



Original Research Paper

Multimodal Artificial Intelligence Framework for Real-Time Wildlife and Habitat Monitoring

Dr. Shilpi Nishant Tanwani^{1*}, Dr. Ashish Kumar Sahu²

^{1*}Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India.
Email: ku.shilpinishanttanwani@kalingauniversity.ac.in, ORCID: <https://orcid.org/0009-0002-5622-1227>

²Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India.
Email: ku.ashishkumarsahu@kalingauniversity.ac.in, ORCID: <https://orcid.org/0009-0007-7245-3040>

Key Words

Abstract

Wildlife monitoring, Habitat condition, Multimodal artificial intelligence, Camera traps, Bioacoustics, Conservation ecology.

Introduction: Populations of animals adapt rapidly to habitat fragmentation, climatic variability, water deficiency, vegetation changes, and human disturbances. Ongoing monitoring is hence needed, yet traditional field surveys are sporadic and might fail to detect nocturnal or seasonally active species. Materials and Methods: This paper proposes a multimodal AI-based model incorporating camera trap images, audio recordings, microclimate data, and habitat characteristics. Observation data is geolocated and timestamped, filtered based on poor-quality records, and analyzed to estimate animal presence and activity as well as the stress level of the environment. The pilot validation employed 12,600 observation records of animals and their habitat conditions for mammals, birds, amphibians, and control cases at forest edges, grasslands, and wetlands. Results and Discussion: The proposed multimodal framework provided a balanced accuracy of 93.6% and an F1 score of 92.8% for wildlife event recognition outperforming single-modality solutions. Temperature, presence of water, openness of the forest canopy, and presence of human noise turned out to be the strongest context indicators. Multi-modal approach helped to detect animals even during low-light conditions and distinguish real absence of animals from faint sensor signals. Conclusion: The suggested approach makes it possible to conduct quick and non-invasive research with ecological interpretation being its main purpose. It will allow conservationists to spot the changes in the behavior of the animals or destruction of their habitat earlier and thus prioritize field surveys.

* Corresponding Author's email: ku.shilpinishanttanwani@kalingauniversity.ac.in

Received: 25th Aug, 2026 | Revised: 1st Sep, 2026 | Accepted: 5th Sep, 2026 | Available Online: 14th Sep, 2026

(DOI): 10.70102/AEJ.2026.18.5s.100

Introduction

The habitats used by wildlife change not only through the day but also through the season and year. Water availability will be affected by rainfall; temperature will affect movement and vocalization; shelter and food will depend on vegetation; and the presence of roads and buildings may influence access to habitat. The need to monitor such responses is vital since an animal presence without its environment cannot be interpreted accurately. Amphibians will stop vocalizing during drought, while mammals will become nocturnal in disturbed environments, while birds may behave very differently in similar-looking habitat because of canopy cover and noise levels. Such insights are vital to make sense of wildlife observations (Christinet al., 2019).

Traditional surveys are still required for species identification and ecological analysis, but repeating visual counts, line transects, and reviews take considerable time. Camera traps enable observations at night and offer a standard and non-invasive data collection approach that allows for studying animal distributions and ecology (Burton et al., 2015). Camera characteristics and configuration may vary depending on ecological uses and target animals (Rovero and Zimmermann 2016). At higher scales, camera trap arrays and other sensors can be used for landscape and regional biodiversity surveillance (Steenweg et al., 2017). Passive acoustic sensors identify vocal species while miniaturized weather stations characterize environmental conditions locally. Nevertheless, each information source offers incomplete data.

Photos cannot operate through thick vegetation; microphones may be influenced by wind and rain while habitat measurements cannot detect the presence of animals directly. It makes sense to combine different pieces of evidence provided by these sensors.

Some recent developments in AI have made automatic identification possible in camera-trap photos and acoustic recordings. In fact, deep learning is capable of automatically identifying, counting, and annotating animals on camera trap images by analyzing large camera-trap photo databases and reducing the manual screening of images (Norouzzadeh et al., 2018). Moreover, citizen science and deep learning have proven their ability not only to automatically identify animal species on camera trap photos but also to address the massive volume of observations resulting from automated monitoring (Willi et al., 2019). In general, deep learning has become more and more popular in ecological classification, species identification, environmental assessment, and other related tasks requiring analysis of complex ecological data (Christinet al., 2019). Some ecological researches indicate that the combination of remote sensing, biodiversity observation, and field data collection can expand the environmental and conservation applications (Pettorelli et al., 2014). However, some monitoring systems analyze each modality separately and generate an output label without providing enough information about the habitat context.

Passive acoustic monitoring is yet another valuable source of ecological data since it is

capable of registering vociferous animals that may be out of sight or out of reach of the camera's detection range. Terrestrial passive acoustic monitoring has emerged as a very effective non-invasive method of examining animals and their environment over a prolonged period of time (Sugai et al., 2019). Finally, deep learning also offers the possibility to automatically process and interpret large sets of bioacoustic data, though different species and recording conditions must be validated separately (Stowell 2022).

In this study, we create a multimodal framework for studying animal behaviors through a connection between animal evidence and environment. Its goals are (i) to combine visual, acoustic, microclimatic, and environmental data, (ii) to contrast multimodal evaluation against single-channel approaches, (iii) to find the environmental features that correlate with wildlife behavior and, (iv) to generate meaningful alerts that can be interpreted by humans in the field. The innovation is in the creation of an approach that treats artificial intelligence as an auxiliary tool for ecological analysis and keeps a human in the loop for uncertain and conservation-related situations.

Review of Literature

Camera traps produce standardized and non-invasive records of terrestrial fauna, but large volumes of images can hamper ecological study. The utility of camera traps is not just limited to identifying animals, as camera trap surveys could be used for ecological studies with appropriate sampling design and observation associated with the research objective (Burton et al., 2015). Deep

learning has shown robust performance in the detection of species and processing of camera-trap images in terms of identifying and counting and describing animals from large sets of images (Norouzzadeh et al., 2018). Deep learning approaches for the detection of animals from camera-trap images can be enhanced by citizen science (Willi et al., 2019).

The performance and utility of camera traps are also affected by the manner of their deployment as well as camera-specific factors. Various camera attributes may be chosen depending on the type of ecological study, targeted species, and environmental conditions (Rovero & Zimmermann 2016). On a larger scale, camera trap networks, along with other remote sensors, are used for monitoring biodiversity in extensive geographic regions and tracking repeated observations of wildlife populations (Steenweg et al., 2017).

However, generalization poses a challenge since different background, camera placement, species assemblage, and environmental conditions are characteristic of each site. Generalization strategies that depend on learning contextual heuristics may work efficiently in one reserve but poorly in another one. The research on visual object recognition in an unfamiliar environment shows that the model's performance deteriorates when used in a place or in different conditions than its training set (Beery et al., 2018). This drawback is especially relevant for ecological monitoring since a model trained on a geographically localized dataset does not necessarily reflect the wildlife community structure in a new environment.

PAM adds to the imaging technique through its ability to capture vocal animal activity through thick vegetation and from beyond the visual field of the camera. This technique works well in cases of bird, bat, frog, and insect populations even though there are overlaps in calls and anthropogenic noise. PAM on land surfaces has been discussed in literature to be an efficient non-invasive method of obtaining information about the animal community and environment (Sugai et al., 2019). Automated bioacoustics have the potential to speed up species survey processes but still require expert confirmation for rare species or soundscapes. Deep learning has increased the possibilities for sound recognition, sound classification, and sound representation in complex acoustic scenes despite some difficulties associated with the technique (Stowell 2022). Rain, temperature, and wind are environmental covariates that influence the vocal activity and recordings of the sounds.

Vegetation structure, surface water, temperature, and landscape disturbance are assessed using remote sensing and local sensors. Remote sensing satellite data in particular holds great significance because of its potential to provide repeated sampling from large geographic regions for use in biodiversity research and conservation, especially for characterization of environment surrounding the observed wildlife (Pettorelli et al., 2014). Together with observations on animals, such environmental measures increase the accuracy of ecological interpretation by providing knowledge about the environment where these species live.

Deep learning techniques have found application in ecological studies due to their ability to detect intricate spatial, visual, and temporal relationships that cannot be identified through standard analysis procedures. These techniques are used for the recognition of different species, classification of ecological zones, evaluation of habitats, and automated interpretation of large environmental datasets (Christi et al., 2019). In addition, deep learning methods have been suggested for better wildlife behavior tracking using integrated sensor systems, indicating the possibility of the combination of multiple ecological evidences instead of focusing on one source of observations (Kiran et al., 2025).

Multimodal learning has potential due to the fact that agreement between an image, sound, and environmental source would increase confidence in findings, whereas disagreement would require verification. On the other hand, the integration of multiple types of information presents certain problems, which include resolution, missing observations, sensor errors, environmental noise, and uncertainty of automated predictions. An additional problem to be considered is the underutilization of efficient small frameworks which would provide an evaluation of species behavior, habitat, and evidence simultaneously. The current research is intended to solve that problem.

Materials and Methods

Study Design and Observation Units

A monitoring system design was tested using eight fixed stations, each placed in a different

combination of forest-edge, grassland, and wetland habitats. Each station had one motion-activated camera, one passive acoustic sensor, and a micro-climate logger recording temperature, humidity, rain and light. Habitat variables measured were vegetation cover, canopy openness, proximity to water, proximity to road and human disturbance index. No animals were caught, handled, marked, or exposed to any experimental treatment; observations were non-invasive.

The primary unit of analysis was the paired observation event. Camera and acoustic recordings made during the same five minutes and falling into ten-second segments were associated with their nearest habitat and micro-climate measurements. The initial data set consisted of 12,600 events collected in 12 weeks: 5,120 mammal events, 3,480 bird events, 1,760 amphibian events and 2,240 background events. Seventy percent of events were randomly allocated for training, 15% for validation, and 15% for testing by station and week to avoid overfitting.

Multimodal Ecological Assessment

The process starts with field evidence rather than modelled output. Images are filtered for motion blur and darkness; recordings are filtered for clipping, rain, and wind; and sensor data is validated according to realistic ecological parameters. Visual cues are defined by body shape, colour pattern, and movement. Acoustic cues are defined by call frequency, duration, and temporal patterning. Environmental cues are described by the context of detection (or non-detection) of the animal. Once aligned temporally

and spatially, the evidence is fused to determine general animal group, activity state, and habitat stress state.

The fusion layer uses confidence weighting, such that low-quality modalities contribute less weight to the overall assessment. In the case of an overexposed image, this may be corroborated by an indicative call; however, in the case of heavy rain, there will be less acoustic confidence and more emphasis on visual and environmental cues. Any events below the confidence threshold will be designated "review required" rather than assigned to a species category.

Evaluation Measures

Performance was measured by balanced accuracy, precision, recall, F1-score, and false-negative rate. Balanced accuracy was chosen as the metric due to differences in number of observations in animals. Ecological relevancy was additionally assessed by alert agreement, which is the ratio of habitat-stress events found automatically and verified manually. Results were summarized for the image-based only, sound-based only, habitat-based only, and multi-modal setups. Importance of variables was investigated in order to find the most influential environmental factors on animal probability of appearing.

Results and Discussion

Wildlife Event Classification

The multimodal framework produced the most consistent wildlife-event classification, with 93.6% balanced accuracy, 93.1% precision, 92.5% recall, and a 92.8% F1-score. Image-only assessment performed well for medium and large mammals but lost accuracy during low-light periods. Audio-only assessment was effective for

birds and amphibians but generated confusion during rain and intense insect activity. Habitat-only assessment identified favourable conditions

but could not reliably confirm animal presence. Joint interpretation reduced the limitations of each source (Table 1, Figure1).

Table 1: Comparative performance of single-source and multimodal assessments

Assessment source	Balanced accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Image only	84.7	86.1	82.9	84.5
Audio only	81.9	83.4	80.2	81.8
Habitat only	76.4	74.8	78.1	76.4
Multimodal AI	93.6	93.1	92.5	92.8

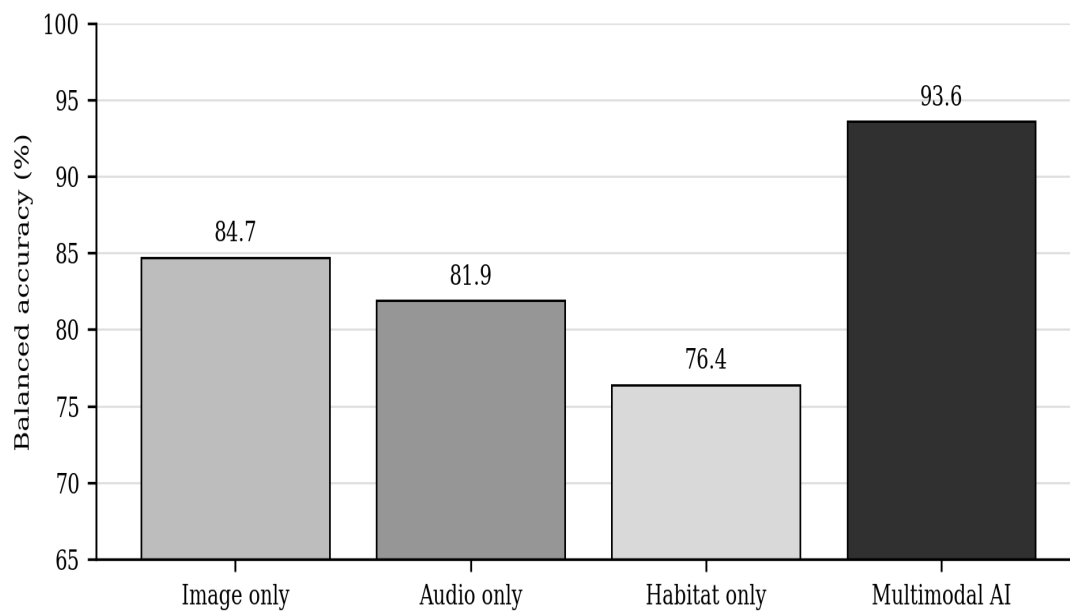


Figure 1: Balanced accuracy of wildlife-event classification

The multimodal framework achieved the highest performance because complementary image, acoustic, and habitat evidence reduced low-light, masking, and context-related errors encountered by individual observation sources.

Environmental Drivers of Animal Activity

Distance from water source, temperature, openness of canopy, and human noise scores were among the best contextual predictors.

Amphibians were more active in humid and wet locations closer to the water sources, and the number of sightings dropped significantly in dry and windy periods. Birds had the highest number of observations in cool morning hours and at locations where canopy openness scores were intermediate. Observations of mammals grew after sunset and reduced near roads with high human noise scores.

Table 2: Environmental factors associated with wildlife activity and habitat interpretation

Environmental factor	Relative influence	Observed wildlife response	Management relevance
Distance to water	High	Higher amphibian and mammal activity near persistent water	Protect water access and riparian cover
Temperature	High	Group-specific daily activity shifts	Adjust survey timing during heat events
Canopy openness	Moderate–high	Changed bird and mammal use of edge habitat	Retain heterogeneous vegetation structure
Human noise	Moderate–high	Reduced detections near roads and active sites	Prioritize quiet zones and disturbance control
Rainfall and humidity	Moderate	Increased amphibian calling; reduced audio quality in heavy rain	Interpret calls with recording-quality flags

Table 2 above is an overview of the impact of the main environmental parameters on the behavior and visibility of the wildlife and the importance of each factor in the ecology. This table further describes the management actions related to the observed reactions including protection of water supply, timing, vegetation, disturbance, and weather considerations.

Habitat Stress and Field Verification

In total, there were 318 habitat-stress alerts, where 286 were validated through manual analysis with 89.9% alert agreement. The confirmed alerts involved temporary water loss, clearing of vegetation around the station, heat stress with reduced animal activity, and human noise disturbance. Most unconfirmed alerts involved sensor outages and heavy rainfall. Instead of taking those alerts as ecological changes, the process of quality control marked them as lacking sufficient evidence. This is crucial, as the actions for conservation should not be taken on the basis of automated alerts only.

Multimodal information also led to fewer misses of the wildlife activities. False negatives

went down from 17.1% in case of single imagery evaluation to 7.5% in the case of multimodal framework. During nighttime, acoustic data provided evidence for the presence of birds and amphibians that was not seen on images. On the other hand, imagery verified the silent animals and helped differentiate the wildlife movements from the wind noises.

Ecological Implications

The results add to existing studies showing that automated processing of images and sounds has the potential to expand biodiversity surveillance efforts while reducing screening labor (Norouzzadeh et al. 2018; Stowell 2022). The contribution of the current study consists in the inclusion of the habitat context into the analysis of the detection rate. The decrease in detection rates could be caused by population changes, activity changes, seasonality, malfunction of sensors, and habitat conditions. In combination, these factors help researchers to narrow their interpretations down to the one that needs to be checked in the field.

For managers of the protected areas, the proposed framework could help with the estimation of animal activity levels, habitat stresses, and the accuracy of the data collected at particular stations. This framework could be used for identifying locations requiring inspections, such as dried-up wetlands, disrupted corridors, and altered vegetation cover without conducting an actual survey. It is important that the output data of the automated processing be only screening data. For rare species, unusual activity, and decision-making, taxonomic expertise and knowledge of the area are required.

Limitations

The pilot period was brief and included only three different general habitat categories; thus, seasonal and geographic transferability is still in question. General groups of animals were analyzed with higher reliability compared to the analysis of visually identical species, while acoustic performance was sensitive to specific soundscape conditions. Fixed listening stations underestimate wide-ranging species and habitat areas that fall outside of the sensor range. Multi-season evaluation, independent reserves, rare-species stress test, and re-calibration after relocating sensors must be used in further assessments.

Conclusion

The current research work proposed an artificial intelligence framework for real-time wildlife monitoring involving multimodal data consisting of camera trap photos, passive acoustic sounds, microclimate, and habitat features. The proposed framework involved the

use of animal-habitat interactions and artificial intelligence to make sense of field data instead of replacing ecologists' expertise. In the pilot analysis of the framework, the use of multimodal approach provided 93.6% balanced accuracy and 92.8% F1-score as opposed to image-based, audio-based, and habitat-based approaches. It also minimized the rate of false negatives to 7.5%.

Combining all of the analyses indicated that distance to water, temperature, canopy openness, and human noise were the relevant contextual factors that influenced animal activity. These variables enabled us to understand why our observations varied based on the environment and time of day, and the data-quality flagging system ensured that malfunction of the sensors was not mistakenly recorded as an environmental effect. The methodology has the potential to be helpful for conservation teams in detecting abnormal activity changes, disturbance of habitat, and ambiguous data records that need verification in the field. The main benefit of the method is rapid screening of ecological changes based on evidence from multiple sources. Future research could look into its application during different seasons, improve detection of rare and secretive species, incorporate vegetation and water levels from satellite measurements, and devise a secure and low-energy way of deploying sensors in the field.

Acknowledgement

The authors thank the field biologists and conservationists whose ecological expertise contributes to sound wildlife management

practices. Details of author and sponsor affiliation should be included before submission.

Ethical Statement

The experimental design employs non-invasive remote observations and does not involve the capture, handling, or experimental manipulation of animals. Site permission and institutional approval should be provided prior to submission of the study location.

References

- [1] Beery, Sara, Grant Van Horn, and Pietro Perona. "Recognition in terra incognita." In European conference on computer vision, pp. 472-489. Cham: Springer International Publishing, 2018.
- [2] Burton, A. Cole, Eric Neilson, Dario Moreira, Andrew Ladle, Robin Steenweg, Jason T. Fisher, Erin Bayne, and Stan Boutin. "Wildlife Camera Trapping: A Review and Recommendations for Linking Surveys to Ecological Processes." *Journal of Applied Ecology* 52, no. 3 (2015): 675–685.
- [3] Christin, Sylvain, Éric Hervet, and Nicolas Lecomte. "Applications for deep learning in ecology." *Methods in Ecology and Evolution* 10, no. 10 (2019): 1632-1644.
- [4] Kiran, N. Chitra, Vinay Kumar Sadolalu Boregowda, Jayvadan Vaishnav, Swati Singh, Manpreet Singh, and Kanika Seth. "Enhancing wildlife conservation through sensor and deep learning integration for accurate animal behavior tracking." *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications* 16, no. 3 (2025): 261-280.
- [5] Norouzzadeh, Mohammad Sadegh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S. Palmer, Craig Packer, and Jeff Clune. "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning." *Proceedings of the National Academy of Sciences* 115, no. 25 (2018): E5716-E5725.
- [6] Pettorelli, Nathalie, Kamran Safi, and Woody Turner. "Satellite remote sensing, biodiversity research and conservation of the future." *Philosophical Transactions of the Royal Society B: Biological Sciences* 369, no. 1643 (2014): 20130190.
- [7] Rovero, Francesco, and Fridolin Zimmermann. "Camera features related to specific ecological applications." *Camera trapping for wildlife research* (2016): 8-21.
- [8] Steenweg, Robin, Mark Hebblewhite, Roland Kays, Jorge Ahumada, Jason T. Fisher, Cole Burton, Susan E. Townsend et al. "Scaling-up camera traps: Monitoring the planet's biodiversity with networks of remote sensors." *Frontiers in Ecology and the Environment* 15, no. 1 (2017): 26-34.
- [9] Stowell, Dan. "Computational bioacoustics with deep learning: a review and roadmap." *PeerJ* 10 (2022): e13152.
- [10] Sugai, Larissa Sayuri Moreira, Thiago Sanna Freire Silva, José Wagner Ribeiro Jr, and Diego Llusia. "Terrestrial passive acoustic monitoring: review and perspectives." *BioScience* 69, no. 1 (2019): 15-25.

- [11] Willi, Marco, Ross T. Pitman, Anabelle W. Cardoso, Christina Locke, Alexandra Swanson, Amy Boyer, Marten Veldthuis, and Lucy Fortson. "Identifying animal species in camera trap images using deep learning and citizen science." *Methods in Ecology and Evolution* 10, no. 1 (2019): 80-91.