

Recent Advances in AI-Based Wildlife Monitoring: A Review of Multimodal Detection, Edge Computing, and Conservation Applications

Mrs. Pradnya Kasture¹, Jidnyasa Kale², Neha Patil³, Vaishnavi Karle⁴, Aayush Bobade⁵

¹Computer Engineering & RMDSTIC

²Computer Engineering & RMDSTIC

³Computer Engineering & RMDSTIC

⁴Computer Engineering & RMDSTIC

⁵Computer Engineering & RMDSTIC

Abstract - The rising level of human-wildlife conflict and the desperate necessity to preserve biodiversity have established high demand of intelligent, automated wildlife monitoring systems. This review assesses recent developments in artificial intelligence used in ecological informatics, and its application in the integration of computer vision, bioacoustics, Internet of Things (IoT), and edge computing technologies. The relevant literature on animal detection by images, automation of camera traps, species recognition, sound-based recognition, intrusion warning systems, and low-power deployment devices were examined. Popular ones are YOLO, convolutional neural networks, EfficientNet, transfer learning, Wav2Vec2 transformers, and sensor-based alert systems. Some studies have reported detection rates of more than 90 percent and the lightweight models have enhanced speed, less memory and can be deployed in remote environments. The results indicate that visual deep learning models are highly accurate in detecting, counting, and identifying species under controlled environments but fail to work effectively in nocturnal scenes, camouflage, occlusion, dense vegetation, and so on. Conversely, hybrid systems that integrate image data, acoustic sensing as well as real-time alerts contribute significantly to detection reliability. Audio models based on transformers, and few-shot learning methods, also demonstrate good potential in recognizing rare species with sparse information. This review suggests that multimodal sensor fusion, predictive analytics, and energy efficient edge processing should be the focus of future systems to enable sustainable wildlife monitoring.

Key Words: Wildlife monitoring; human-wildlife conflict; artificial intelligence; deep learning; computer vision; bioacoustics; Internet of Things (IoT); edge computing; YOLO; conservation technology.

1.INTRODUCTION

Human-wildlife conflict (HWC) has emerged as a major global issue due to the rapid urbanization, deforestation, agricultural development and climate change. With the shrinkage of the natural habitats, the breaking up of the ecological boundaries, the wild animals are increasingly making their way into villages, farms, and peri-urban areas in search of food, water, or migration routes. Such interactions have been related to crop destruction, livestock loss, property destruction, human injuries and death of animals, which have caused drastic ecological, social, and economic repercussions [1]. The

main species of these conflicts are often the elephants, tigers, leopards, bears, wolves and wild boars, particularly the developing regions that are rich in biodiversity with human settlements found in and near forests and other protective areas [2]. At the same time, the loss in biodiversity has been growing at a very high rate in the world as most of the species are under threat of being poached, habitat destroyed, illegal trade, and environmental stress. Proper continued and large scale monitoring systems capable of identifying the presence of animals, their movements and behaviors in real time are thus proposed to be the effective conservation of wildlife [3]. The traditional techniques such as hands-on patrolling, camera surveillance, collar tagging, fencing, slow walking or even difficult to navigate over extensive areas include manual patrolling, camera observation, collar tagging, fencing, slow walking or even difficult to navigate over large areas [4]. The inaccessibility and infrastructures in remote areas further reduce the performance of conventional monitoring systems [5]. Recent advancements in Artificial Intelligence (AI) have opened up new possibilities of ecological informatics and conservation of wildlife. Automated animal recognition using images, videos, sounds and sensor data have now become possible with the help of deep learning, computer vision, bioacoustics, Internet of Things (IoT) and edge computing technologies [6]. Cameras with in-built traps and AI applications can automatically classify species, as well as counting animals and identifying behavioral patterns automatically. Similarly, acoustic monitoring systems can determine species-specific calls and are able to detect hidden or nocturnal animals in dense forests where visual monitoring systems may not be effective [7]. The alert systems based on the IoT could be integrated to provide real-time alerts to the residents of the villages or the forest authorities in case the dangerous animals are on their way to the human settlements [8]. Computer vision methods, such as convolutional neural network (CNN) and models of object detectors, including YOLO (You Only Look Once) have shown to perform well in wildlife image recognition. It has been reported that through camera trap or surveillance pictures, high detection rate of species such as lions, tigers, elephants, bears and leopards has been reported [9]. The application of EfficientNet, ResNet and other such architectures as the transfer learning algorithms has also improved the performance in the cases where there is a lack of labelled and unlabeled

data [10]. Lightweight detection models have also been developed to be deployed in remote sites without having to rely on the cloud [11]. Simultaneously, bioacoustic monitoring has also gained significance as a non-invasive instrument of biodiversity measurement. The vocalizations of animals present important data about the presence of the species, its density, its migration, reproduction as well as the stressful conditions. Recent transformer-based models such as Wav2Vec2 and few-shot learning models have reported good performance in recognizing species with a limited amount of audio information even in noisy field settings [12]. These are particularly effective in the process of locating elusive, nocturnal or camouflaged species that are difficult to observe [13]. Although these improvements have been made, there are still significant issues. Under low light conditions, thick vegetation, camouflage, motion blur, and animals in the same place may cause visual models to feel a reduced degree of accuracy. Acoustic systems may be influenced by the wind, rain, anthropogenic noise and overlaps of calls [14]. In addition, majority of ecological data sets have been characterized by a class imbalance, no annotations, domain shift across habitats as well as low availability of rare species data [15]. The next existing problem is whether a unimodal system (vision only or audio-only) suffices to give the field a reputable performance or whether a multimodal system (cameras, microphones, thermal sensors and environmental signals) is required [16]. As a growing body of evidence emerges, showing that multimodal approaches can also be employed to achieve robustness, there is also an increasing body of evidence that multimodal approaches can be used to achieve robustness too, but it also comes at the cost of increased cost and complexity of the system. The issue of energy consumption and sustainability is another burning issue. The use of wildlife monitoring devices is mostly used in remote locations that use battery or solar panels as a power source. The large computing capacity of existing AI models can reduce the operating time and increase the maintenance cost. Accordingly, model compression, efficient inference, and edge processing approaches have entered the limelight of the researchers interested in the balancing of accuracy and energy efficiency [17]. Therefore, the main aim of the review is to compare and contrast the recent advances in AI-based wildlife surveillance systems, comprising of image-based detection, bioacoustic analysis, IoT enhanced alert system, and energy-efficient edge computing solutions. This paper introduces the key methodologies, evaluates the effectiveness of the methods, walks gaps in the research, and discusses the future perspectives. Contrary to other past surveys, this review provides a single coherent analysis of multimodal and energy-efficient intelligent systems and the main conclusion is that multimodal and energy-efficient intelligent systems are needed to provide reliable, scalable and real time monitoring of wildlife and ultimately lead to sustainable conservation and reduction of risks to human communities [18].

2. Materials and Methods

2.1. Datasets and Data Acquisition

The methodologies which have been evaluated in this review are founded on a broad, and huge datasets in both visual and acoustic modalities. In the Snapshot Serengeti dataset, which has approximately 3.2 million labeled images of 48 wildlife species, there were large repositories in the case of visual detection. Custom datasets were also utilized, which contained 8,250 images of elephants and 1,619 annotated samples, which represented such species as lions, tigers, bears, and leopards. These datasets can be used to perform object detection, classification, counting, and behavioral analysis activities. Curated bioacoustic data sets were used to monitor acoustically, and these data sets included recordings of approximately 15 different species of birds, mammals and amphibians. There are also other studies which used long-term hydrophone recordings of marine animals including baleen whales. These datasets are often obtained through publicly available repositories or collected in the field, enabling reproducibility and comparative analysis. In several studies, the limitations of datasets, such as class imbalance and insufficient samples of rare species were alleviated using synthetic data generation and augmentation algorithms.

2.2. Hardware and Sensor Infrastructure

Reviewed wildlife monitoring systems in this study are usually implemented with low-power embedded systems, like Raspberry Pi 3 Model B and Raspberry Pi Zero. These systems are complimented by Passive Infrared (PIR) motion sensors and camera modules to obtain real time images. Other modules are used in the case of alert-based systems, in this case Global Positioning System (GPS) modules and Global System for Mobile communications (GSM) modules are used to communicate real time notifications either through SMS or through the network-based communication. Omnidirectional microphones and underwater hydrophones are used in acoustic monitoring systems to record the soundscapes in the environment. The sampling rates depend on the target species, usually between 250 Hz to 16 kHz to effectively record species-specific vocalizations.

2.3. Data Preprocessing and Augmentation

Visual Data:

Image preprocessing and augmentation are commonly used to enhance model generalization and to deal with the imbalance in a dataset. These consist of random transformations like horizontal flipping, scaling, rotation, adjustments of hue and saturation, and mosaic augmentation. Bounding box annotations are converted to standard formats like the YOLO (x, y, width, height) format to train object detection models.

Acoustic Data:

Frequency-domain transformations are used to process audio signals, such as Short-Time Fourier Transform

(STFT) to extract features like Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, and energy distributions. To increase the diversity of training data, augmentation techniques are used, including pitch shifting, time stretching, and simulated environmental distortions (e.g., Doppler effects).

2.4. Algorithmic Architectures and Model Training

Visual Detection Models:

Classification and feature extraction tasks are performed using deep learning models based on Convolutional Neural Networks (CNNs), and including ResNet and VGG architectures. YOLO-based models (YOLOv5, YOLOv8) are highly used due to their high accuracy and speed in real-time detection. Lightweight architectures like WildARe-YOLO introduce depth-wise separable convolutions and mobile bottleneck layers to cut computational complexity, enabling them to be deployed to the edge. Fine-grained classification tasks, such as the identification of an individual animal, based on its unique visual features, use transfer learning methods using pretrained models such as EfficientNetB3.

Acoustic Detection Models:

The techniques of animal sound recognition can be classified into classical and deep learning techniques. Classifier models used in the traditional models include Random Forest which is of course trained on man-made acoustic features. More recent systems are based on transformer-based architectures, such as Wav2Vec2, which learns representations of the features directly on top of the raw audio signals. Few-shot learning algorithms have also been welcomed and they enable one to detect with a limited number of labeled samples. Also, to learn temporal relationships in animal vocalizations, Convolutional Recurrent Neural Networks (CRNNs) with recurrent units like GRUs are implemented.

2.5. Energy Reduction and Edge Deployment Strategies

Energy efficient processing strategies are deployed in order to facilitate long-term deployment in the remote and off-grid settings. Frame differencing is a motion detection technique which is used to filter the irrelevant frames and hence unnecessary deep learning inferencing is minimized. System architectures are often decomposed into distributed components, where local nodes sense and process information, and central nodes communicate with each other. This segregation minimizes the use of energy when there is no patient. In addition, model optimization techniques are optimized further by compression and conversion into lightweight format (e.g., TensorFlow Lite) as the means to make them compatible with low-memory embedded devices. The effects of these optimizations are significant in reducing power consumption, and increasing system scalability.

2.6. Evaluation Metrics

The studies reviewed measure the system performance by standard measures such as accuracy, precision, recall, F1-score, and mean Average Precision (mAP). Frames Per Second (FPS) are used to measure the real-time performance, whereas the reduction of power consumption and the efficiency of systems are used to evaluate the energy efficiency.

2.7. Availability of Data and Resources

The datasets of the reviewed studies consist of both publicly available repositories and specially collected field data. A number of studies offer access to trained models, code implementations and experimental setups to facilitate reproducibility. In the case of proprietary datasets, the original authors describe in detail the methods used to collect the data.

2.8. Ethical Considerations

All the reviewed methodologies are founded on non-invasive techniques of monitoring wildlife including camera traps and passive acoustic records, which cause minimal disturbance to wildlife. The relevant original studies have acquired ethical approvals and permissions in the field of data collection.

3. Results

The section includes a critical analysis of the studies reviewed, which will focus on the effectiveness of various detection modalities, system effectiveness, and practical implementation results. These outcomes are divided into visual detecting, acoustic recognition, and system level optimization.

3.1. Testing of Detection Modalities.

3.1.1. Visual Detection Accuracy and Efficiency

A deep Convolutional Neural Networks (CNNs) applied to camera trap images has shown impressive performance in automating the process of detecting, classifying, and counting wildlife. Top-1 accuracy of 93.8% and top-5 accuracy of 99.1% on the 3.2 million image Snapshot Serengeti dataset using heavyweight architectures, such as ResNet-152. With the addition of the confidence thresholding, the model was able to automate 99.3% of the identification process and thus a significant portion of the manual labeling workload, estimated to be around 8.4 years of human labor, was removed. In situations where real-time detection is required, the YOLOv8x model proved to be quite robust with a mean Average Precision (mAP) of 94.3% and a working rate of around 20 frames per second (FPS). This renders it applicable in the continuous observation in dynamic environments that involve the use of lions, tigers, leopards and bears. Besides detection at the species level, transfer learning techniques have allowed fine-

grained identification. The EfficientNetB3 model was used to achieve an 88.49% accuracy rate in classifying 98 individual tigers, showing the viability of intra species tracking using distinctive visual designs like stripe patterns.

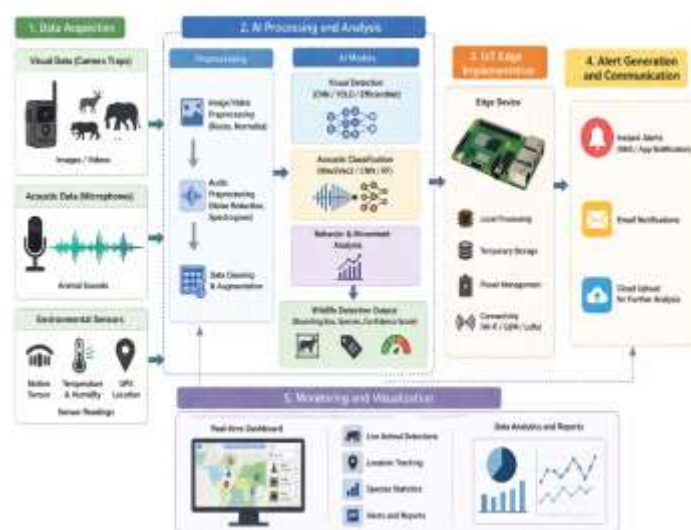


Fig -1: AI-based wildlife monitoring framework system architecture.

Table -1: List of Detection Model Performance.

Modality	Architecture	Key Performance Metric
Visual (Camera Trap)	ResNet-152	93.8% Top-1 Accuracy
Visual (Real-Time)	YOLOv8x	94.3% mAP @ 20 FPS
Visual (Individual)	EfficientNetB3	88.49% Accuracy
Acoustic (Waveform)	Wav2Vec2	92.75% Accuracy ¹
Acoustic (Handcrafted)	Random Forest	78.62% Accuracy ¹

¹Acoustic models have variability in their performance based on the environmental noise conditions.

3.1.2. Acoustic Recognition and Few-Shot Learning

Acoustic monitoring systems have been shown to have high performances of wildlife sensing on a challenging environment to observe. Models based on deep learning can perform much better in the complex soundscapes compared to classical machine learning techniques. In Wav2Vec2, transformer based architectures, a 92.75 percent classification accuracy was observed when using raw audio signals and a set of 15 animal species. By comparison with the conventional random forest models trained on manually engineered features such as Mel-Frequency Cepstral Coefficients (MFCCs) only reach an accuracy of 78.62 percent and it is evident that end-to-end deep learning models can perform better than this model. The Few-shot Sound Event Detection (SED) systems were used to improve the capability of the systems since

they are able to recognize the rare or variable animal sounds with high accuracy using as few as five labeled examples, which are very effective in data-sparse ecological settings.

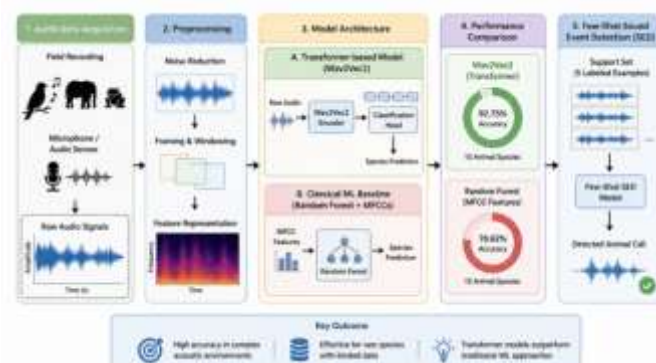


Fig -2: Architecture of acoustic recognition and few-shot learning model of wildlife monitoring.

3.1.3. Limitations Important Findings and Shortcomings.

The review of the studies analyzed reveals that deep learning models are generally highly performing, with the level of accuracy in most cases being above 90 percent in various modalities. Specifically, the YOLO based architecture has been shown to be very effective in real time wildlife detection because of their tradeoff between speed and accuracy. Also, audio models based on transformers are shown to be more effective than classical machine learning techniques, particularly in the complex acoustic environments. Transfer learning methods also boost the functionality of the system as they allow learning the separate animals even with a small amount of training data. Although these developments have taken place, there are a number of constraints. Visual detection models show worse performance in harsh conditions like low-light situations, dense vegetation, occlusion, and camouflage. Likewise, there are environmental noise effects and overlapping sound signals on the acoustic models, which can influence the classification accuracy. Moreover, access to balanced and large enough datasets is also a major problem, especially with the rarest and most endangered species, which can cause issues such as imbalanced classes and poor generalization of the models.

3.2. Performance of Edge Deployment Models.

Besides accuracy of detection, computational efficiency and consumption of energy is a crucial consideration in real world applications. Models have been made lightweight to allow them to be run on edge devices which are resource-constrained.

Table-2: Lightweight detection models that are more efficient.

Model	FLOPs Reduction	Parameter Reduction	FPS Increase
WildARe YOLO	50.92%	28.55%	17.65%
YOLOv5n	32.40%	15.20%	10.10%

WildARe-YOLO uses depth-wise separable convolutions (DWConv) that can be used to decrease computational complexity, although it does not affect detection accuracy.

3.3. Mathematical Elements Formatting

To measure the effectiveness of the edge-based wildlife monitoring systems, the total energy used by the device will be defined as:

$$E_{total} = E_{s2dl} + E_{s2i} + E_s + E_{idle}$$

(1)

where E_{s2dl} is the energy expended between sensing and deep learning inference, E_{s2i} denotes the energy used in sensing to lightweight processing (e.g., frame differencing), E_s is the energy required for sensing only, and E_{idle} corresponds to idle system energy. In a traditional system, inference of deep learning is activated continuously and hence a high cost of computing. Nevertheless, optimized systems add steps of intermediate filtering like motion detection and greatly minimizes unnecessary inference operations. This relationship can be further expressed as:

$$E_{s2dl} = N_{dl} \cdot P_{dl} \cdot t_{dl}$$

(2)

where N_{dl} is the number of deep learning inferences, P_{dl} is the power used during each inference, and t_{dl} is the time per inference. Considering the reduction in N_{dl} by means of motion-based filtering, the overall energy consumption is reduced considerably.

Theorem 1. Energy Thresholding Proposition.

Under the condition of a multimodal wildlife monitoring system motion-based filtering prior to deep learning inference, the idle energy component (E_{idle}) is the dominant component, and the entire system will be more energy efficient.

Proof of Theorem 1. Using the classic type of wildlife monitoring systems, image or audio data is continuously processed by deep learning models with or without the incidence of pertinent events. This leads to many high-power computational units being switched on and hence leads to a high energy consumption with the preponderance of E_{s2dl} . A hierarchical approach to processing is used in optimized systems. At first, sensing mechanisms that consume low-power such as Passive Infrared (PIR) sensors or frame differencing algorithms detect motion or a change in the environment. Deep learning inference is triggered by the system only when a major event is detected. This makes the number of inference operations N_{dl} , reduced, and this minimizes E_{s2dl} .

Furthermore, the field of system architecture is often divided into edge nodes (in charge of sensing and preprocessing) and communication nodes (in charge of the transmission of data). The isolation of these parts will not allow to continue operating the high-energy communication modules (e.g., GSM or cloud transmission), and will reduce unnecessary power consumption. Consequently, the overall energy balance is taken over by low-energy processes:

$$E_{total} \approx E_{idle} + E_s + E_{s2i}$$

where E_{idle} becomes the greatest contributor in idle times. As the E_{idle} is considerably less than the E_{s2dl} , the total energy consumption is greatly reduced. Experimental demonstrations show that this optimization can reduce the overall energy consumption by up to 60.6, which is tantamount to doubling the working life of battery-driven devices in remote locations. Therefore, motion-based filtering ensures that the idle energy state is dominant in the system hence proving the theorem.

4. Discussion

4.1. Detection Performance Interpretation.

The studies under review show a consistent result as deep learning models are much better at identifying wildlife than traditional methods. The visual detection systems based on convolutional neural network and YOLO architecture have high rates of accuracy of above 90 percent and ensure their applicability in the area of species recognition and also in the application of real-time monitoring. These results are in line with larger improvements in computer vision, where deep neural networks have excelled over handcrafted feature-based approach in complex classification problems. However, the results also indicate that still the visual detection systems remain very sensitive to the environmental conditions. The reality that even use of image-based methods in low light conditions, dense vegetation, and occlusion may not yield high reliability in the real-world.

4.2. Bioacoustic Monitoring role.

Bioacoustic monitoring is one such complement technique, which would overcome most of the shortcomings of visual detection systems. Transformer-based models such as Wav2Vec2 show good performance in recognition of species based on raw audio signals even in difficult and noisy conditions. Deep learning methods are much more accurate than traditional machine learning approaches, such as Random Forest models fitted on handcrafted MFCC features. The few-shot learning approaches also enable species to be detected using very small labeled datasets, which is especially important in the ecological monitoring setting, where annotated datasets are extremely scarce, especially with rare or endangered species.

4.3. Multimodal System Integration

The key finding of this review is the fact that multimodal systems involving the combination of visual, acoustic and sensor-based input may be more robust and reliable in

their performance than unimodal approaches. These systems are capable of overcoming the disadvantages of each of the modalities, and improving the detection accuracy in different environmental conditions. The heterogeneous source of data, however, comes with challenges including; complexity of the system, higher computation requirement and the difficulty in coordination and processing multimodal data streams. Despite all these, the results have a good sign that in the future, multimodal sensing strategies ought to be implemented as a form of wildlife monitoring system.

4.4. Power Saving and Side Implementation

The aspect of energy efficiency is especially critical towards the practical implementation of wildlife monitoring systems and especially in remote and off-grid locations. The reviewed literature reveals that edge computing and energy optimization methods can greatly decrease power usage with no effect on system functioning. Motion-based filtering and hierarchical processing implementation will cut down unnecessary deep learning inference, saving up to 60.6 of energy. These results validate that event-based system architectures are very effective in reducing the computational overhead and changing the device operating life. In practical view, the results reveal that model accuracy is not so crucial as system efficiency and hardware optimization to real-life applications.

4.5. Implications for Wildlife Conservation

The implementation of AI-based monitoring systems has significant implications to the wildlife conservation and human security. Early warning systems can be established in terms of real-time detection systems that can enable alerting of the villagers and authorities in the event of animal approaching human settlements thereby minimizing the chances of conflicts between humans and wild animals. The use of automated monitoring also reduces the need to rely on manual observations and makes it possible to have continuous data collection, which is important in ecological research, biodiversity research, and policy development. The developments help in more efficient conservation policies, and facilitate sustainable co-existence between people and animals.

4.6. Rebates of Existing Solutions

Though the performance of the AI-based systems that is currently guaranteed is quite promising, there is still a certain number of limitations. A lot of research is based on datasets which are either small in size or do not adequately represent rare species, and as a result challenges in generalization of the model. The performance of the system is still affected by other environmental factors such as changes in lighting, weather and the noise in the background. Also, the complexity and cost of implementing large-scale monitoring systems can be a limiting factor to implementation in resource constrained areas. These limitations are reflective of the necessity to undertake further research to improve on robustness, scalability and accessibility.

4.7. Future Research Directions

The following study is to be carried out on the development of multimodal artificial intelligence systems that combine visual, acoustic and sensor-based information to enhance the trustworthiness of detection in various environments. The emergence of light weight deep learning models which can be deployed to remote location with limited resources will be very critical in enabling real time processing in remote locations with limited resources. New technologies, like drone-based monitoring systems will be used to expand the range and improve the collection of the data in the inaccessible locations. Also, animal-behavior and movement predictive analytics can be utilized in supplement to early-warner systems to allow proactive as opposed to reactive reaction to potential conflicts. Technically, another critical challenge is also how to enhance resiliency of a system with changes in the environment. The data augmentation techniques will be forced to be effective in order to cope with the scarcity of labeled data and in particular rare and endangered species. More active harmonization of real time alert systems will also be needed to make sure that timely communication of monitoring systems and end users will be achieved. Finally, as part of the process of defining the ground work to massive adoption of these technologies to actual real world conservation work, scalable and cost-effective deployment plans will need to be developed.

4.8. Overall Interpretation

The findings of this review demonstrate that the artificial intelligence has been able to a high extent to realize the potential of the wildlife surveillance systems. But the fact is that none of the solutions is a full solution regardless of the environmental factors. The most promising future development is the combination of multimodal sensing, low-energy edge computing, and a system of real-time communication. These integrated systems can help in improved conservation actions, improved ecological information, and minimization of the chance of human-wildlife contacts.

5. CONCLUSIONS

Recent advancements in the field of artificial intelligence-based wildlife monitoring systems, based on image-based detection, bioacoustic analysis, IoT-enabled alert system, and energy efficient edge deployment, were discussed in this review. The findings demonstrate that deep learning models, specifically CNN and YOLO, are very accurate on visual detection tasks, but transformer-based models, such as Wav2Vec2, were much better on acoustic recognition. The approach of transfer learning also allows recognizing a particular animal in detail, which demonstrates that it is possible to conduct in-depth ecological monitoring. Despite these advancements, none of the detection modalities can be applied to consistently achieve good results in all environmental conditions. Visual systems are influenced by lighting, occlusion and

camouflage, and acoustic systems are influenced by environmental noise and redundant signals. Multimodal methods, which are based on visual, acoustic and sensor-based information, have proven to be the most effective approach in improving reliability and robustness in real life scenarios. System design by energy-efficient design, especially through edge computing and motion-based filtering techniques is a key focus enabling long term deployment in remote environments. These strategies reduce the amount of computation required, and dramatically extend the operational life, allowing large-scale monitoring to become more feasible. Overall, the field of wildlife monitoring has been revolutionized by artificial intelligence through offering automated, real-time and scalable solutions. Multi-modal integration, model design and affordability deployment strategies should be prioritized in the future to facilitate sustainable conservation measures and eliminate human-wildlife conflicts.

ACKNOWLEDGEMENT

The authors would like to credit the useful contributions to the previous research works and datasets that are publicly available and formed the basis of this review. The authors would also like to take this opportunity to acknowledge the fact that during the drafting process of this paper, they had the great privilege of being guided by the Department of Computer Engineering, RMDSTIC, Pune in terms of the academic advice and institutional support.

REFERENCES

1. Norouzzadeh, M.S.; Nguyen, A.; Kosmala, M.; Swanson, A.; Palmer, M.S.; Packer, C.; Clune, J. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl. Acad. Sci. USA* 2018, 115, E5716–E5725. <https://doi.org/10.1073/pnas.1719367115>
2. Premarathna, K.S.P.; Rathnayaka, R.M.K.T.; Charles, J. To solve the issue of human-elephant conflict and to track the movements of an elephant with the help of deep learning, an elephant detection system. In *Proceedings of the IEEE International Conference on Industrial and Information Systems (ICIIS)*, Colombo, Sri Lanka, 2020. <https://doi.org/10.1109/ICITR51448.2020.9310798>
3. Sato, R.; Saito, H.; Tomioka, Y.; Kohira, Y. The energy reduction method in detecting devices of wild animals. *IEEE Access* 2022, 10, 24149–24160. <https://doi.org/10.1109/ACCESS.2022.3155242>
4. Redmon, J.; Farhadi, A. YOLOv3: A slight improvement. *arXiv* 2018, arXiv:1804.02767.
5. Jocher, G.; Chaurasia, A.; Qiu, J. YOLOv5 by Ultralytics. Available online: <https://github.com/ultralytics/yolov5> (accessed on 10 April 2026). Jocher, G.; Qiu, J.; Chaurasia, A. YOLOv8: Ultralytics next-generation object detection model. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 10 April 2026).
6. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. This can be found in *Proceedings of the International Conference on Machine Learning (ICML)*, Long Beach, CA, USA, 2019. <https://doi.org/10.48550/arXiv.1905.11946>
7. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. Wav2Vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2020. <https://doi.org/10.48550/arXiv.2006.11477>
8. Chollet, F. *Deep learning with Python*; Manning Publications: New York, NY, USA, 2017. 10.
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Published in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016. <https://doi.org/10.1109/CVPR.2016.90>
10. Very convolutional networks, very large-scale image recognition. Simonyan, K.; Zisserman, A. *arXiv* 2015, arXiv:1409.1556.
11. Breiman, L. Random forests. *Mach. Learn.* 2001, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
12. Stowell, D.; Plumbley, M.D. Unsupervised learning of features is of great benefit to automatic large-scale classification of bird sounds. *PeerJ* 2014, 2, e488. <https://doi.org/10.7717/peerj.488>
13. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An ontology and human labeled audio event dataset. Between 2017 and 2018, over 3000 articles were available for purchase. Between 2017 and 2018, approximately 3000 articles were on sale. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017. <https://doi.org/10.1109/ICASSP.2017.7952261>
14. Howard, J.; Ruder, S. Fine-tuning a universal language model to text classification. In *Proceedings of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, 2018. <https://doi.org/10.18653/v1/P18.1031>
15. Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; Vinyals, O. To know about deep learning one has to revisit the concept of generalization. *arXiv* 2017, arXiv:1611.03530.
16. TensorFlow Lite. TensorFlow Lite Mobile and embedded. Available online: <https://www.tensorflow.org/lite> (accessed on 10 April 2026).
17. Alumentations Library. Alumentations: Library of fast image augmentation. Available online: <https://alumentations.ai> (accessed on 10 April 2026).