Multimodal Deep Learning Based Wildlife Intrusion Perception Using YOLOv12 and YAMNet

Vamshi Krishna Velpula[1], Arun Kumar Ankeshwarapu[1], Madhu Kumar Bolle[1], Dr. B. Venkat Raman[2]

[1]Student, Department of Computer Science and Engineering, Rajiv Gandhi University of Knowledge Technologies, Basar, India

[2]Assistant Professor, HOD, Department of Computer Science and Engineering, Rajiv Gandhi University of Knowledge Technologies, Basar, India

DOI: https://doi.org/10.51244/IJRSI.2026.1304000040

Received: 20 March 2026; Accepted: 26 March 2026; Published: 27 April 2026

ABSTRACT

Crop damage caused by wildlife intrusion is a major challenge for farmers near forest boundaries. Traditional monitoring methods are labor-intensive and ineffective under poor visibility conditions. This paper proposes a multi-modal wildlife intrusion detection system that combines visual object detection and environmental sound classification.

The system utilizes the YOLOv12 model for real-time animal detection from surveillance video and YAMNet for identifying animal sounds. By integrating visual and auditory sensing, the proposed framework improves detection reliability in low-light or occluded conditions. Experimental evaluation demonstrates improved detection accuracy compared to single-modal approaches. The system can be deployed on edge devices such as Raspberry Pi or Jetson Nano, enabling real-time monitoring of agricultural fields.

Keywords: Wildlife Intrusion Detection, Deep Learning, Computer Vision, YOLO Object Detection, Environmental Sound Classification, YAM Net, Smart Agriculture, Multimodal Monitoring.

INTRODUCTION

Agriculture plays a crucial role in supporting rural livelihoods and food security across many regions of the world. However, one of the persistent challenges faced by farmers, particularly those located near forest boundaries, is the intrusion of wild animals into agricultural fields. Animals such as elephants, wild boars, deer, and monkeys frequently enter farmlands in search of food, causing severe crop damage and economic loss to farmers. In many cases, farmers rely on traditional methods such as manual guarding, fencing, or noise-making devices to deter animals.

These approaches are often inefficient, labour-intensive, and ineffective during nighttime or adverse weather conditions such as fog or heavy rain. With the rapid advancements in **Deep Learning and Computer Vision**, automated monitoring systems have become a promising solution for addressing wildlife intrusion problems. Modern object detection algorithms can recognize and classify objects in real time with high accuracy. Among these algorithms, the YOLO (You Only Look Once) family of models has gained significant popularity due to its ability to perform fast and accurate object detection in real-time applications.

In this work, we propose a **Deep Learning-based Wildlife Intrusion Perception System** that utilizes the **YOLO-V12 object detection model** to detect animals from real-time webcam or surveillance video feeds. The model is trained to recognize multiple animal species and provide immediate detection results, enabling continuous monitoring of farmland environments.

However, relying solely on visual detection can be challenging in low-light conditions, dense vegetation, or when animals are partially occluded. To address this limitation, the proposed system integrates an additional **audio-**

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

based detection module using **YAM Net**, a pre-trained deep learning model designed for environmental sound classification. This module enables the system to detect characteristic animal sounds such as trumpeting, growling, or screeching, thereby improving detection reliability even when visual cues are limited.

By combining both visual and auditory inputs, proposed system provides a **multi-modal monitoring framework** that enhances detection accuracy and reliability. The system is designed to assist farmers by providing an intelligent monitoring solution capable of reducing crop damage and improving farmland safety. Furthermore, the framework can be extended for deployment on low-cost edge devices such as **Raspberry Pi or Jetson Nano**, enabling practical real-time monitoring in rural agricultural environments.

The development of intelligent monitoring systems can significantly reduce the dependency on manual surveillance in agricultural fields. By leveraging deep learning and real-time sensing technologies, automated systems can detect wildlife presence quickly and accurately. Such systems enable farmers to receive early alerts when animals approach farmland areas. This helps in minimizing crop damage and improving overall farm security.

Therefore, the integration of advanced detection models provides an effective solution for modern smart agriculture.

RELATED WORK

M. N. Kishore *et al.* [1] proposed a real-time wild animal detection and classification system using deep learning techniques to reduce human–wildlife conflicts. Their approach utilized convolutional neural network models to detect and classify animals entering agricultural areas, enabling early alerts to farmers, and helping to reduce crop damage.

The study demonstrated that deep learning-based detection systems can significantly improve monitoring efficiency compared to traditional surveillance methods.

O. Gnanasekar *et al.* [2] developed an image processing-based animal intrusion detection system for agricultural fields using deep learning. The system processes images captured from surveillance cameras and applied object detection algorithms to identify animals approaching farmland regions.

Their work showed that automated vision-based monitoring can reduce the dependency on manual field guardingT. S. Dilwar and S. Mukhopadhyay [3] presented a real-time farm surveillance system integrating Internet of Things (IoT) technology with the YOLOv8 object detection model for detecting animal intrusions. The system used cameras connected through an IoT framework to continuously monitor farmland areas and generate alerts when animals were detected.

Their results highlighted the effectiveness of combining IoT infrastructure with deep learning models for smart agricultural monitoring.

Apart from vision-based systems, audio-based detection methods have also been explored for environmental monitoring. C. Malmberg [4] implemented a real-time audio classification system on edge devices using YAM Net and TensorFlow Lite. The system demonstrated the ability to classify environmental sounds, including animal calls, in real-time even on low-power devices. The YAM Net model available on TensorFlow Hub [5] provides a pre-trained deep neural network capable of recognizing a wide variety of environmental sounds.

Additionally, Robo flow [6] offers datasets and tools that facilitate the training and deployment of computer vision models for object detection applications.

PROPOSED SYSTEM / METHODOLOGY

The proposed system utilizes a multi-modal deep learning framework to detect wildlife intrusion in agricultural fields using both visual and audio inputs. A YOLO-based object detection model processes real-time webcam feeds to identify animals, while the YAM Net model analyzes environmental sounds captured from a

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

microphone. By combining visual and auditory information, the system improves detection of reliability and enables continuous monitoring of farmland environments.

System Architecture

The proposed wildlife intrusion detection system follows a multi-layer architecture consisting of input, processing, integration, and output layers. The **input layer** collects real-time environmental data through a webcam for video capture and a microphone for audio capture.

In the **processing layer**, the video stream is analyzed using the **YOLOv12 object detection model** to identify animals in the camera feed, while the audio input is processed using the **YAM Net sound classification model** to detect animal-related sounds. The outputs from both models are then combined in the **integration layer.**

Finally, the **output layer** displays detection results to the farmer, generates alerts through sound or notifications, and stores detection logs in a CSV file for future analysis and monitoring.

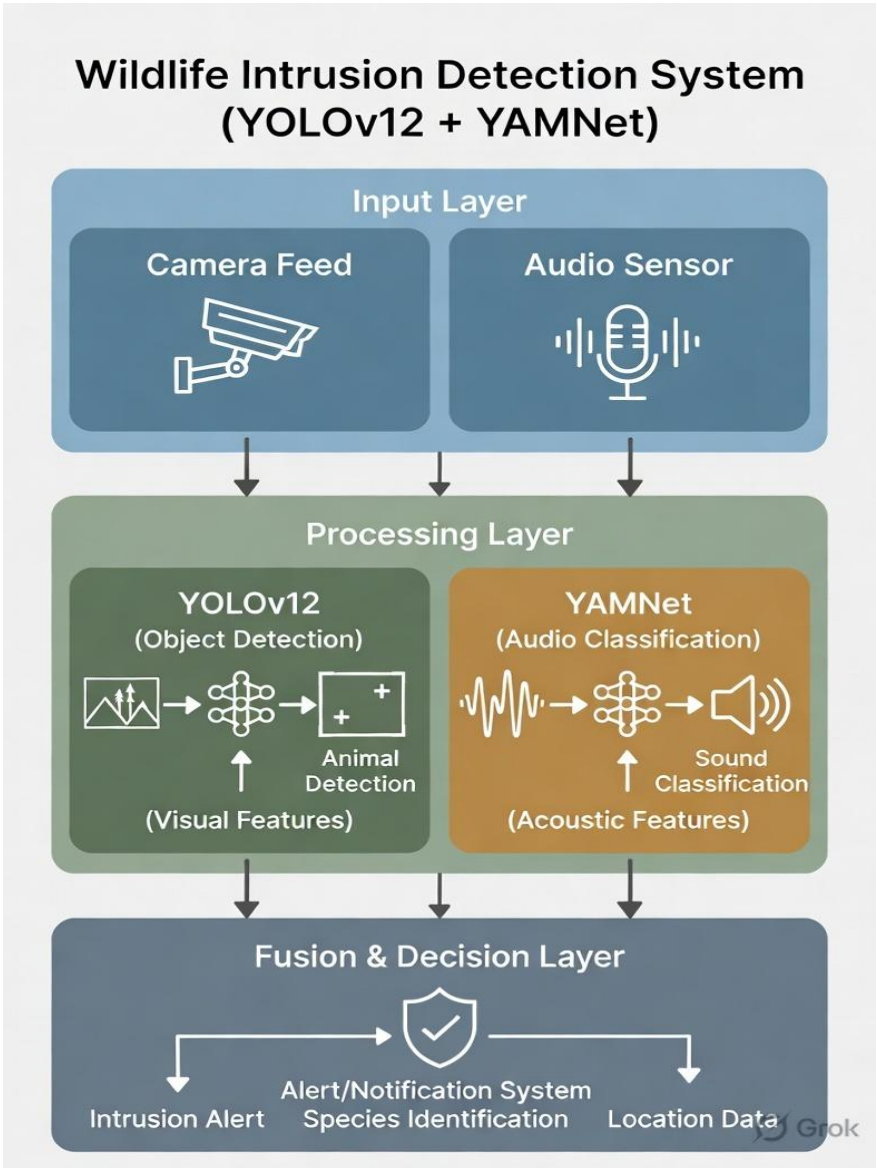


Fig. 1. Proposed system architecture for wildlife intrusion detection using YOLOv12 and YAM Net.

B. YOLO-V12 Animal Detection

The proposed system uses the YOLO-V12 object detection model to identify animals from real-time video captured through a webcam. YOLO-V12 is a deep learning–based single-stage object detector that performs object localization and classification simultaneously, making it suitable for real-time wildlife monitoring.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

The YOLO-V12 architecture consists of three main components: **Backbone, Neck, and Head**. The **backbone network** extracts important visual features from the input image using convolution layers and R-ELAN modules. These layers help capture spatial patterns such as shapes, edges, and textures of animals in the scene.

The **neck component** performs feature aggregation using sampling and concatenation operations to combine information from different feature scales. This allows the model to detect animals of varying sizes and distances within the frame.

Finally, the **detection head** predicts bounding boxes and class probabilities for detected animals. The model outputs the location of animals along with confidence scores, enabling the system to identify wildlife intrusion in agricultural fields in real time.

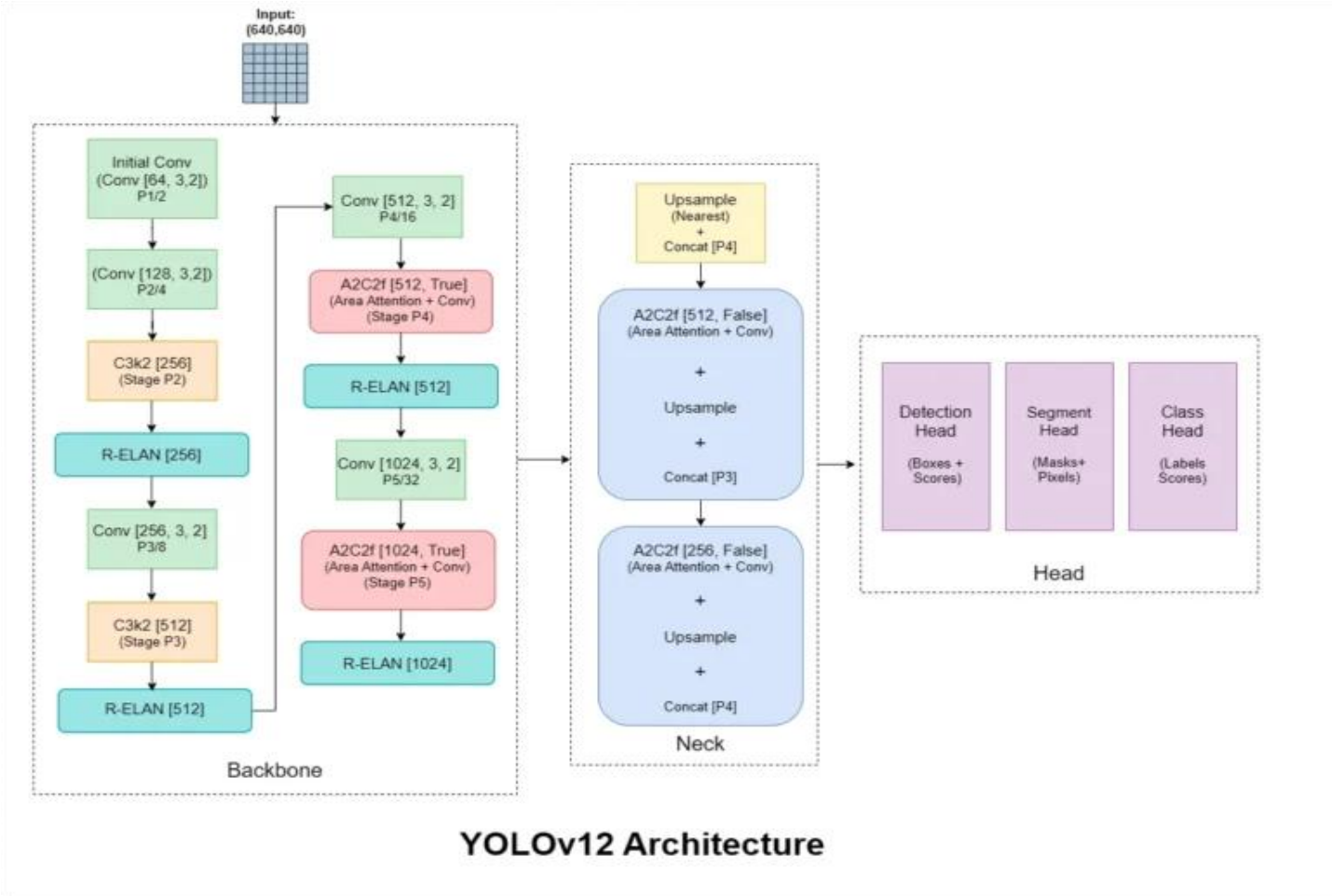


Fig. 2. Architecture of the YOLO-V12 object detection model showing backbone, neck, and head components.

C. YAM Net Audio Classification

YAM Net is an efficient, pre-trained convolutional neural network by Google for audio event classification. It classifies 521 sound classes from the Audio Set ontology, including animal calls, environmental sounds, and intrusion indicators such as gunshots or distress vocalizations. The model takes 16 kHz mono audio as input, converts it to log-mel spectrogram patches (96×64, ≈0.96 seconds), and performs frame-level multi-label prediction, making it well-suited for continuous real-time wildlife monitoring on edge devices.

The architecture leverages MobileNetV1's depth-wise separable convolutions to achieve low latency and small model footprint. Each separable block applies to a depth-wise convolution followed by a pointwise (1×1) convolution, drastically reducing parameters compared to standard convolutions. The diagram shows the first 27 layers dedicated to these convolutional operations, each including batch normalization for stable training and Re LU activation to introduce non-linearity.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Subsequent layers perform average pooling to condense spatial features, followed by fully connected layers that generate a 1024-dimensional embedding. The final SoftMax layer outputs probabilities across the 521 classes.

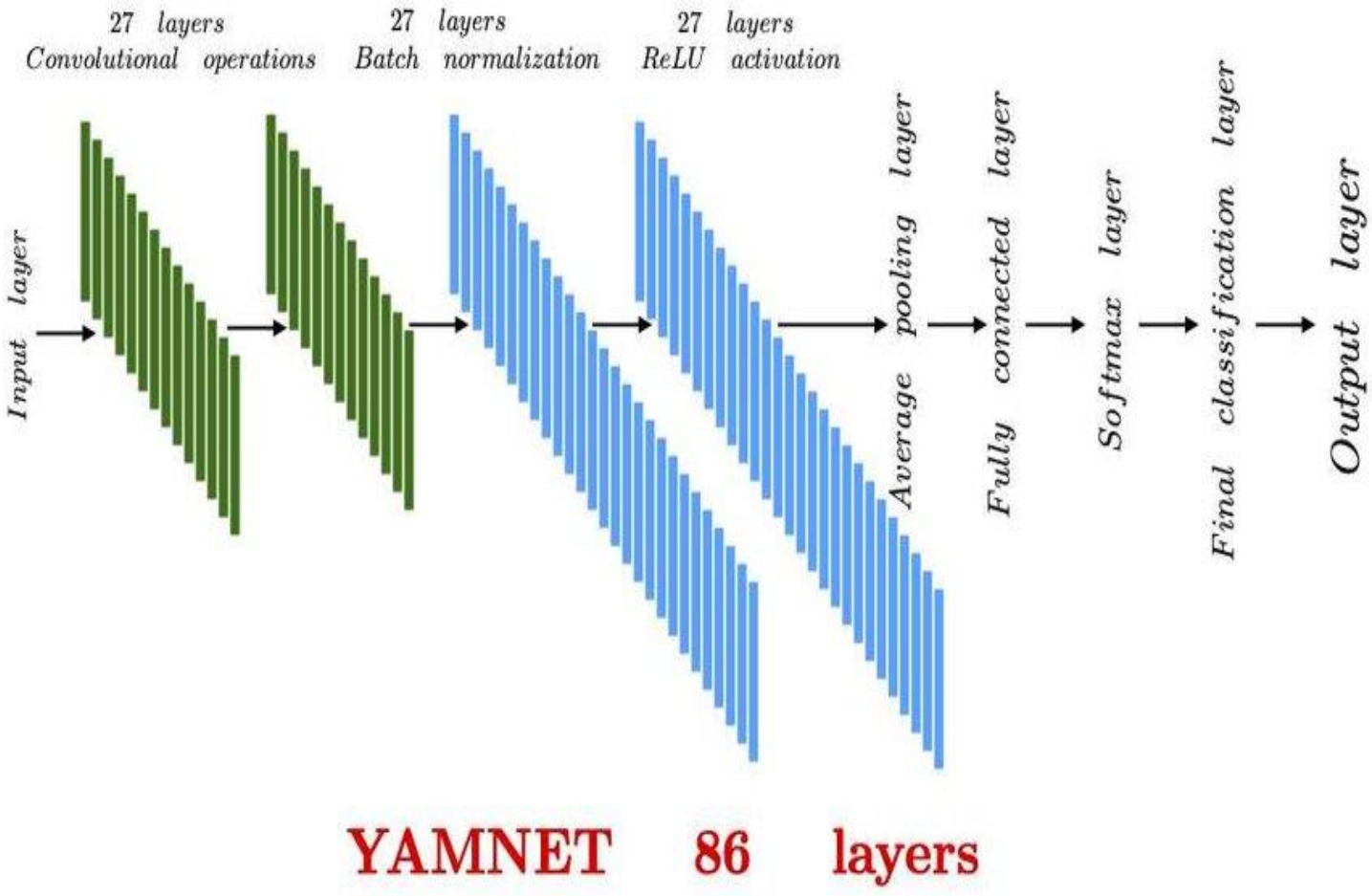


Fig 3. Architecture of YAM Net for audio classification in the wildlife intrusion detection system.

D. Multimodal Detection Process

The proposed system employs a multimodal detection approach that integrates both visual and audio data to improve the accuracy of wildlife intrusion detection. The webcam continuously captures video frames from the monitored area, while a microphone records environmental sounds. These inputs allow the system to analyze both animal movement and animal-generated sounds simultaneously.

The captured video frames are processed using the YOLO-V12 object detection model to identify animals based on their visual features such as shape, size, and motion. At the same time, the recorded audio signals are analyzed using the YAM Net model to classify sounds that may indicate the presence of wildlife, such as animal calls or movement noises.

The outputs from the visual and audio models are combined in a fusion module that verifies detection events using both sources of information. When both modalities confirm the presence of an animal, the system triggers alerts and logs the detection results. This multimodal approach enhances detection of reliability and reduces false alarms compared to systems that rely on only a single data source.

EXPERIMENTAL SETUP

The proposed wildlife intrusion detection system was implemented using a computer equipped with a webcam and microphone to capture real-time video and audio data from the surrounding environment.

The hardware setup enables continuous monitoring of the farmland area for detecting animal movement and sounds.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

The system was developed using **Python** along with several deep learning and computer vision libraries. The **YOLO-V12** model was used for real-time object detection, while **YAM Net** was used for environmental sound classification.

Libraries such as **OpenCV** were used for video processing and frame extraction, and **TensorFlow** was used to run the audio classification model.

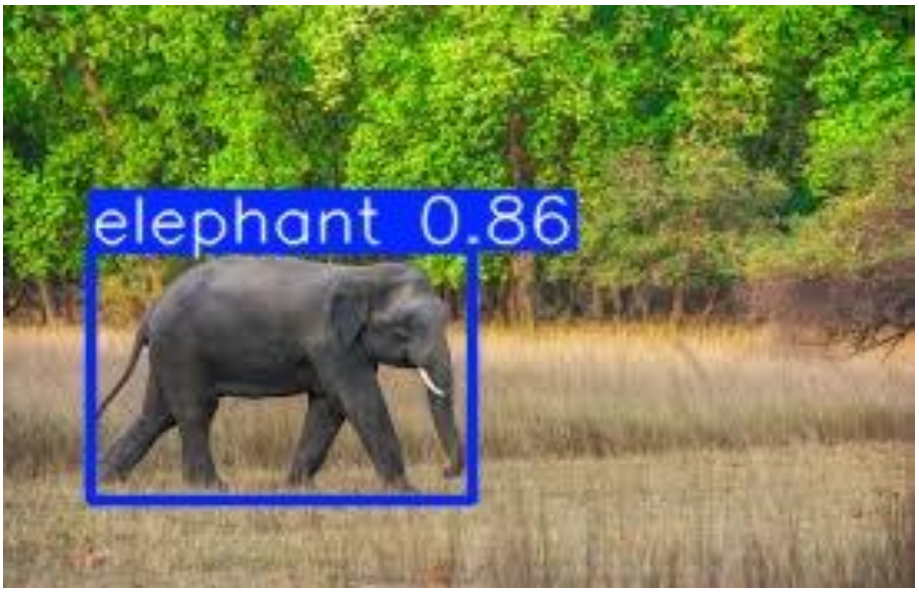
For training and testing the visual detection model, image datasets containing various animal classes were used and processed through the **Robo flow platform** for dataset preparation.

The YOLOv12 model was used for object detection on wildlife datasets. Training was conducted using image size 320×320, batch size 16, and 100 epochs..

The system was tested in a real-time monitoring environment where both visual and audio inputs were analyzed simultaneously to evaluate the effectiveness of the proposed multimodal wildlife intrusion detection approach.

RESULTS AND DISCUSSION

The proposed wildlife intrusion detection system was tested using real-time video and audio inputs captured through a webcam and microphone. The YOLO-V12 model successfully detected animals in the video frames and generated bounding boxes around detected objects with confidence scores. This enabled the system to identify wildlife presence in the monitored environment.



ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Fig. 4. Animal detection results using YOLO-V12 showing bounding box and confidence score.

The YAM Net model analyzed environmental audio signals and classified sounds related to animals. This helped the system detect wildlife even when visual conditions were poor, such as during low-light or partially occluded scenes.

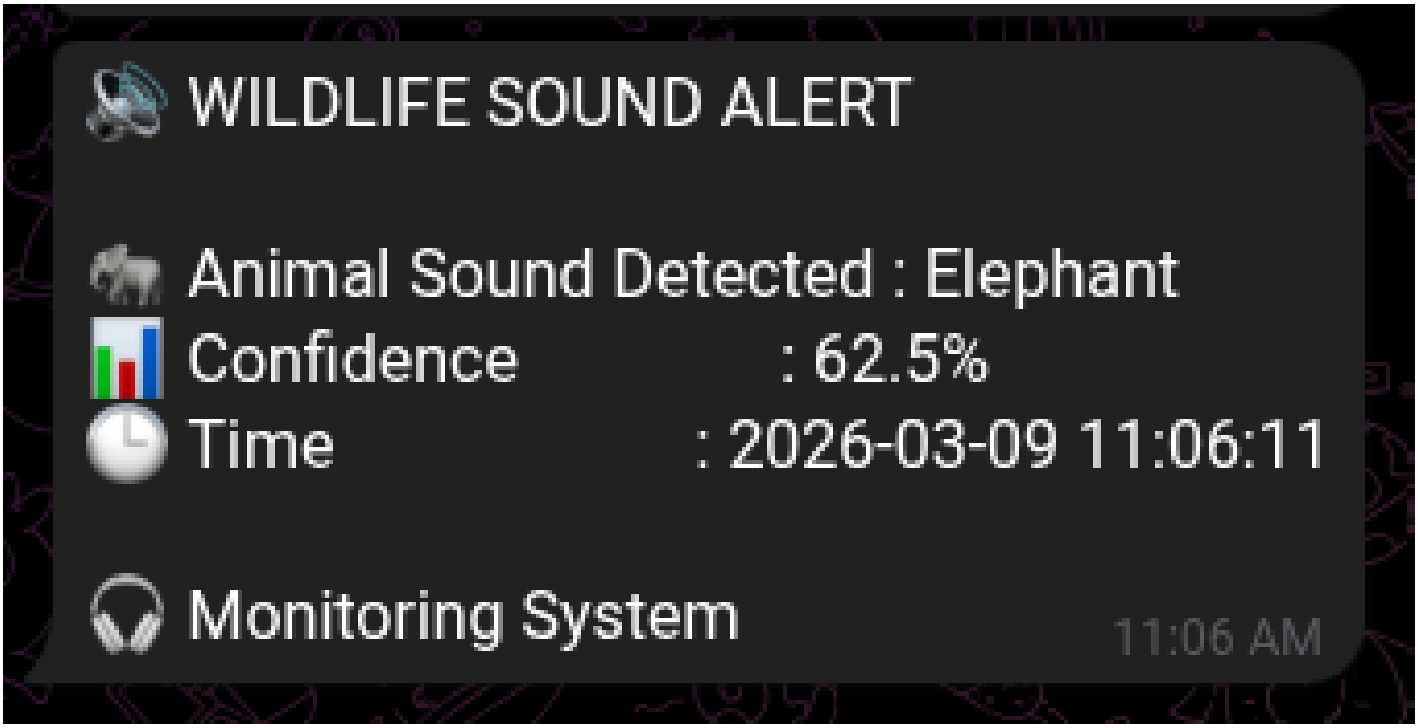


Fig. 5. YAM Net audio classification output showing detected environmental sounds.

The integration of both visual and audio detection methods formed a multimodal detection framework that improved the overall eliability of the system. When both the YOLO-V12 and YAM Net models indicated the presence of wildlife, the system confirmed the intrusion event and triggered alerts. This combination helped reduce false detections that could occur when using only a single detection method.

Experimental observations indicate that the system can monitor agricultural environments effectively and detect potential wildlife intrusions in real time. The system continuously processes incoming data streams and logs detection results for further analysis. These logs can help farmers understand patterns of wildlife movement and take preventive measures to protect crops.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

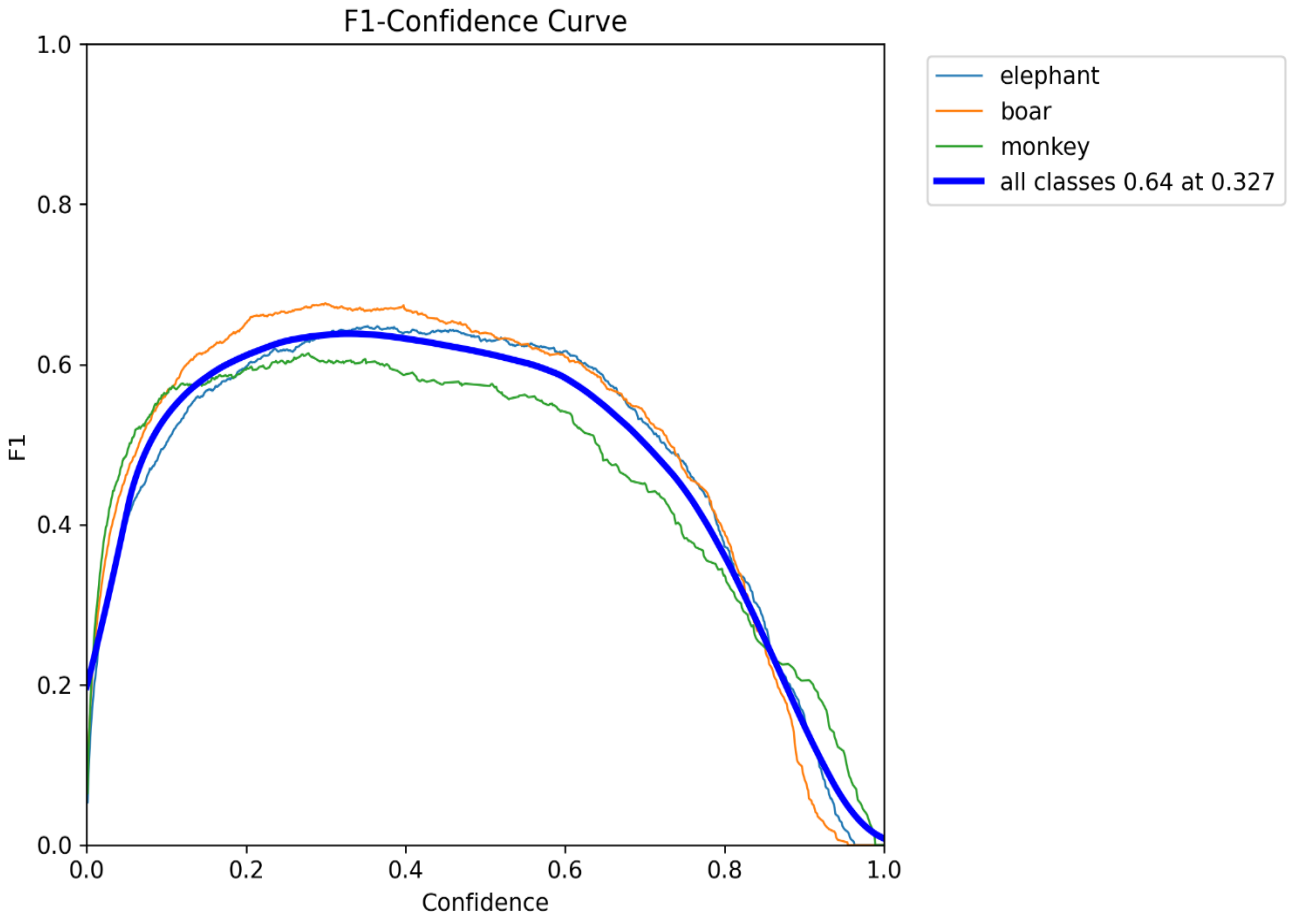


Fig. 6. F1-score performance of the proposed model.

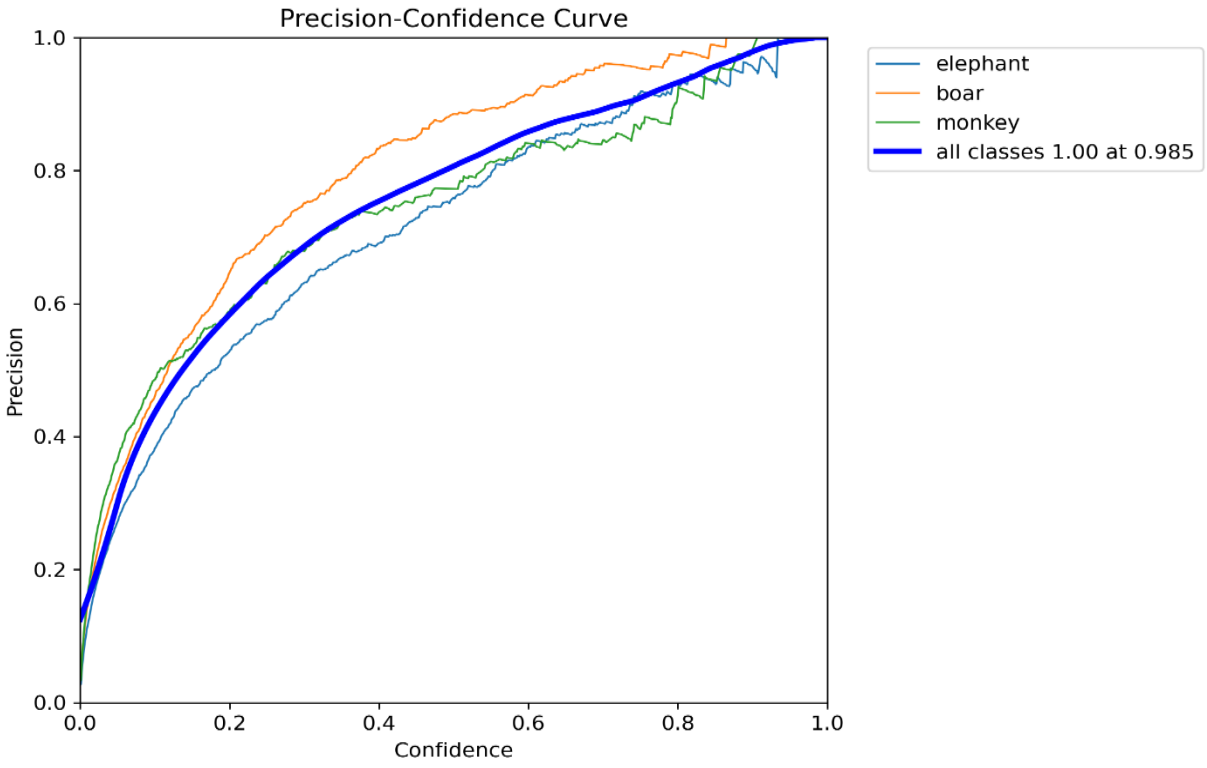


Fig. 7. Precision performance of the proposed model.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

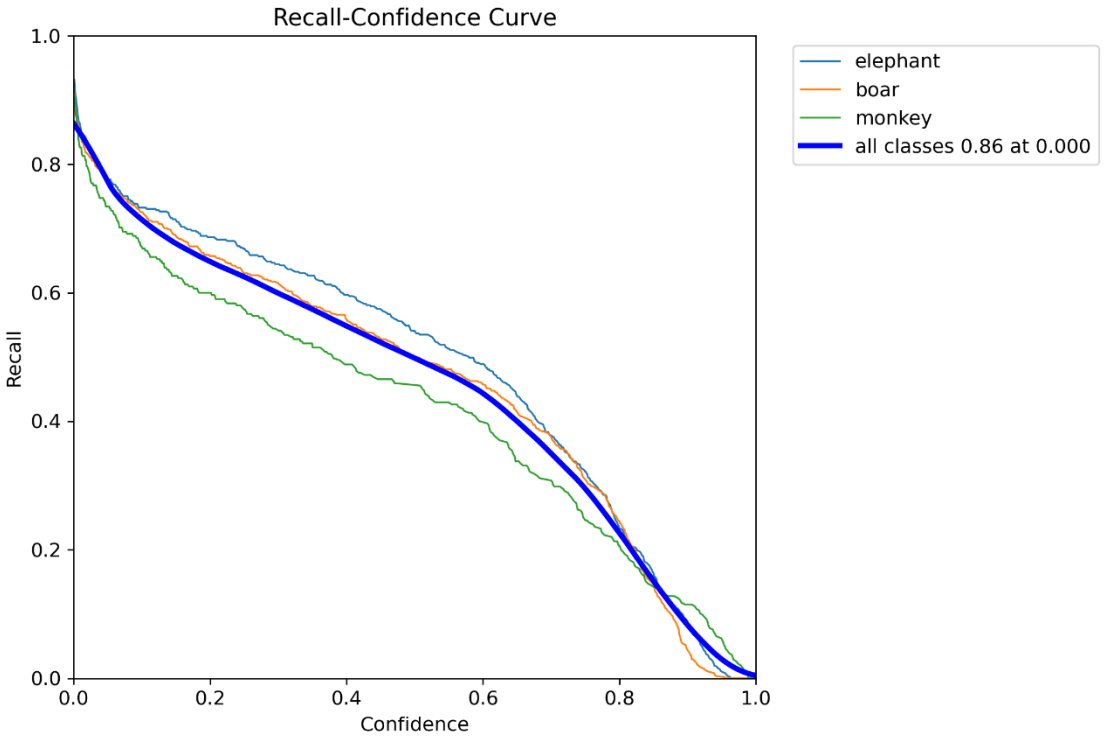


Fig. 8. Recall performance of the proposed model.

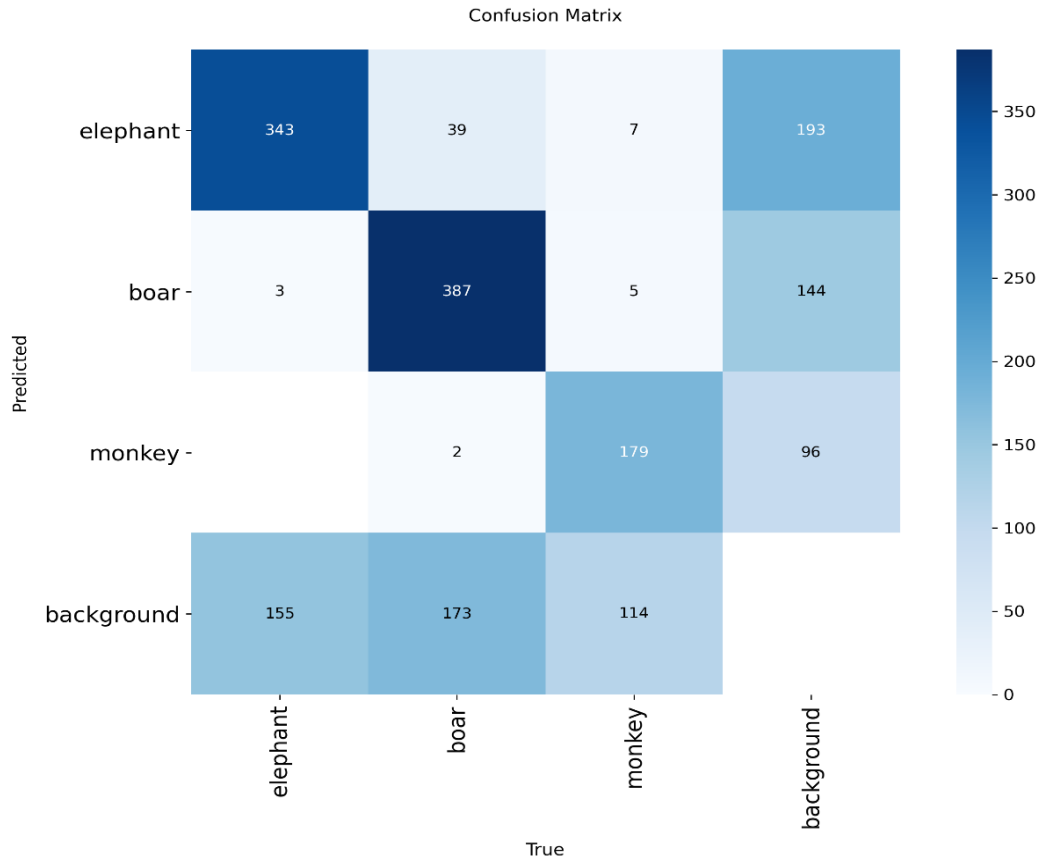


Fig. 9. Confusion matrix of the proposed model showing classification performance across different animal classes.

Overall, the results demonstrate that the proposed multimodal approach can serve as an efficient and intelligent monitoring solution for reducing crop damage caused by wild animals.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

CONCLUSION AND FUTURE WORK

This paper presented a multimodal wildlife intrusion detection system designed to assist farmers in monitoring agricultural fields and reducing crop damage caused by wild animals.

The proposed system integrates the YOLO-V12 object detection model for visual animal detection with the YAM Net model for environmental sound classification.

By analyzing both video and audio inputs simultaneously, the system can detect wildlife presence more accurately and reliably.

The experimental results demonstrate that the integration of visual and audio detection improves the performance of wildlife monitoring systems compared to single-modality approaches.

The system is capable of identifying animals in real-time and generating alerts when intrusion events occur, thereby helping farmers take preventive actions.

In the future, the system can be extended by deploying it on edge computing devices such as Raspberry Pi or NVIDIA Jetson for field-level implementation.

Additionally, further improvements can be made by expanding the dataset to detect a wider range of animal species and integrating automated deterrent mechanisms to prevent animals from entering farmlands.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to **Dr.B. Venkat Raman** for his valuable guidance, continuous support, and insightful suggestions throughout the development of this project. We also thank the **Department of Computer Science and Engineering, Rajiv Gandhi University of Knowledge Technologies, Basar**, for providing the necessary facilities and resources to successfully carry out this research work.

This work was conducted as part of the **final year's major project**, and the authors greatly appreciate the academic environment and encouragement provided by the institution.

REFERENCES

1. Kishore, M. N., Mahesh Babu, B., & M. D. (2025). Real-time wild animal detection and classification using deep learning for human–wildlife conflict mitigation. International Journal of Research Publication and Reviews (IJRPR).
2. Gnanasekar, O., Dinesh, P., & S. K. (2024). Image processing based animal intrusion detection system in agricultural field using deep learning. In Proceedings of IEEE Conference.
3. Delwar, T. S., & Mukhopadhyay, S. (2025). Real-time farm surveillance using IoT and YOLOv8 for animal intrusion detection. MDPI.
4. Malmberg, C. (2021). Real-time audio classification on an edge device using YAMNet and TensorFlow Lite [Online].
5. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 779–788).
6. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection [Online]. https://arxiv.org/abs/2004.10934
7. TensorFlow Hub. (2025). YAMNet: Audio event classification [Online]. Available at: https://tfhub.dev/google/yamnet/ (Accessed: Oct. 2025).
8. Roboflow. (2025). Roboflow: The universal dataset platform for computer vision [Online]. Available at: https://roboflow.com/ (Accessed: Oct. 2025).