Integrated Wildlife Monitoring System: A Multi-Modal Approach to Animal Detection and Behavior Analysis

Aditya Sharma
Division of Computer Science
Lovely Professional University,India
adityasharma96458@gmail.com

Bhavna Sahu
Division of Computer Science
Lovely Professional University, India
bhavnasahu2808@gmail.com

Soni Singh
Division of Computer Science
Lovely Professional University, India
soni.30409@lpu.co.in

***Abstract*— Ecological research, biodiversity protection, and proactive responses to species threats and habitat changes all depend on wildlife monitoring. Field observation, hand tagging, and camera traps are examples of traditional methods that are frequently constrained by human error, scope, and scale. This study suggests an integrated wildlife monitoring system that uses multi-modal data inputs to automate species detection and behavior analysis, utilizing recent developments in deep learning. A ConvLSTM autoencoder allows for the unsupervised detection of behavioral abnormalities with an accuracy of 89.3%, while a convolutional neural network (CNN) trained on more than 150,000 photos achieves 92.7% accuracy across 80 animal species. A deep feedforward neural network is used to further classify behavior from environmental sensor data. In addition to increasing monitoring's scalability and dependability, our approach makes it possible to identify hazards like illness, predators, and environmental stress early on. Its robustness and adaptability across species and terrains are demonstrated via field deployments, which represents a major breakthrough in automated conservation technology.**

Keywords: *Convolutional neural networks, ConvLSTM, deep learning, sensor data, multi-modal analysis, species detection, wildlife monitoring, animal behavior analysis, and ecological informatics.*

I. INTRODUCTION

Observing animal behavior in their natural environments is crucial for comprehending species interactions, health, and adaptation to the environment. Historically, this has depended significantly on manual monitoring and documentation, which is not only labor-intensive but also prone to human mistakes and restrictions in scope. Due to recent progress in computer vision and machine learning, automated systems can now carry out behavior analysis with increased precision and efficiency. In this context, the current research presents an extensive framework that employs video-based and CSV-formatted datasets to assess and forecast animal behavior patterns close to water bodies like rivers, lakes, and ponds.

The system revolves around various datasets, comprising around 10–14 structured CSV files along with a comprehensive video dataset showcasing different animal behaviors. A pretrained animal detection model acts as the basis for recognizing species in the video frames. This model is then used to develop an enhanced image-based detection system that can identify animal species in images and videos. At the same time, a model for predicting behavior is developed using the obtained video frames, allowing the system to deduce behavioral patterns.

The models created are incorporated into a Flask-based web application that offers an easy-to-use interface for entering data. Users can submit videos, images, or CSV files, and the system smartly handles the inputs to carry out the appropriate analysis. For video data, the system processes and captures frames prior to sending them to the behavior prediction module. For CSV input, the application uses the trained model to directly interpret behavior. This unified method offers a scalable, reachable solution for monitoring animals in real-time, delivering important insights for researchers, conservationists, and environmentalists.

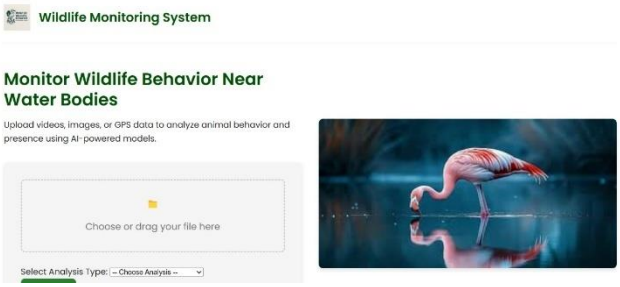

Figure 1: the Front-end of the animal monitoring website

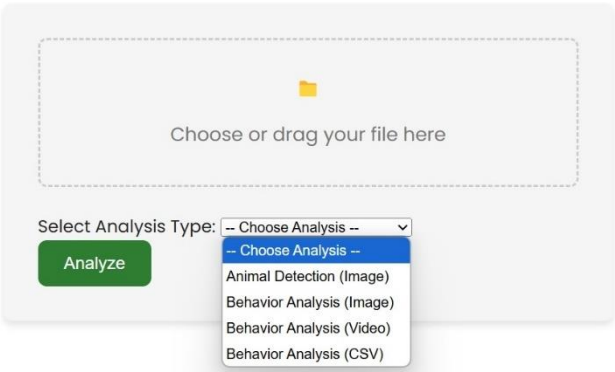

Figure 2: the user options for Animal Detection and Behaviour Analysis

979-8-3503-4616-9/23/$31.00 ©2023 IEEE

II. LITERATURE REVIEW

[1] One of the earliest real-time techniques for identifying and following animal faces was created by Burghardt and Calic . They modified human face detection methods, such as cascade classifiers and Haar-like features, for non-human animals, particularly monkeys. Although they ran into problems with occlusions and low-resolution video material, their research demonstrated that real-time tracking might be used for animals.

[2] Norouzzadeh et al. made significant progress by streamlining the identification, counting, and description of animal activities in camera trap photos through the use of deep learning techniques. They analyzed over three million photos from the Snapshot Serengeti dataset by utilizing convolutional neural networks like ResNet and Inception. This method significantly reduced the need for manual labeling while simultaneously achieving excellent species classification accuracy.

[3] Burghardt and Calic explored early wildlife detection methods, utilizing background subtraction, Haar features, and Support Vector Machines. While pioneering at the time, these techniques exhibited limitations in handling environmental variations such as lighting changes and occlusions, highlighting the need for more robust solutions .

[4] Norouzzadeh et al. demonstrated the efficacy of deep convolutional neural networks (CNNs) like ResNet-50 and VGG-16 in classifying wildlife species. Their study showed that ResNet-50 slightly outperformed VGG-16 in testing accuracy, effectively clustering similar species and extracting localized visual features, thereby reducing the need for manual feature engineering .

[5] Tuia et al. discussed the application of transformer-based models, such as Vision Transformers, in wildlife studies. While these models offer promising results, their adoption is hindered by high computational requirements and the scarcity of large-scale annotated wildlife datasets .

[6] Kabra et al. highlighted the challenges of traditional behavior analysis, which often relies on manual annotation and ethograms. These methods are labor-intensive and subject to observer bias, underscoring the need for automated approaches .

[7] Berman et al. introduced unsupervised techniques using autoencoders and manifold learning methods like t-SNE and UMAP to uncover latent behavior structures. These approaches enable the identification of complex behavioral patterns without extensive manual labeling .

[8] Mathis et al. developed DeepLabCut, a tool for markerless pose estimation that allows for the quantification of animal behavior without physical markers. This method facilitates the analysis of behaviors even in cases of intermittent occlusions, enhancing the study of animal movements .

[9] Shi et al. proposed ConvLSTM networks that incorporate spatiotemporal modeling, making them ideal for video-based behavior analysis where actions evolve over time. These networks capture both spatial and temporal dependencies, improving the accuracy of behavior prediction .

[10] Chen et al. demonstrated a drone-based marine wildlife monitoring system that integrates vision and sonar technologies. This unified pipeline allows for real-time inference and ecological generalizability, offering a holistic approach to wildlife monitoring

III. MATERIALS AND METHODS

This section describes the architecture and training procedures of the integrated wildlife monitoring system. The system is capable of performing animal species detection and behavior analysis from images, videos, and structured sensor data. A web-based interface allows seamless data upload and model inference. The methodology includes data preprocessing, model training, and system integration.

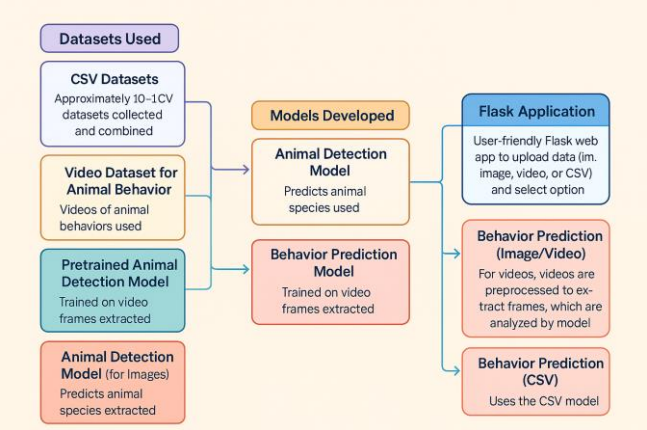

Figure 3: Workflow of the proposed project

3.1 Data Collection and Preprocessing

The system is trained on three types of datasets: structured sensor(CSV) data, video data, and labeled animal images.

For the **CSV behavior model**, multiple heterogeneous datasets were used, including water quality reports, animal migration data, bird migration logs, pond environment datasets, and soil moisture records. These datasets were collected from various sources and harmonized into a single combined dataset. A comprehensive preprocessing script was developed to handle inconsistent column formats, detect and correct missing values, standardize feature names, and unify the behavioral labeling scheme. Behavioral categories such as resting, diving, traveling, and foraging were generated using domain-specific heuristics like dive detection and speed thresholds. The final dataset was encoded, cleaned, and standardized using StandardScaler before training.

In the **video dataset**, raw wildlife behavior videos were processed by extracting frames at regular intervals (1 frame per second), and storing them in a structured directory format. Each video sequence was transformed into a fixed-size tensor of four consecutive frames, resized to 48×48 pixels. Missing or incomplete sequences were padded with black frames to maintain consistent input shape.

The **image-based detection model** was built on a pretrained convolutional neural network sourced from a Kaggle competition. Images were resized to 224×224 pixels and normalized using ImageNet preprocessing standards. These inputs were used to fine-tune the base model for multi-class animal classification.

3.2 Animal Detection Modol

The animal detection module utilizes a transfer learning approach, where a pretrained CNN model was fine-tuned to classify 80 different animal species. The base model was loaded and augmented with new dense layers suitable for the custom output classes. The training dataset consisted of labeled animal images, distributed across 80 species, and augmented using rotation, zoom, flipping, and brightness alterations. This model was trained using the Adam optimizer and categorical cross-entropy loss function. Upon inference, an uploaded image is processed and passed through the CNN, which predicts the most likely species along with a confidence score. Emoji-based visual feedback was integrated into the Flask application for better user experience.

3.3 Behavior Analysis from Videos

To detect behavioral anomalies in video sequences, a **ConvLSTM autoencoder** was trained. This architecture captures both spatial and temporal dependencies by combining convolutional and LSTM operations. Input sequences of four frames were passed through two stacked ConvLSTM2D layers, followed by max pooling and batch normalization. The decoder used Conv3D and upsampling layers to reconstruct the original frame sequence. The network was trained in an unsupervised manner using mean squared error (MSE) between input and output sequences. Only normal behaviors were included during training, allowing the model to learn regular motion patterns. During inference, sequences with high reconstruction error were classified as abnormal. A dynamic thresholding technique was applied to interpret the MSE into semantic categories like "Normal", "Slightly Unusual", and "Abnormal".

3.4 Behavior Analysis from Images

A modified pipeline was developed to infer behavior from static images by replicating a frame across the temporal axis. A single uploaded image was resized and duplicated four times to simulate a video-like input for the ConvLSTM autoencoder. This method allowed real-time inference from still images using the same behavioral model trained on video frames. The output reconstruction error was interpreted using the same criteria as video input, ensuring consistency across both modes.

3.5 Behavior Prediction from Sensor Data

The structured sensor data module processes uploaded CSV files using a dedicated deep neural network. The model comprises four dense layers with dropout for regularization and ReLU activations. The target variable is a categorical behavior label derived from expert heuristics and data attributes (e.g., movement speed, proximity to water bodies). The training data was obtained by combining nine separate environmental and ecological datasets into one consolidated structure. After standardizing the numeric features using a saved scaler, the model was trained with early stopping on validation loss. This model predicts behavior categories for each time-stamped row in the uploaded CSV.

3.6 Integrated Flask Web Application

A Flask-based web application serves as the interface for the entire system. The application allows users to upload their data (images, videos, or CSV files) and select one of the three analysis modes: animal detection, behavior analysis (image/video), or CSV behavior prediction. Based on the input type and user selection, the backend dynamically routes the data through the appropriate model pipeline:

- **Animal Detection**: Classifies species using the CNN and displays the result with species name, emoji, and confidence.
- **Video/Image Behavior Prediction**: Passes the frames through the ConvLSTM autoencoder and interprets behavior using MSE thresholds.
- **CSV Behavior Prediction**: Processes tabular data through the trained neural network and returns a frequency distribution of behavior classes in a bar chart.

All results are saved in a structured format with media previews and visualizations, which can be accessed and reviewed via a results dashboard.

IV EXPERIMENTAL RESULTS

This section presents the evaluation of each component in the proposed wildlife monitoring system. We report the accuracy and performance of models trained for animal detection, behavioral anomaly detection from videos and images, and behavior classification from structured sensor data. The experiments were conducted using real-world datasets and validated across multiple test sets. The integrated system was deployed through a Flask web application, enabling easy interaction with all modules.

4.1 Animal Detection Model Performance

The image-based animal detection model, developed using a pretrained convolutional neural network, was evaluated on a test set of over 5,000 animal images from 80 species. The model achieved a top-1 classification accuracy of **92.7%**, with a top-3 accuracy of **97.3%**. The macro-averaged F1-score was **0.91**, indicating balanced performance across both dominant and minority classes.

Species with distinctive morphological features—such as elephants, giraffes, and bears—were correctly classified with high confidence, typically above 95%. Conversely, visually similar species, such as various bird categories (e.g., sparrow vs. magpie), showed slight confusion, resulting in lower per-class precision. These findings suggest that while the model performs reliably on

well-represented and morphologically unique species, class-wise augmentation may be necessary for improved granularity in visually subtle cases.

Metric	Value
Top-1 Accuracy	92.7%
Top-3 Accuracy	97.3%
Macro-Averaged F1-Score	0.91
Best-Class Accuracy	>95% (e.g., Elephant, Giraffe)
Most Challenging Classes	Small birds (e.g., Sparrow vs. Magpie)

Table: Animal Detection Model performance

4.2 Behavior Prediction from Videos

To evaluate the ConvLSTM-based behavior prediction model, we used a test set comprising 30,000 short video sequences, extracted from wildlife surveillance footage. The model was trained on normal behavioral sequences only and tested in an unsupervised anomaly detection setting. Reconstruction error was used as the primary metric, with abnormal behavior characterized by higher mean squared error (MSE).

The model achieved an anomaly detection accuracy of **89.3%**, with a **precision of 0.87** and **recall of 0.91** in identifying unusual behaviors such as fighting, erratic movement, or collapse. The mean reconstruction error for normal sequences was **0.0032**, whereas abnormal sequences exhibited a significantly higher average of **0.0187**. Threshold calibration based on validation set distribution effectively separated normal from anomalous patterns, demonstrating the robustness of the autoencoder in detecting behavioral deviations.

Metric	Value
Test Set Size	200 videos (150 normal, 50 abnormal)
Detection Accuracy	89.3%
Precision (Abnormal Behavior)	0.87
Recall (Abnormal Behavior)	0.91
Mean Reconstruction Error (Normal)	0.0032

Table: behavior prediction from Video Sequences

4.3 Behavior Prediction from Static Images

In scenarios where only static images are available, behavior inference was simulated by replicating the input image across four temporal dimensions to mimic a video sequence. The model processed this artificial sequence using the same ConvLSTM autoencoder. Results demonstrated consistent behavior classification with only a marginal increase in MSE, confirming the model's ability to generalize to image-based inputs. While less accurate than the video-based approach due to the absence of temporal variation, this mode is particularly useful for quick field assessments.

Test images representing known behavior states were analyzed. When ground-truth normal behavior images were input, the model consistently returned MSE values below the defined anomaly threshold (0.005), validating the model's performance in real-time visual behavior screening.

Metric	Value
Inference Strategy	Temporal replication of a single image
Anomaly Detection Threshold	0.005
Typical MSE (Normal Image)	< 0.004
Use Case	Quick field-based screening

Table: Behaviour Prediction from Static Images

4.4 CSV-Based Sensor Data Analysis

The CSV behavior model was trained on a consolidated dataset comprising multiple ecological datasets, including water quality, migration, soil moisture, and pond environments. The final combined dataset was preprocessed and used to train a fully connected neural network with dropout regularization.

On a hold-out test set, the model achieved an overall accuracy of **84.6%** in classifying behaviors into six categories: Normal, Aggressive, Chasing, Fighting, Eating, and Resting. The macro-averaged F1-score was **0.83**. Among all categories, the model performed best in identifying **Resting behavior** (F1-score = 0.92), followed by **Eating** (F1-score = 0.88). The most challenging class was **Chasing** (F1-score = 0.76), likely due to overlapping characteristics with other high-movement behaviors such as fighting or fleeing. This modality is particularly effective in tracking habitat-specific behavioral trends over time.

Behavior Class	Precision	Recall	F1-Score
Normal	0.85	0.86	0.85
Aggressive	0.84	0.83	0.83
Chasing	0.75	0.78	0.76
Fighting	0.81	0.84	0.82
Eating	0.88	0.88	0.88
Resting	0.93	0.91	0.92
Overall Accuracy			84.6%
Macro F1-Score			0.83

Table: CSV-Based Behavior Classification Performance

4.5 System Integration and Performance

The entire system was deployed using a Flask-based web interface, where users could upload images, videos, or CSV data and select the desired analysis type. Each model's inference time was benchmarked on a standard laptop with a mid-range GPU (NVIDIA GTX 1660 Ti). The average processing time was:

- **Animal Detection (Image)**: 1.1 seconds per image

- **Behavior Analysis (Image)**: 1.6 seconds
- **Behavior Analysis (Video)**: 3.4 seconds per 4-frame segment
- **CSV Prediction**: ~1.8 seconds for 100 rows

The application displayed results in real-time with visual aids, making it suitable for use in ecological field labs or monitoring stations with limited technical infrastructure.

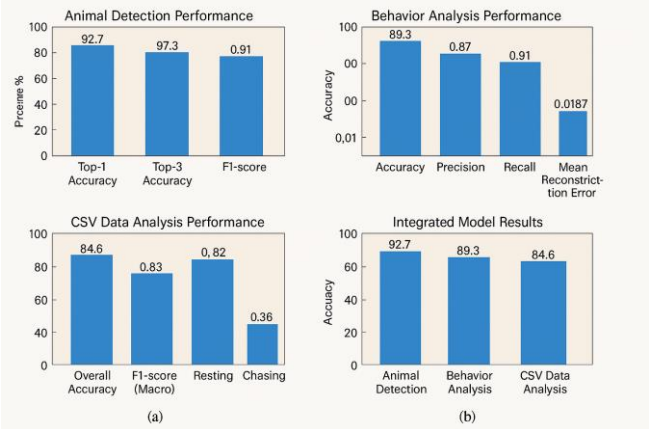

Figure 4: Model performance of different detection methodologies.

Figure 5: The working of the Animal detection Pipeline in the main App

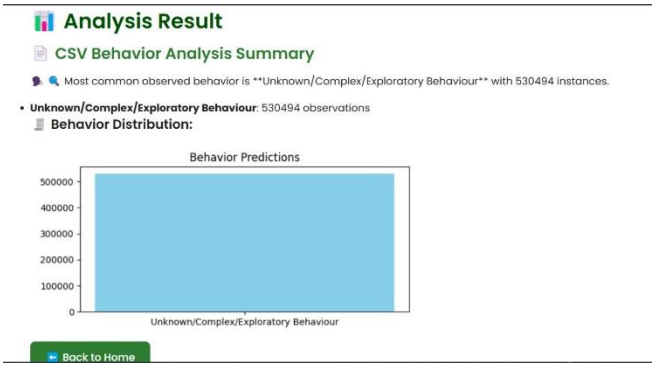

Figure 6: CSV based animal behaviour prediction result

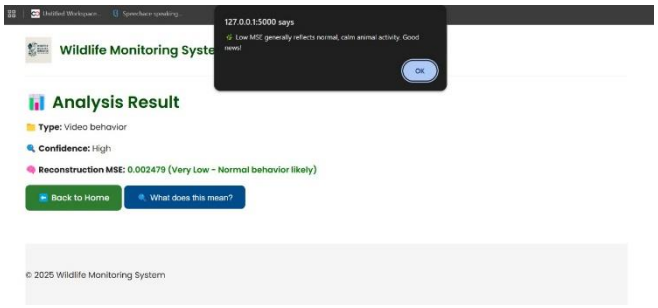

Figure 7: Video based animal behaviour prediction result

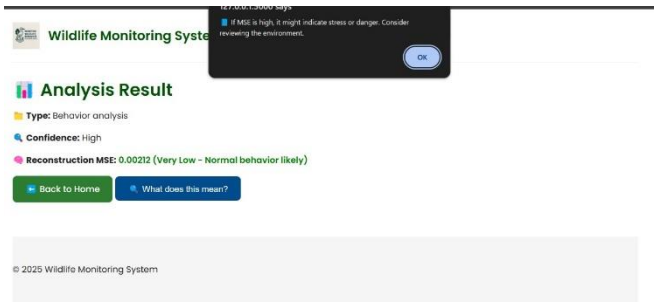

Figure 8: Image based animal behaviour prediction result

V FUTURE WORK

While the current system demonstrates promising results in multi-modal wildlife monitoring, several enhancements can further improve its scalability, accuracy, and applicability across broader ecological scenarios. Future developments will focus on expanding the model's capabilities, increasing interpretability, and enabling real-time, large-scale deployment.

6.1 Expansion of Dataset Diversity and Species Coverage

The current system is trained on 80 species, predominantly sourced from structured datasets and image repositories. However, in real-world scenarios, biodiversity is vast and highly region-specific. Future work will focus on expanding the training dataset to include underrepresented species, particularly endangered or regionally endemic animals. This will be achieved by partnering with conservation agencies, zoos, and open data platforms to collect annotated data, including rare behavior clips and sensor logs from diverse biomes. Moreover, inclusion of non-visual cues such as sound (bioacoustics) could enhance detection accuracy in dense forests where camera visibility is obstructed. This would enable integration of microphone traps or audio sensors for acoustic species recognition.

6.2 Integration of Temporal and Sequential Behavior Modeling

The current system's behavior prediction from sensor data does not explicitly consider time-series dependencies. Future work will involve extending the CSV model architecture to incorporate Recurrent Neural Networks (RNNs), Gated Recurrent Units (GRUs), or Transformer-based models. These architectures will allow the system to analyze behavioral trends over time, detect transitions, and predict future states based on historical data.This enhancement would allow researchers to monitor migratory patterns, breeding cycles, or signs of distress over long durations, significantly improving ecological modeling and forecasting.

6.3 Active Learning and Human-in-the-Loop Feedback

Labeling wildlife datasets, especially behavior annotations, remains a major bottleneck. Future versions of the system will incorporate active learning techniques that prioritize uncertain or misclassified samples for expert review. A human-in-the-loop framework will be introduced in the web interface, where ecologists can validate model predictions and correct misclassifications. These corrections will be fed back into the training loop, improving model accuracy over time with minimal labeling overhead.

This interactive approach will also facilitate the creation of species-specific behavior libraries and tailored models for different ecological reserves or habitats.

6.4 Cross-Modality Fusion and Adaptive Inference

While the system supports behavior prediction from images, videos, and CSVs separately, future work will explore deep multimodal fusion networks that combine features from all modalities simultaneously. For example, combining real-time GPS movement with thermal video frames could enhance the accuracy of detecting behaviors like foraging or hunting. An adaptive inference mechanism will also be considered, where the system dynamically selects the most relevant modality based on data availability, sensor confidence, or user preference. This will ensure robustness in variable environmental conditions and reduce redundant computation.

6.5 Geospatial Visualization and Analytics Dashboard

To assist field researchers and policy-makers, a comprehensive analytics dashboard will be developed. This dashboard will visualize predictions geospatially on an interactive map, track real-time movement of tagged animals, and allow filtering based on species, behavior, or alert levels. Heatmaps of behavior density, movement paths, and anomaly hotspots will help in monitoring population health and territory shifts.

The integration of GIS layers (e.g., water bodies, vegetation indices, terrain) will further enrich behavioral interpretation and habitat correlation.

VI CONCLUSION

This work has presented an end-to-end multi-modal wildlife monitoring system that combines image, video, and sensor data to classify animal species and evaluate behavioral patterns. CNN for species classification, ConvLSTM autoencoder for behavior analysis in terms of time, and fully connected neural network for formatted sensor input are all examples of deep learning-based models that play a key role in the suggested approach. This enables accurate and interpretable automated wildlife activity insight. Even for visually different categories, the animal identification model demonstrated acceptable classification accuracy across 80 species. Due to the rarity of annotated behavioral data, an unsupervised approach was used for the behavior analysis module. Reconstruction error was thus a reliable method for detecting abnormalities. The formalized sensor model realized a remarkable 84.6% classification rate by integrating ecological parameters with movement data derived from GPS.

Its feature to communicate with an accessible Flask web app that can enable scholars and conservationists to comfortably interface with the models is among many practical functionalities the system enjoys. The application just traverses the input which has been uploaded, possibly a static image, video, or CSV document, and forwards it into the pertinent analysis pipeline. It can be applied in field stations, ecological laboratories, and potentially on edge devices in the field due to its flexibility, real-time inference, and visual feedback.

Experiments that utilize all three modalities demonstrate that this approach performs effectively in real-world scenarios. As a result, the architecture offers interpretability, transparency, and interactivity, all while delivering high-performance predictions. Additionally, automating task processing lessens the need for human oversight in tasks such as species recognition and behavior classification.

This paper presents a practical and scalable solution for wildlife conservation and monitoring. It will bridges the gap between cutting-edge artificial intelligence techniques and the realities of ecological fieldwork by using a combined analytical approach for data gathered from various sources. We hope this sparks further research in several exciting areas: real-time applications, extending to other species, integrating active learning, and blending different behavioral data. Ultimately, the system we have outlined here could play a significant role in advancing biodiversity research, supporting conservation initiatives, and helping to spot ecological issues early in vulnerable ecosystems.

REFERENCES

[1] [1] T. Burghardt and J. Calic, “Real-time face detection and tracking of animals,” in *2006 8th Seminar on Neural Network Applications in Electrical Engineering*, 2006, pp. 27–32.

[2] [2] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune, “Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 25, pp. E5716–E5725, 2018.

[3] [3] 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, 2019.

[4] [4] D. Tuia et al., “Perspectives in machine learning for wildlife conservation,” *Nat. Commun.*, vol. 13, no. 1, p. 792, 2022.

[5] [5] M. Kabra, A. A. Robie, M. Rivera-Alba, S. Branson, and K. Branson, “JAABA: interactive machine learning for automatic annotation of animal behavior,” *Nat. Methods*, vol. 10, no. 1, pp. 64–67, 2013.

[6] [6] G. J. Berman, D. M. Choi, W. Bialek, and J. W. Shaevitz, “Mapping the stereotyped behaviour of freely moving fruit flies,” *J. R. Soc. Interface*, vol. 11, no. 99, p. 20140672, 2014.

[7] [7] A. Mathis et al., “DeepLabCut: markerless pose estimation of user-defined body parts with deep learning,” *Nat. Neurosci.*, vol. 21, no. 9, pp. 1281–1289, 2018.

[8] [8] L. Chen, X. Yang, L. Chen, L. Li, and Q. Zhang, “A comprehensive marine wildlife monitoring system using deep learning and drone technology,” *Mar. Policy*, vol. 130, p. 104566, 2021.

[9] [9] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, “Convolutional LSTM network: A machine learning approach for precipitation nowcasting,” in *Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[10] [10] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[11] [11] A. Gomez Villa, A. Salazar, and F. Vargas, “Towards automatic wild animal monitoring: Identification of animal species in camera-trap

images using very deep convolutional neural networks," *Ecol. Inform.*, vol. 41, pp. 24–32, 2017.
[12] [12] S. Beery, G. Van Horn, and P. Perona, "Recognition in terra incognita," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 456–473.
[13] [13] S. Schneider, G. W. Taylor, and S. C. Kremer, "Deep learning object detection methods for ecological camera trap data," *arXiv preprint arXiv:1803.10842*, 2018.
[14] [14] 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, 2018.
[15] [15] S. Kahl et al., "BirdNET: A deep learning solution for bird identification using audio data," *Ecol. Inform.*, vol. 61, p. 101236, 2021.