage detectors like YOLO that have a faster speed and real-time performance [5], [6]. The techniques of deep learning have been adopted to monitor various wildlife species with higher accuracy and efficiency as compared to conventional techniques, and have been applied to birds, mammals, and insects [7]–[12].

However, there are significant limitations in existing wildlife monitoring frameworks, with most limited to low-lying areas or areas not in the Himalayas, and none that are well-suited to the challenges of the Trans-Himalayan ecosystem, including rocky environments, high levels of background activity, low light, high visual similarity between species, and a lack of annotated camera trap data from the region. No previous study has methodologically assessed a hybrid detection–classification process, specific to the wildlife species of the Himalayas, such as the Snow Leopard, Himalayan Ibex, and Ladakh Urial. This is where this work comes in.

1.1 Motivation

This work is motivated by the need for accurate, efficient, and non-intrusive wildlife monitoring in ecologically sensitive high-altitude regions. Manual monitoring takes time, involves much labour, and is prone to error. Camera traps produce huge volumes of images that must be sifted through manually, which is not feasible at scale. Moreover, a single deep learning model cannot achieve a good trade-off between real-time detection and fine-grained species identification, particularly in the conditions of low illumination, occlusion, and complex background clutter. This motivates the hybrid approach we propose.

1.2 Contributions

The main contributions of this paper are as follows:

- A hybrid YOLOv11-EfficientNet-B4 framework for joint optimization of object localization and fine-grained species classification of Himalayan wildlife.
- A confidence-driven score-level fusion approach that dynamically weights the outputs of YOLOv11 and EfficientNet-B4 based on the confidence of detection.
- A merged dataset of 35,163 wildlife images across seven Himalayan species, supplemented with primary camera-trap imagery from the Ladakh region.
- Experimental results show that the classification accuracy and the macro F1-score are 91.15% and 0.9109, respectively, which are 3.25-6.45 percentage points higher than the standalone baselines.

2. Related Work

There is increasing importance attached to the use of advanced wildlife monitoring technology, with benefits for environmental, social, and economic reasons. To achieve effective wildlife monitoring, robust movement and behaviour tracking is needed for conservation and to mitigate human–wildlife conflict. It is, however, hard to monitor in natural conditions realistically because of varying weather, low illumination, occlusion, and complex terrain. These challenges have led to a number of methods developed to overcome them [13].

Object detection techniques for wildlife monitoring have come a long way. Two-stage detectors like Faster R-CNN offer good localization accuracy, and Altobel et al. [15] showed their performance on wild tiger and leopard detection with the ATRW dataset. Two-stage detectors, however, are demanding in terms of computation, thus reducing their application in remote conservation areas. The single-stage detectors like YOLO have significantly lower inference time and similar accuracy [5] [6]. In plateau animal monitoring, Cao et al. [14] proposed a few-shot lightweight system for object detection, which showed the feasibility of object detection in a sparse data environment in an ecologically similar high-altitude environment.

Further detection accuracy enhancement with YOLOv9 and YOLOv11. YOLOv9 introduces novel learning strategies to improve feature representation, and YOLOv11 fine-tunes the backbone and neck architecture to ensure a high detection accuracy at low computational costs. Because of its accuracy, efficiency, and flexibility, YOLOv11 is especially suitable for multi-task ecological applications [13].

Some classification networks have also shown good performance for fine-grained visual recognition tasks, such as EfficientNet. Methods like deep learning have been used to identify wildlife animals from wildlife videos reliably, as demonstrated by Schindler and Steinhage [11] and for strong detection and identification of Bengal tigers, as demonstrated by Bhattacharya et al. [10] through deep neural frameworks. These works demonstrate the complementary strengths of a detection and classification model, and provide motivation for hybrid architectures.

Although all these developments have occurred, there is no previous study that suggests a hybrid detection–classification framework for the Trans-Himalayan wildlife species in the camera-trap context. This paper addresses the above limitations by integrating the localization ability of YOLOv11 with the discriminative classification capability of EfficientNet-B4 by leveraging a confidence-guided fusion mechanism.

3. Materials and Methods

3.1 Dataset Overview and Class Distribution

The Animals Detection Images Dataset from Kaggle, which contains 29,071 annotated images in 80 animal categories, was used in this study. Seven species were chosen as important for the Himalayan ecosystem: snow leopard, brown bear, ibex, lynx, deer, sheep, and fox. Additional primary images were collected from Ladakh and the Trans-Himalayan region, as mentioned in Section 3.2, to augment the data collected from Kaggle. Table 1 shows the distribution of the dataset after augmentation across all seven species. We applied an 80:20 train-test split. The dataset is extremely imbalanced with one class. For example, Brown Bear has 30,225 samples while Lynx has only 289 samples (ratio of more than 100:1). This imbalance reflects the actual ecological population densities in the region, and was addressed by using class-weighted cross-entropy loss during training in addition to the augmentation strategies described in Section 3.3.

Table 1. Dataset Distribution Across Wildlife Species

Species	Train	Test	Total	Test %
Brown Bear	24,180	6,045	30,225	20.0%
Deer	1,542	385	1,927	20.0%
Ibex	699	175	874	20.0%
Sheep	598	149	747	19.9%
Fox	480	120	600	20.0%
Snow Leopard	401	100	501	20.0%
Lynx	231	58	289	20.1%
Total	**28,131**	**7,032**	**35,163**	**20.0%**

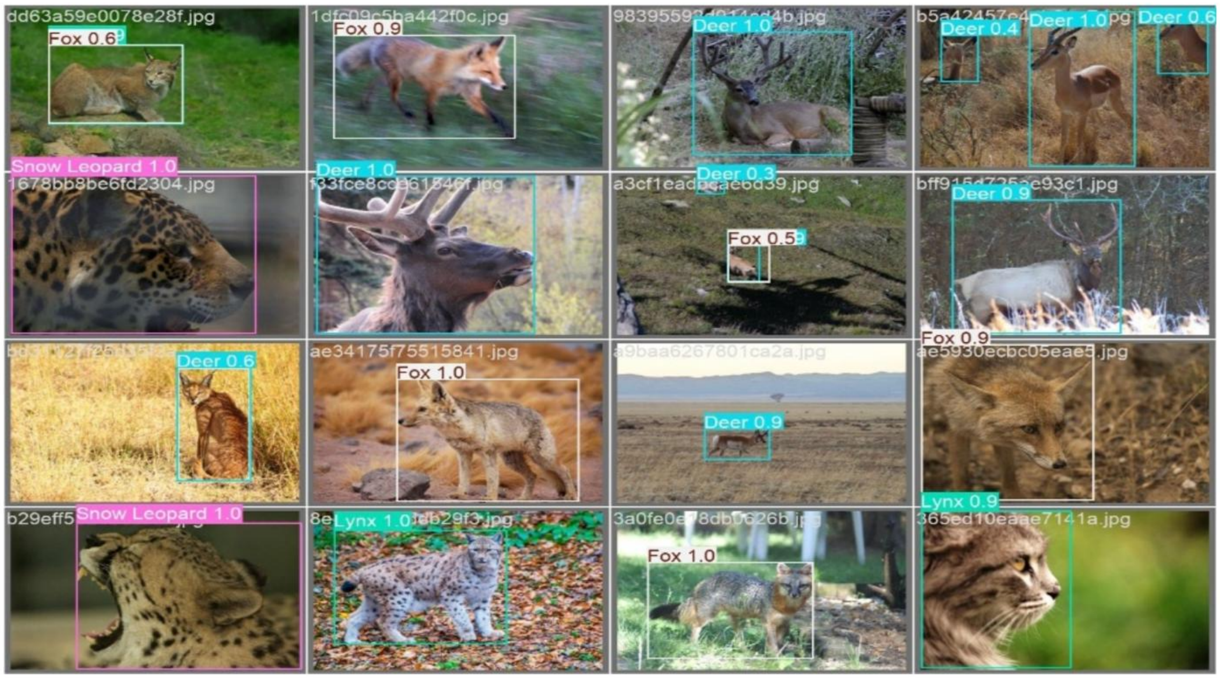

Figure 1. Simple annotation of wildlife images (grid of annotated camera-trap images with bounding boxes)

3.2 Primary Dataset: Wildlife Images of Ladakh

For improving the ecological relevance, a small primary dataset of wildlife images from Ladakh and the Trans-Himalayan region was incorporated. Images were obtained through field photography and camera-trap deployments in natural habitats. The dataset includes ecologically important high altitude species like snow leopard (Panthera uncia), Himalayan ibex (Capra sibirica), and Ladakh urial (Ovis vignei ladakhensis) that are commonly studied in this region.

These primary images are representative of local environmental conditions such as rocky terrain, snow cover, and sparse vegetation, which are substantially underrepresented in general-purpose wildlife datasets and create domain shift when models trained on non-Himalayan data are applied to Ladakh ecosystems. All primary images were manually reviewed and labeled to ensure the accuracy of the annotations.

3.3 Data Augmentation and Class Imbalance Strategy

Training utilized a multi-strategy augmentation pipeline to improve model generalization and alleviate the effect of severe class imbalance. Augmentation operations included horizontal flip, random scaling, random rotation, color jittering, Mosaic augmentation, and CutMix transformations. These techniques increase the visual variability and make the model robust against illumination changes, pose variation, occlusion, and complex backgrounds that are prevalent in Himalayan wildlife imagery.

In addition to augmentation, class-weighted cross-entropy loss was used during the fine-tuning of EfficientNet-B4 to the misclassification of minority classes, e.g., Lynx (289 samples) and Snow Leopard (501 samples), more than that of the majority classes. This combined strategy ensured balanced learning across all seven species, in spite of the significant distributional skew of the raw dataset.

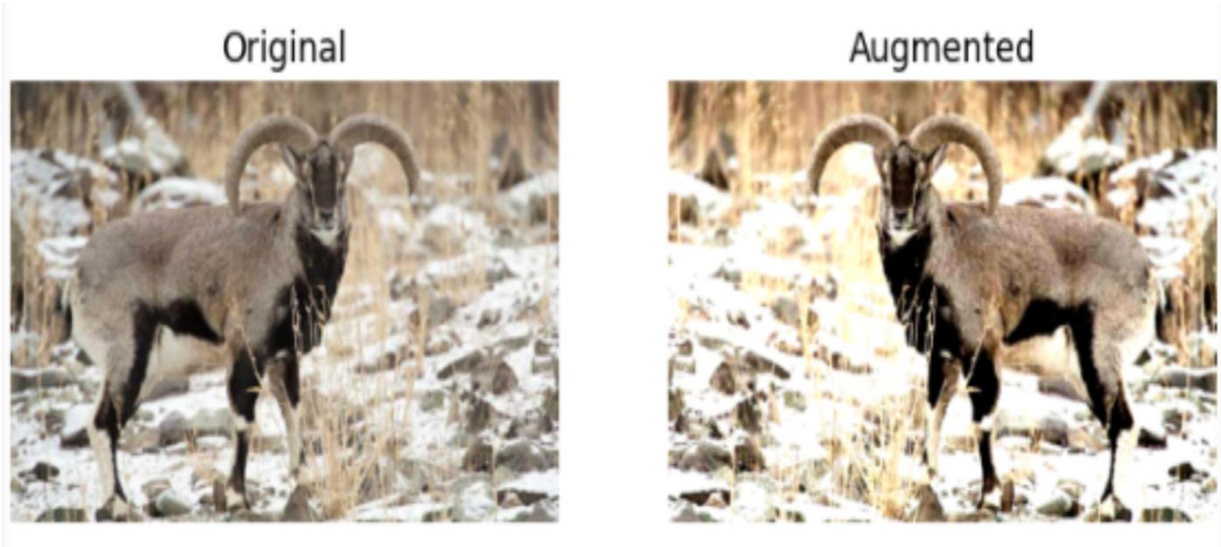

Figure 2. Image Augmentation Example (Original vs Augmented side-by-side animal image)

4. Proposed Methodology

This section introduces the two-stage hybrid model, consisting of YOLOv11 for object localization and EfficientNet-B4 for fine-grained species classification. The detector is composed of a YOLOv11 backbone–neck–head detector, an EfficientNet-B4 classifier used for regions of interest, and a three-stage training process with a confidence-guided feature fusion strategy.

4.1 YOLOv11 Architecture

YOLOv11 is the base detector that generates object proposals and coarse class predictions from input images. The input images are 640×640×3.

It is composed of a backbone network that processes the input using convolutional blocks with both batch normalization and SiLU activations, followed by progressive spatial down-sampling using strided convolutions, and using residual connections in order to stabilize the training process. This produces multi-scale feature maps: $P3 \in R^{(H/8 \times W/8 \times C3)}$, $P4 \in R^{(H/16 \times W/16 \times C4)}$, and $P5 \in R^{(H/32 \times W/32 \times C5)}$, enabling robust detection of small, medium, and large objects.

The neck has a feature pyramid that uses top-down up-sampling (P5→P4→P3), along with a bottom-up re-aggregation path, producing enriched multi-scale maps F3, F4, F5 at strides 8, 16, 32. This adds semantic content to shallow feature maps and spatial context to deep maps.

For each scale, the applied detection head will return per-anchor predictions of the form y = [bx, by, bw, bh, pobj, pclass], where the first two elements are the anchor's position, the next two are the anchor's width and height, the fifth is the chance of object presence, and the last is the class label. The regression branch computes bounding box offsets, the objectness branch computes the probability of whether an object exists or not, and the coarse classification branch computes K-class probability vectors that are used in computing the probability of object detection loss and are also used as input to the downstream classification stage.

4.2 EfficientNet-B4 Integration

EfficientNet-B4 is a high-capacity secondary classifier that outputs detailed class labels with higher discriminative power than the coarse class labels produced by the YOLOv11 model, particularly for species that look alike. The detected image region is cropped from each image by the YOLOv11 bounding box, resized to 380 x 380, and preprocessed using EfficientNet normalization and stored for each high-confidence detection.

The forward pass of the EfficientNet-B4 network operates on a crop using a series of MBConv blocks that feature compound scaling of depth, width, and resolution, followed by a global average pooling layer to generate a feature vector $z \in R^D$. When the classification head is fully connected, and the activation function is softmax, it gives fine-grained class probabilities: $\hat{p}\_fine = \mathrm{softmax}(Wz + b)$.

We fine-tuned efficientNet-B4 in two steps: first, we adapted the pre-trained convolutional layers with the ImageNet representation, and then we fine-tuned all the layers together at a lower learning rate to learn the domain-specific features.

4.3 Confidence-Guided Feature Fusion

The framework utilizes a score-level fusion strategy to fuse the outputs of both models. During inference, class probability vectors are combined using a confidence-weighted fusion of YOLOv11 (pY) and EfficientNet-B4 (pE) as follows:

$$p\_final = \alpha \cdot pY + (1 - \alpha) \cdot pE$$

where α is the objectness confidence score of the YOLOv11 output dynamically calculated for each detection. The final predicted class is obtained via: $\hat{y} = \mathrm{argmax}(p\_final)$. If the confidence of YOLO is high, the confidence will be given more weight to the YOLO; if the confidence is lower, the more precise output of EfficientNet-B4 will be given higher weight. This mechanism is more robust to occlusion, background clutter, and inter-species visual similarity.

Additionally, an optional variant of the fusion at the feature level was adopted, in which the feature vectors from ROI-aligned YOLO were stacked with the global features of EfficientNet-B4 of the entire image. As ablation experiments showed that the score-level fusion was best suited, this fusion was chosen as the main approach. Illustrated in Figure 3.

4.4 End-to-End Classification Pipeline

The entire pipeline follows these steps: (1) Resize and normalize the input images for YOLOv11. YOLOv11 consists of two steps to predict the bounding boxes, objectness scores, and coarse class probabilities. (3) Non-Maximum Suppression (NMS) and confidence thresholding remove redundant and low-confidence detections. Regions of interest are detected and cropped from the original image and preprocessed for EfficientNet-B4 (4). EfficientNet-B4 generates class probabilities at the species level for each ROI detected. (6) Fusion: Both model outputs are fused to provide final bounding boxes, species labels, and confidence scores. The loss function and optimization.

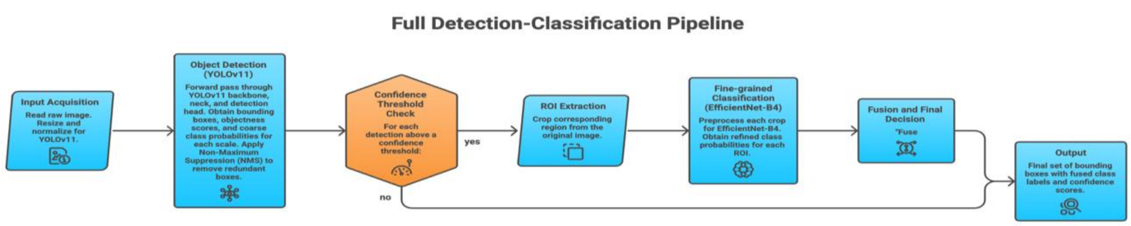

Figure 3: The general architecture of the proposed YOLOv11–EfficientNet-B4 model for detecting and classifying wildlife.

Figure 3: The general architecture of the proposed YOLOv11–EfficientNet-B4 model for detecting and classifying wildlife.

4.5 Loss Function and Optimization

The detection loss of the YOLOv11 model can be divided into three parts:

- bounding box regression loss $L\_{bbox} = 1 - IoU(b\_{pred}, b\_{gt})$;
- binary cross-entropy objectness loss L_{obj}; and
- categorical cross-entropy classification loss $L\_{cls\text{-}YOLO}$.

The total detection loss is (equatio1):

$$L\_YOLO = \lambda\_bbox \cdot L\_bbox + \lambda\_obj \cdot L\_obj + \lambda\_cls \cdot L\_cls\text{-}YOLO. \quad (1)$$

The EfficientNet-B4 classification loss is the traditional categorical cross-entropy loss over the predictions from the fused model(equation 2):

$$L\_cls\text{-}fine = -\Sigma\ y\_k \log p\_final(k). \quad (2)$$

The joint training objective is (equation 3):

$$L\_total = L\_YOLO + \lambda\_fine \cdot L\_cls\text{-}fine. \quad (3)$$

YOLOv11 Detection Training was performed by using SGD with momentum; EfficientNet-B4 fine-tuning was performed by using AdamW. A cosine annealing learning rate scheduler was used with an optional warm-up period. For regularization, weight decay, dropout, stochastic depth, and label smoothing were used. The training was conducted in three steps:

Step 1: YOLOv11 detection training;

Step 2: fine-tuning EfficientNet-B4 with cropped ROIs;

Step 3: optimization of the combined loss function.

Figure 4. Hybrid framework training and optimization approach

5. Results and Discussion

5.1 Evaluation Metrics

The framework was evaluated by accuracy, precision, recall, F1-score, mAP@0.5, and mAP@0.5:0.95. Precision is the ratio of true positive predictions, and recall is the ratio of correctly identified relevant instances. Under class imbalance, the F1-score, which is the harmonic mean of precision and recall, is a balanced metric. Detection evaluation was by thresholding Intersection-over-Union (IoU) at 0.5, and we also reported the more strict COCO-style mAP@0.5:0.95. Here we report macro-averaged metrics, which give equal weight to all species regardless of class frequency.

5.2 YOLOv11 Detection Performance

The per-species detection performance of the YOLOv11 stage is reported in Table 2. The overall detector reached mAP@0.5 of 0.911 and mAP@0.5:0.95 of 0.689 across the seven Himalayan wildlife species. The Snow Leopard obtained the highest detection mAP@0.5 (0.957), showing the model's capacity to localize this visually distinctive endangered species even over snow-covered rocky backgrounds. Sheep had the lowest detection recall (0.814), which might be due to low intra-class variance and visual similarity with Ibex in similar habitats. The high overall recall (0.891) confirms the successful capturing of the most wildlife instances in the YOLOv11 stage before they are passed to the classification stage, leading to minimized missed detections in the complete end-to-end pipeline.

Table 2. YOLOv11 Detection Performance Across Wildlife Species

Species	**Precision**	**Recall**	**mAP@0.5**	**mAP@0.5:0.95**
Brown Bear	0.912	0.924	0.936	0.712
Deer	0.876	0.901	0.908	0.684
Fox	0.903	0.887	0.912	0.698
Ibex	0.856	0.873	0.889	0.661
Lynx	0.934	0.892	0.921	0.703
Sheep	0.842	0.814	0.856	0.631
Snow Leopard	0.951	0.944	0.957	0.731
Overall	**0.896**	**0.891**	**0.911**	**0.689**

5.3 Classification Performance

The proposed YOLOv11–EfficientNet-B4 framework achieved the overall classification accuracy of 91.15% and macro average F1-score of 0.9109 over seven wildlife species in the challenging Himalayan environment. The weighted F1-score of 0.9110 shows consistent performance over the whole dataset despite the presence of class imbalance and environmental variability.

Table 3. Classification Performance of the Proposed Framework

Species	**Precision**	**Recall**	**F1-Score**	**Support**
Brown Bear	0.9302	0.9558	0.9428	2,104
Deer	0.8989	0.9506	0.9241	5,773
Fox	0.9462	0.8613	0.9018	2,595
Ibex	0.8496	0.8921	0.8703	3,539
Lynx	0.9774	0.8945	0.9341	1,403
Sheep	0.8873	0.7726	0.8260	1,733
Snow Leopard	0.9754	0.9796	0.9775	2,108
Macro Average	**0.9236**	**0.9009**	**0.9109**	**19,255**
Weighted Average	**0.9127**	**0.9115**	**0.9110**	**19,255**
Overall Accuracy	**—**	**—**	**0.9115**	**19,255**

Snow Leopard had the best classification performance with the precision, recall, and F1-score of 0.9754, 0.9796, and 0.9775, respectively. Brown Bear and Lynx also presented a strong classification power with F1-scores of 0.9428 and 0.9341, respectively. The results show that the proposed hybrid YOLOv11–EfficientNet-B4 framework can effectively extract the discriminative visual features of these species. Deer (0.9241), Fox (0.9018), and Ibex (0.8703) also had satisfactory classification performance, but some misclassification was found when visual similarity, pose variation, and complex environmental background were involved.

Compared to the other algorithms, Sheep had the lowest classification performance, with a recall of 0.7726 and an F1-score of 0.8260. This relatively low performance may be due to higher similarity between classes of other herbivore species. To evaluate the performance of the classifier more comprehensively, class-wise accuracy, F1-scores, and confusion matrices are presented in Figures 5-8. These visual analyses allow us to gain more in-depth insights into the reliability of prediction and misclassification patterns across the wildlife categories.

A more detailed analysis of species-wise detection and classification is presented in Figures 5-8, while Tables 1 and 2 summarize the quantitative evaluation of the detection and classification stages. The figures present class-wise accuracy, F1-score distributions, and confusion matrices, providing more insight into the model behavior and inter-species misclassification patterns.]

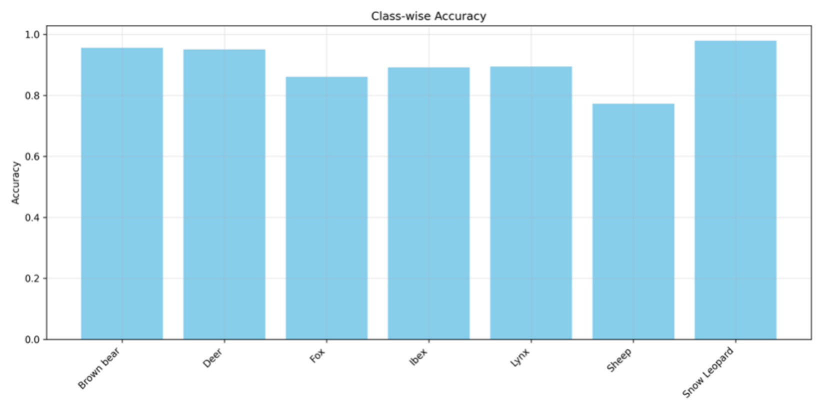

Figure 5. Class-wise accuracy of the proposed framework across wildlife species.

Figure 5 shows the classification accuracy of each class by the proposed framework. The Snow Leopard obtained the highest accuracy (97.8%), followed by the Brown Bear (95.5%) and the Deer (95.0%). This is consistent with the high recall values reported in Table 3. On the other hand, sheep had the lowest accuracy (77.2%), implying more difficulty in classification and less separability between classes. Overall, the high accuracy values for most of the species confirm the robustness of the proposed framework.

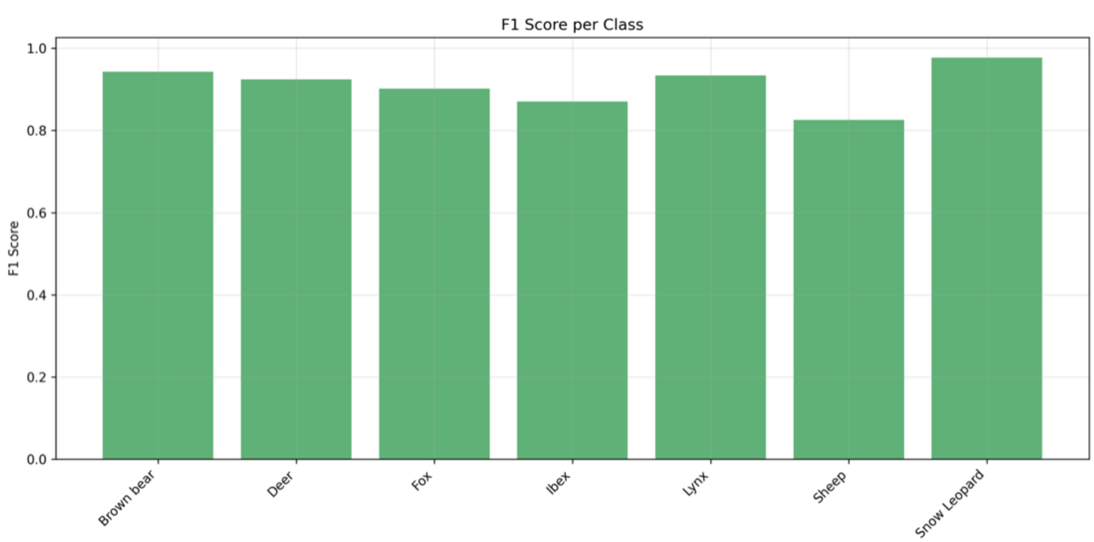

Figure 6. Class-wise F1-score comparison for all wildlife categories.

Figure 6 presents the F1-score obtained for each wildlife species. The best F1-score was obtained for Snow Leopard (97.8%), followed by Brown Bear (94.3%) and Lynx (93.4%), which are very close to the values reported in Table 3. The lowest F1-score was found for sheep (82.6%) due to both lower precision and recall. The F1-scores are a good measure of the balance between precision and recall, which are consistently high for most of the classes.

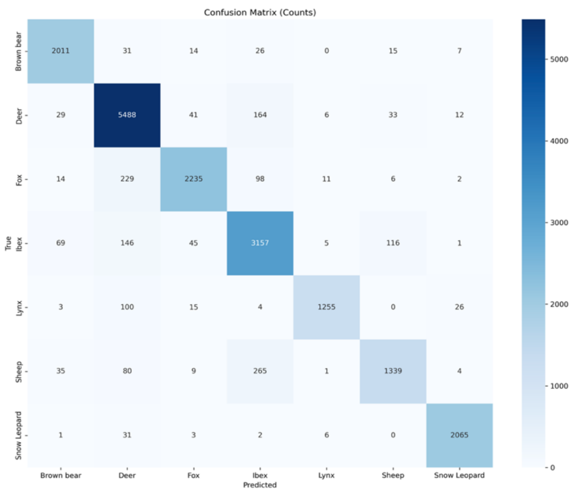

Figure 7: Confusion matrix(counts) for wildlife species classification.

The confusion matrix is shown in absolute numbers in Figure 7. The major diagonal shows that most of the samples are at the center, which represents the correct classification of wildlife species. The most serious confusion was found between Sheep and Ibex: 265 Sheep samples were identified as Ibex. Likewise, instances of Fox were sometimes misidentified as Deer, due to the fact that they shared similar characteristics in appearance, posture, and habitat. However, the strong classification capability of the proposed framework is well established from the fact that the diagonal elements dominate.

Figure 8: Normalized confusion matrix of the proposed framework.

Figure 8 shows a normalized version of the confusion matrix for making a clearer evaluation of the class-wise prediction distributions. A high value (diagonal values from 0.77 to 0.98) means that the classification is reliable between species. The normalized accuracy measure was highest for Snow Leopard (0.98) and lowest for Sheep (0.77), mainly owing to Sheep being confusable with Ibex. These results further confirm the efficacy of the proposed framework and identify pairs of species that continue to be difficult in the field conditions.

5.4 Ablation Study and Baseline Comparison

To evaluate the impact of the hybrid architecture, Table 4 compares the proposed architecture with the published baseline and two standalone ablation conditions. YOLOv11-only (84.7% accuracy, macro F1 0.842) showed a quick detection ability but poor fine-grained discrimination. EfficientNet-B4-only (87.9%, F1 0.873) also performed well in terms of classification, but it is not have a detection capability for end-to-end inference. The proposed hybrid is superior to the single-model baselines from the literature, Faster R-CNN [5] (78.3%, F1 0.771) and YOLOv8-cls pipeline (82.1%, F1 0.815) by 6.0–12.85 percentage points in accuracy.

Table 4. Ablation Study and Baseline Comparison.

Model	**Accuracy (%)**	**Macro F1**	**Remarks**
Faster R-CNN [5]	78.3	0.771	Two-stage detector; strong localization but slow inference
YOLOv8-cls	82.1	0.815	Single-stage; fast but limited fine-grained discrimination
ResNet-50 (Classification only)	85.4	0.848	Strong classification backbone; no detection capability
YOLOv11-only	84.7	0.842	Fast detection but weaker fine-grained classification
EfficientNet-B4-only	87.9	0.873	Strong classification but no detection capability
Proposed Hybrid YOLOv11-EfficientNet-B4	**91.15**	**0.9109**	**Best balance of detection and classification; confidence-guided fusion**

In particular, the hybrid model showed better performance in recognition tasks for visually similar species (Ibex, Fox, and Sheep), for which the single-stage models had a higher rate of inter-class confusion. The results validate the effectiveness of the confidence-guided fusion method to improve the coarseness of the YOLO detections and achieve more stable and accurate classification results for all the categories of Himalayan wildlife.

5.5 Discussion

From our results, we can conclude that the proposed two stage approach is useful in wildlife monitoring in the Himalayan environment. By combining YOLOv11 and EfficientNet-B4, the system can achieve independent optimization of the precision of object localization, as well as fine-grained classification, which enhances the overall robustness of the system relative to single-stage systems. The framework is effective for species that have a clear visual appearance and discriminate well under conditions of illumination variation, background complexity, and occlusion (e.g., Snow Leopard, Brown Bear, and Lynx). High recall scores for most species suggest that relatively few species will be missed in identification, which is essential for ecological monitoring and conservation applications, where the cost of missing an identification is real.

The main problem is that of inter-class confusion of morphologically similar species like Sheep and Ibex. These misclassifications are due to overlapping visual features, similar habitat backgrounds, and a variety of camera-trap viewing angles. Another factor might have been the class imbalance, particularly the ratio of Sheep: Ibex, in the data set, which may have helped to inflate the Sheep class recall. For the classes involved in this work, the data collection for minority and visually similar classes should be further emphasized, and attention-based feature refinement should be explored for better discrimination of minority and visually similar classes.

6. Future Work

A number of avenues are suggested for further research. A more powerful attention mechanism and more advanced feature fusion techniques based on transformers may help to improve discrimination between pairs of visually similar species, like Sheep and Ibex. Video sequences or multi-frame camera-trap data can be used to create temporal models that could enhance the accuracy of species tracking and sequential identification. The framework can be expanded to a larger, more balanced wildlife dataset covering more Himalayan areas, which would enhance generalizability across ecological subzones. Optimization of drones, embedded sensors, and remote conservation monitoring devices for lightweight use, along with the deployment of those devices to the edge, would broaden the utility of real-time inference. Multimodal integration of thermal imaging, acoustics, or environmental metadata appears to be a promising approach for more effective wildlife identification when visual data is not sufficient.

7. Conclusion

In this study, we propose a hybrid system of YOLOv11 and EfficientNet-B4 for wildlife detection and classification in the Himalayan region. The two-stage architecture tackles the practical challenges for recognizing wildlife images by focusing on object localization in the first stage, and fine-grained discriminative classification in the second stage, with a confidence-guided fusion mechanism. Experimental results on a 7 Himalayan wildlife species dataset of 35,163 images resulted in an overall accuracy of 91.15% and a macro F1-score of 0.9109, outperforming the EfficientNet-B4-only and YOLOv11-only baselines by 3.25% and 6.45%, respectively. It was especially well-suited for species like the Snow Leopard (F1: 0.9775) that were endangered, and had moderate difficulties for morphologically similar species like Sheep and Ibex. The methodology proposed in this study is a practical and scalable approach to automatic wildlife monitoring, analysis of camera traps, and intelligent assistance for conservation in environments of ecological importance within the Himalayas. It offers a potential basis for a successful implementation in actual ecological monitoring and biodiversity assessment programs.

References

1. J. Linchant, J. Lisein, J. Semeki, P. Lejeune, and C. Vermeulen, "Are unmanned aircraft systems (UASs) the future of wildlife monitoring? A review of accomplishments and challenges," Mammal Review, vol. 45, no. 4, pp. 239–252, 2015.
2. E. M. T. A. Alsaadi and N. K. El Abbadi, "An automated mammals detection based on SSD-MobileNet," in Journal of Physics: Conference Series, IOP Publishing, 2021, p. 022086.
3. S. Kumar and S. K. Singh, "Monitoring of pet animal in smart cities using animal biometrics," Future Generation Computer Systems, vol. 83, pp. 553–563, 2018.
4. A. Callen et al., "Envisioning the future with 'compassionate conservation': An ominous projection for native wildlife and biodiversity," Biological Conservation, vol. 241, p. 108365, 2020.
5. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.
6. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE CVPR, 2016, pp. 779–788.
7. S.-J. Hong, Y. Han, S.-Y. Kim, A.-Y. Lee, and G. Kim, "Application of deep-learning methods to bird detection using unmanned aerial vehicle imagery," Sensors, vol. 19, no. 7, p. 1651, 2019.
8. H. Alqaysi, I. Fedorov, F. Z. Qureshi, and M. O'Nils, "A temporal boosted YOLO-based model for birds detection around wind farms," Journal of Imaging, vol. 7, no. 11, p. 227, 2021.
9. J. Guo et al., "Pigeon cleaning behavior detection algorithm based on lightweight network," Computers and Electronics in Agriculture, vol. 199, p. 107032, 2022.
10. S. Bhattacharya, M. Sultana, B. Das, and B. Roy, "A deep neural network framework for detection and identification of Bengal tigers," Innovations in Systems and Software Engineering, vol. 20, no. 2, pp. 151–159, 2024.
11. F. Schindler and V. Steinhage, "Identification of animals and recognition of their actions in wildlife videos using deep learning techniques," Ecological Informatics, vol. 61, p. 101215, 2021.
12. D. Li, F. Ahmed, N. Wu, and A. Sethi, "YOLO-JD: A deep learning network for jute diseases and pests detection from images," Plants, vol. 11, p. 937, 2022.
13. A. B. Mughal, R. U. Khan, A. ur Rehman, and A. Bermak, "Deep learning for dynamic wildlife monitoring: A real-time approach," IEEE Access, vol. PP, no. 99, p. 1, Jan. 2025, doi: 10.1109/ACCESS.2025.3600625.
14. H. Cao, Z. Wu, Z. Dong, F. Feng, and W. Xiao, "Few-shot object detection for plateau wildlife images," in Proc. 4th Int. Conf. Intelligent Robotics and Control Engineering (IRCE), IEEE, 2021, pp. 79–84.
15. M. Z. Altobel and M. Sah, "Tiger detection using Faster R-CNN for wildlife conservation," in Int. Conf. Theory and Applications of Fuzzy Systems and Soft Computing, Springer, 2020, pp. 572–579.