IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 31, No. 6, June, 2026, pp. 2123-2133
Journal homepage: http://iieta.org/journals/isi

A Multimodal Deep Learning and Internet of Things Framework for Real-Time Stray Dog Behavior Detection and Child Safety Monitoring

Anushree Chandrappa*, Chaitra Mrutyunjaya, Rajanishree Manjappa, Aishwarya Meti

Department of Computer Science Engineering, BNM Institute of Technology, affiliated with Visvesvaraya Technological University (VTU), Bangalore 560070, India

Corresponding Author Email: anushreechandrappa5@gmail.com

https://doi.org/10.18280/isi.310630

Received: 11 May 2026
Revised: 10 June 2026
Accepted: 17 June 2026
Available online: 30 June 2026

Keywords:
stray dog behavior recognition, multimodal deep learning, audio-visual fusion, Internet of Things, object detection, smart safety monitoring, real-time alert system

ABSTRACT

Stray dog attacks in public and residential environments remain a safety concern, particularly for children and vulnerable individuals. Existing monitoring approaches mainly focus on animal detection or motion surveillance, while providing limited capability for understanding aggressive behavior and generating timely warnings. This study proposes a multimodal deep learning (DL) and Internet of Things (IoT) framework for real-time stray dog behavior detection and child safety monitoring. The proposed system integrates audio-based vocal analysis, vision-based behavior recognition, and IoT-enabled alert communication within an edge-oriented architecture. A convolutional neural network (CNN) is employed to classify dog vocal patterns using acoustic features extracted from recorded sounds, while a YOLOv5-based detection module identifies dogs and analyzes visual behavioral characteristics from images and video streams. The outputs from audio and visual modalities are fused through a decision-level mechanism to improve detection reliability and reduce false alarms. An ESP32-based IoT platform equipped with sensing modules, warning devices, and Telegram notifications is developed to enable real-time user alerts. Experimental evaluation demonstrates that the proposed multimodal framework achieves an overall accuracy of approximately 92%, outperforming individual audio-only or image-only approaches. The results indicate that combining DL-based behavioral analysis with IoT communication provides an effective approach for intelligent animal monitoring and child safety protection in real-world environments.

1. INTRODUCTION

The aggression shown by dogs in roads, parks, and other public places is a serious issue that poses a danger to people, especially children, elderly people, and pedestrians. Such aggressive actions are sudden, leaving little room for precaution, hence the need for a system that can detect such aggression to ensure safety. The current approaches for detecting dog behavior include human inspection, rudimentary CCTV cameras, and motion detection devices, but they do not analyze behaviors or provide timely alerts. With the advancement of artificial intelligence (AI), machine learning (ML), Computer Vision, and Internet of Things (IoT), intelligent automated systems can now detect animals, analyze behavior, and generate alerts in real time, thereby improving safety and response time.

Recent research has focused on recognizing dog behavior using deep learning (DL) techniques. Hussain et al. [1] proposed a DL-based activity detection system using wearable sensors to monitor dog behavior and wellbeing, demonstrating that AI models can effectively identify behavioral patterns. Nagy and Korondi [2] introduced a DL-based recognition system for analyzing limb-independent dog behavior, showing improved accuracy in behavior classification tasks. Li et al. [3] developed a DL-based dog expression recognition system that identifies emotional states using visual cues such as facial features and posture. These studies highlight the importance of visual and behavioral analysis for understanding dog aggression.

Audio-based classification has also been widely explored for detecting suspicious or aggressive events. Barkana et al. [4] introduced auditory suspicious event databases for sound-based detection systems, emphasizing the role of audio classification in identifying abnormal events. Seo et al. [5] proposed a convolutional neural network (CNN) using log mel-spectrograms for audio event classification, demonstrating high accuracy in sound-based recognition tasks. Talha et al. [6] presented a unified approach to voice classification using spectrogram and mel-spectrogram features, showing that DL models can effectively classify different sound patterns. These works motivate the use of CNN-based audio classification for detecting aggressive dog vocalizations such as growling and intense barking.

Computer vision-based animal detection has also shown promising results. Ke et al. [7] proposed a pet recognition algorithm using DL and edge devices, demonstrating real-time detection capability for animal monitoring systems. Parkavi et al. [8] developed a system for detecting animals in real-time

environments to enhance road safety, highlighting the effectiveness of automated vision-based monitoring. Additionally, YOLOv5-based stray dog detection systems have demonstrated high-speed and accurate detection of dogs in real-world scenarios, making them suitable for real-time safety applications [9].

Motivated by these works, the proposed Aggressive Dog Detection System integrates both audio and visual intelligence into a single framework. The system uses CNN with Librosa for classifying dog vocal sounds such as bark, growl, and grunt, while YOLOv5 is used for real-time detection of dogs and classification based on behavior analyzing facial cues. The AI models are integrated with an ESP32-based IoT platform that includes a sound sensor, UV sensor, liquid crystal displays (LCD) display, buzzer alert, and Telegram notification system. When aggressive behavior is detected, the system triggers a buzzer and sends instant alerts to users. By combining multimodal detection and IoT-based monitoring, the proposed system improves accuracy, reduces false alarms, and provides a reliable real-time safety solution for residential and public environments.

2. RELATED WORK

Recent advancements in computer vision, DL, and multimodal AI systems have significantly accelerated research in automated animal behavior monitoring and activity recognition. Traditional dog activity recognition systems initially relied on image-based classification models that focused on identifying simple postures and actions from static frames. Nolasco et al. [10] proposed a DL framework for dog activity recognition using the InceptionV3 CNN architecture deployed on a Raspberry Pi platform for edge-based monitoring applications. Their system classified dog activities such as sitting, standing, and lying using captured visual data and achieved approximately 88% classification accuracy. The major contribution of this work was demonstrating that lightweight CNN architectures could be implemented in low-cost embedded systems for pet monitoring. However, the model primarily depended on static image classification and failed to capture temporal movement patterns such as continuous walking, running, or transitions between activities, which are critical for real-world behavioral monitoring.

To address limitations in vision-only systems, researchers explored wearable sensor-based activity recognition approaches. Syeda et al. [11] developed a dog activity detection and recognition framework that utilized wearable accelerometer and motion sensors to capture movement patterns of dogs. ML algorithms were applied to classify activities such as running, walking, sitting, standing, and resting. The study demonstrated promising performance with nearly 79% recognition accuracy, showing that motion-based data can effectively capture behavioral information. However, wearable-based systems introduce practical challenges such as hardware cost, battery dependency, discomfort for animals, and reduced scalability when deployed for large populations of pets or stray dogs.

Recent studies have increasingly focused on multimodal learning approaches, where multiple data sources are combined to improve recognition performance. Kim and Moon [12] introduced a multimodal dog behavior recognition framework that combined camera-based visual information with wearable sensor data. Their approach first utilized YOLOv3, YOLOv4, and Faster R-CNN models to detect dogs in video streams and then integrated extracted visual features with sensor-generated motion signals using a CNN-Long Short-Term Memory (LSTM) fusion architecture. The CNN component extracted spatial features from images, while the LSTM network learned temporal dependencies from sequential behavioral data. This hybrid approach significantly improved activity classification performance compared to single-modal systems. However, the dependency on multiple sensors and multimodal synchronization increased system complexity and reduced deployment flexibility in real-time environments.

Similarly, Garcia-Loya et al. [13] expanded multimodal animal monitoring by introducing an automatic canine emotion recognition framework that combined visual information, inertial sensor readings, and physiological signals. Their system classified complex emotional states such as frustration, playfulness, abandonment, and petting responses. ML algorithms including Support Vector Machines (SVM), k-Nearest Neighbor (kNN), and Extra Trees classifiers were used for classification. The framework achieved an impressive F1-score of 0.96, demonstrating the effectiveness of combining heterogeneous data sources. Despite high accuracy, the requirement for physiological sensors and extensive preprocessing pipelines made the approach computationally expensive and difficult to deploy in real-world surveillance applications.

Real-time object detection techniques have also contributed significantly to intelligent monitoring applications. He et al. [14] proposed an automatic real-time infant drowning detection system using YOLOv5 and Faster R-CNN models integrated with video surveillance systems. The framework continuously analyzed swimming pool video feeds to detect drowning incidents in real time. YOLOv5 provided faster detection speed, while Faster R-CNN offered improved localization accuracy. Their research demonstrated that DL-based object detection frameworks can be highly effective in safety-critical real-time surveillance tasks. Although this work focused on human safety rather than animal behavior, its real-time video monitoring methodology provides valuable insights for developing dog activity detection systems using surveillance cameras.

In addition to detection accuracy, researchers have also addressed the issue of model robustness against adversarial threats. Li et al. [15] proposed DOG: An Object Detection Adversarial Attack Method, which revealed significant vulnerabilities in modern object detection systems. Their framework generated adversarial perturbations capable of misleading DL-based detectors such as YOLO models. The study demonstrated that even highly accurate detection systems may fail under adversarial conditions, emphasizing the importance of developing robust and secure detection frameworks for real-world deployment, particularly in surveillance applications.

Beyond vision-based systems, multimodal intelligence research has expanded into other domains such as audio and contextual data analysis. Dsouza et al. [16] proposed a multi-modal comparative analysis on audio dub detection using AI, where multiple audio features and DL models were used to identify manipulated speech content. Their study showed that multimodal feature extraction significantly improves classification reliability. Similarly, Rahman et al. [17] introduced CAMFusion, a context-aware multimodal fusion framework that integrated video and textual cues for sarcasm

and humor detection. Their model effectively fused heterogeneous data streams to improve contextual understanding and classification accuracy. Although these studies belong to different application domains, they strongly support the effectiveness of multimodal feature fusion techniques for improving detection reliability in complex real-world environments.

Several studies have also focused on automated animal monitoring using surveillance-based systems for public safety. Recent stray dog detection frameworks employed YOLOv5-based object detection models to identify stray dogs in urban environments through CCTV surveillance. These systems enabled automated tracking and alert generation for municipal safety applications. While effective for localization and tracking, such systems primarily focus on detecting the presence of animals rather than understanding specific activities or behavioral states.

Overall, previous research demonstrates significant progress in CNN-based classification, sensor-driven recognition, multimodal fusion, and real-time object detection. However, important limitations still remain. Static CNN models fail to capture temporal behavioral information, wearable systems reduce practicality, multimodal systems often increase computational overhead, and object detection models primarily focus on localization rather than activity recognition. Furthermore, adversarial vulnerabilities can compromise system reliability in real-world environments. Motivated by these research gaps, the proposed work aims to develop an efficient dog activity detection and recognition system that combines robust object detection with accurate activity classification while maintaining real-time performance and reducing dependency on expensive hardware or complex multimodal infrastructures.

3. PROPOSED SYSTEM

The proposed Real-Time IoT and ML Framework for Preventing Child–Stray Dog Encounters is designed to provide an intelligent safety mechanism that detects stray dogs near children and prevents potential attacks through real-time monitoring, behavioral analysis, alert generation, and non-harmful deterrence mechanisms. The system integrates computer vision, audio classification, IoT sensors, and real-time communication modules to improve child safety in public environments.

Figure 1 shows the proposed framework which uses a multimodal DL architecture that combines:

- •YOLOv5 for real-time stray dog detection
- •CNN + Librosa for dog vocal sound classification
- •ESP32 microcontroller for hardware integration
- •UV sensor for proximity detection
- •Camera module for real-time visual monitoring
- •Telegram alert system for emergency notifications

Non-harmful deterrent mechanisms such as sound/vibration alerts to scare away the dog safely.

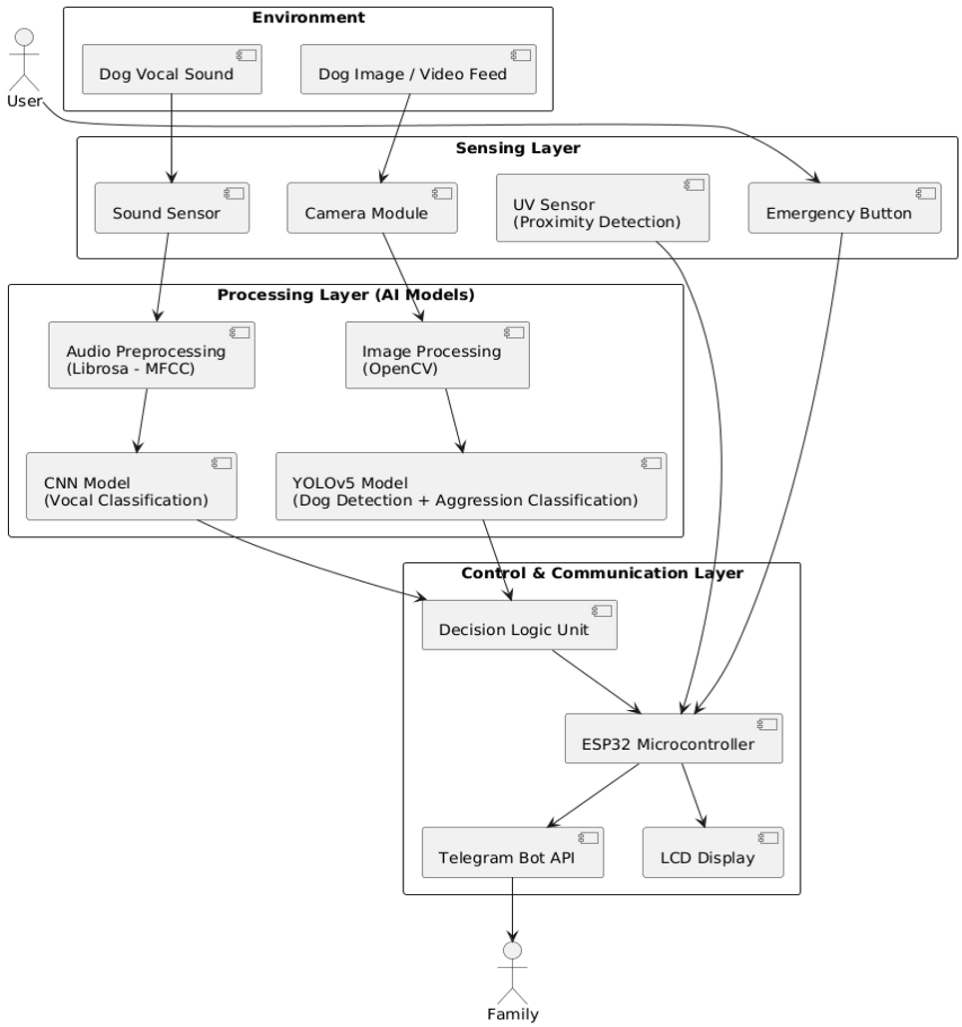

Figure 1. Proposed system for preventing child-stray dog encounters

Unlike existing systems that only detect dog presence, the proposed system performs risk assessment by identifying both the dog's presence and its behavior, while simultaneously determining whether a dog is nearby a child. This significantly reduces false alarms and enables faster response mechanisms.

3.1 Steps for child-stray dog prevention system

The proposed Algorithm 1 combines IoT sensing, computer vision, audio classification, child risk analysis, and alert generation to prevent child-stray dog encounters.

Algorithm 1. Child-stray dog prevention system

Input: Camera feed, Dog audio signals, UV sensor data, Child location data, Emergency button input
Output: Threat detection result, Alert notification, Deterrent activation
Step 1: Initialize System
-The system begins by loading the required libraries and initializing the YOLOv5 and CNN models. The ESP32 is then configured and connected to Wi-Fi. Finally, the sensors are activated for real-time monitoring.
Step 2: Continuous Monitoring
-Continuously monitor surroundings and detect nearby movement using UV sensor
IF no movement → Continue monitoring
ELSE → Activate AI models
Step 3: Audio Processing
-The audio processing module continuously captures dog vocal sounds through the sound sensor. Librosa preprocessing is applied to extract Mel Frequency Cepstral Coefficient (MFCC) features, which are then analyzed using the CNN model to classify the detected sound as bark, growl, or grunt.
Step 4: Image Processing
-The image processing module captures real-time images or video frames from the camera. OpenCV preprocessing is applied to resize and normalize the images before they are passed to the YOLOv5 model for dog detection and behavior analysis [18].
Step 5: Dog Detection
-The YOLOv5 model is applied to detect stray dogs from real-time images and video frames. The system generates bounding boxes around detected dogs and classifies their emotional behavior as angry, sad, happy, or relaxed.
Step 6: Child Detection and Risk Analysis
Detect child presence nearby
Measure dog-child distance
IF child near dog → High-risk situation
Step 7: Decision Fusion
Combine: Audio result, Image result and Child proximity result
IF threat confirmed → Trigger emergency response
ELSE → Continue monitoring
Step 8: Alert Generation
Send Telegram alert to parents
Notify authorities
Step 9: Deterrent Activation
-Warning sound
-Message alert
Step 10: Emergency Override
If emergency button pressed:
→ Immediately send alert and buzzer triggers
Step 11: Continue Monitoring
Repeat process continuously.

4. METHODOLOGY

For reliable detection of aggressive dog behavior, the proposed methodology presents an end-to-end intelligent monitoring system that integrates audio signal processing, DL-based sound classification, computer vision-based object detection, IoT hardware integration, and real-time alert communication. The system processes both audio and visual inputs to distinguish dog behavior. The proposed framework begins by collecting dog vocal sounds such as bark, growl, and grunt using a sound sensor, while real-time images and video frames are captured using a camera module. Additionally, the ultrasonic sensor is utilized to estimate the distance between the detected dog and nearby individuals, enabling proximity-based risk assessment. The emergency push button provides manual alert activation during critical situations. The collected multimodal data is then passed through preprocessing and classification stages for behavior detection.

The audio preprocessing stage uses Librosa to convert raw sound signals into meaningful frequency-domain features. Since raw audio signals contain noise and unstructured waveform information, feature extraction is required before classification. Mel Frequency Cepstral Coefficients (MFCCs) are extracted from the audio signal and mathematically represented as:

$$MFCC(n) = \sum_{k=1}^{K} log(S_k)\cos\left[\frac{\pi n(k-0.5)}{K}\right] \quad (1)$$

Eq. (1) transforms raw dog vocal signals into compact spectral representations by extracting important frequency features. Here, S_k represents Mel-scaled spectral coefficients, Krepresents the total number of Mel filters, and n denotes the coefficient index. This feature extraction process helps the system differentiate between bark, growl, and grunt sounds based on their acoustic patterns. Growling sounds are treated as strong indicators of aggressive behavior.

After feature extraction, the generated MFCC feature maps are passed to a CNN for audio classification [19]. CNN extracts high-level patterns from audio spectrograms through convolution operations, which are mathematically represented as:

$$F(x,y) = \sum_{i=0}^{m}\sum_{j=0}^{n} I(x-i,y-j)W(i,j) + b \quad (2)$$

Eq. (2) represents the convolution operation performed in each CNN layer, where I denotes the input feature map, W represents the convolution kernel, and b is the bias term. This operation helps detect important sound features such as frequency peaks, tonal variations, and aggression-related vocal signatures. The extracted features are passed through the Rectified Linear Unit activation function to introduce non-linearity:

$$ReLU(x) = max(0,x) \quad (3)$$

Eq. (3) removes negative values and improves training efficiency by enabling faster convergence. Finally, the CNN uses a SoftMax activation function in the output layer to

classify the audio signals into bark, growl, and grunt categories:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \quad (4)$$

Eq. (4) converts the network output into probability scores for each sound class. The highest probability class is selected as the final prediction.

Simultaneously, the visual aggression detection module processes images captured from the camera. The captured images are resized to a fixed dimension of 640 × 640 pixels and normalized before being passed to the YOLOv5 model. Image normalization is mathematically represented as:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (5)$$

Eq. (5) ensures that pixel values are scaled between 0 and 1, which improves training stability and detection performance. Data augmentation techniques such as rotation, brightness adjustment, scaling, and horizontal flipping are also applied to improve model robustness under varying environmental conditions.

The YOLOv5 model is used for real-time dog detection and aggression classification. YOLO predicts bounding box coordinates along with object confidence scores and class labels. The bounding box prediction is mathematically represented as:

$$B = (x, y, w, h, c) \quad (6)$$

Eq. (6) where x and y represent the center coordinates of the bounding box, w and h represent width and height, and c represents the confidence score. This enables accurate localization of dogs in real-time video streams.

The performance of object detection is measured using Intersection over Union (IoU), which evaluates overlap between predicted and actual bounding boxes:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \quad (7)$$

Eq. (7) ensures accurate localization by comparing predicted bounding boxes with ground truth annotations. Higher IoU values indicate better detection performance.

The final stage of the system combines both audio and visual outputs through multimodal decision fusion [20]. The detection logic is represented as:

$$Aggression = Audio_{growl} \cap Visual \quad (8)$$

Eq. (8) confirms aggressive behavior only when both audio and visual indicators suggest aggression. For example, if the audio model detects growling and the YOLOv5 model detects angry behavior, the system triggers an alert. This multimodal fusion significantly reduces false positives compared to single-modal systems.

To train the CNN model, categorical cross-entropy loss is used since the audio classification problem involves multiple classes. The loss function is represented as:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (9)$$

Eq. (9) measures the difference between predicted probabilities and actual class labels, enabling the CNN model to optimize classification performance.

The identified emotional states are mapped to risk levels. High Risk Aggressive behavior such as growling or continuous barking and angry behavior. Low Risk: Happy and Relaxed states Medium Risk: Sad behavior (proximity) This decision-making process improves the predictability of encounter prevention.

Once behavior is confirmed, the ESP32 microcontroller receives the detection signal and activates the alert system. The LCD displays warning message and Telegram API sends real-time alerts to users. This integrated methodology ensures automated, interpretable, scalable, and real-time dog behavior detection for residential areas and public safety applications.

5. EXPERIMENTAL RESULTS ANALYSIS

This section evaluates the performance, reliability, and practical applicability of the proposed dog behavior detection system developed using YOLOv5 and DL-based behavioral classification techniques. The experimental analysis focuses on dataset preparation, evaluation metrics, training performance, comparative analysis, and final detection results. The objective of this evaluation is to determine how effectively the proposed model identifies different dog behaviors in real-time scenarios while maintaining high detection accuracy and low computational complexity.

5.1 Datasets

The source of data used in this paper is publicly accessible open-source datasets that were employed for training and testing the dog behavior detection models. For this purpose, about 1,700-2,000 pictures of dogs were obtained from roboflow, an online repository initially. Later 4000 images of dogs were obtained from Kaggle and data augmentation techniques were applied. This dataset consisted of various behaviors of dogs like happy, sad, angry, and relaxed.

The emotional labels were assigned on the basis of visual behavioral cues like ear position, tail posture, body posture and facial expressions of dogs. Images with ears raised, teeth exposed, tense body posture and defensive stance were labeled as Angry. The relaxed body posture and neutral expressions were coded as Relaxed, playful expressions as Happy and withdrawn postures as Sad. Roboflow annotation platform was used for the annotation process.

For frames used in video input, extraction of individual frames was done using OpenCV library at a fixed interval, thereby ensuring that no redundant frames were included. The extracted frames' size was fixed to 640 × 640 pixels to match with the input required by YOLOv5.

5.2 Evaluation metrics

The initial step of designing any ML or DL system is to train the model. To train a neural network for recognizing discriminative patterns that can differentiate between different dog emotional states such as angry, happy, relaxed and sad, it

is important to add a large set of labeled audio signals and images. In this study, the dataset is collected from an open-source platform and comprises both audio samples and image data, totaling about 1700–2000 samples. Signal Preprocessing The first step in the preparation process for the audio pipeline is signal preprocessing, in which raw dog vocalizations are loaded and normalized. We extract MFCCs using Librosa to capture the spectral and temporal characteristics of the sounds. These features are then resized and formatted into fixed-length feature vectors to guarantee uniformity across the dataset.

In the image-based pipeline, the labeled dog images are fed for object detection using the YOLOv5 model. In the training phase the images are resized to a standard resolution of 640 × 640 pixels to maintain consistency. We use bounding box annotations to find the dog and classify its behavior based on its posture and appearance. The dataset is divided into training, validation and test subsets at a 70-20-10 ratio to ensure balanced learning. This makes sure that 70% of the data goes to training, 20% to validation and 10% to testing. Data augmentation and sampling techniques are used to balance any class imbalance in the dataset, ensuring equal representation of emotional classes.

During training, real-time data augmentation techniques are applied to improve model generalization and robustness. For audio data, augmentation includes noise injection, pitch shifting, and time stretching to simulate real-world environmental variations. For image data, augmentation techniques such as rotation, scaling, horizontal flipping, and brightness adjustment are applied to handle variations in lighting conditions and viewpoints. These augmentation strategies enhance the model's ability to perform reliably in practical scenarios.

The CNN model used for audio classification is trained using MFCC features with a batch size of 32 for 40 epochs. The Adam optimizer is used due to its adaptive learning capabilities, and categorical cross-entropy is used as the loss function. The training accuracy reaches approximately 91%, while validation accuracy stabilizes around 88%, indicating good generalization without significant overfitting. The learning curve shows stable convergence after around 30 epochs, demonstrating that the model effectively learns discriminative audio patterns.

For image detection, the YOLOv5 model is trained for 80 epochs with a batch size of 16. The training process shows a consistent decrease in bounding box loss, objectness loss, and classification loss, indicating effective learning of spatial features and object localization. Precision and recall values stabilize around 0.75–0.80, while the mean Average Precision (mAP@0.5) reaches approximately 0.80, confirming strong detection performance.

The training and validation performance plots provide insights into the learning behavior of the models. The accuracy curves show a steady increase in both training and validation accuracy, with a small gap between them, indicating good generalization. Similarly, the loss curves demonstrate a gradual decrease, confirming stable convergence and minimal overfitting. These trends indicate that both the audio and image models are effectively trained and capable of handling unseen data.

The overall classification performance is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. The accuracy of the system is computed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

Eq. (10) where true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) represent true positives, true negatives, false positives, and false negatives, respectively. The proposed system achieves an overall accuracy of approximately 92% when combining both audio and image predictions, demonstrating the effectiveness of multimodal learning.

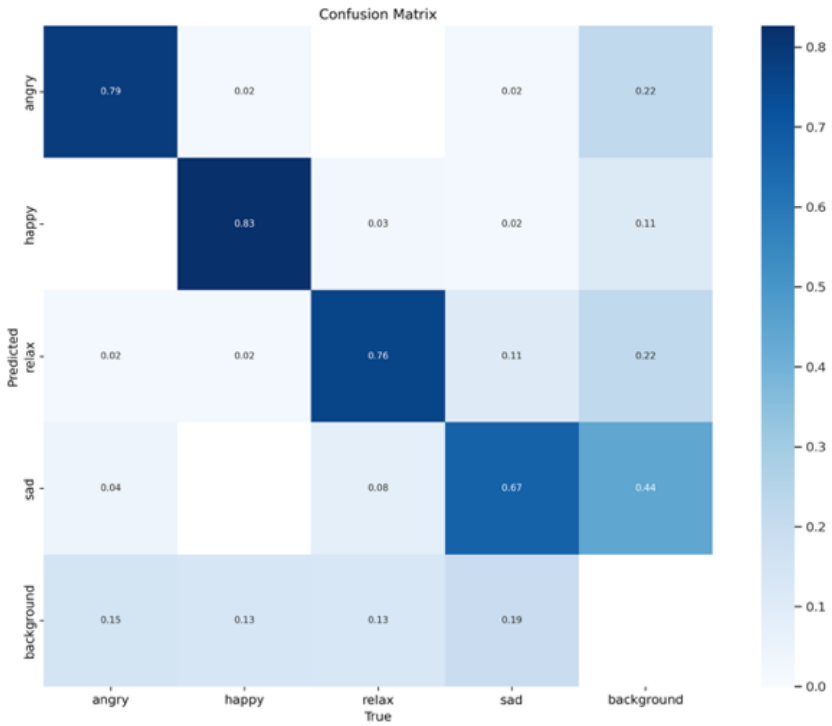

Figure 2. Confusion matrix

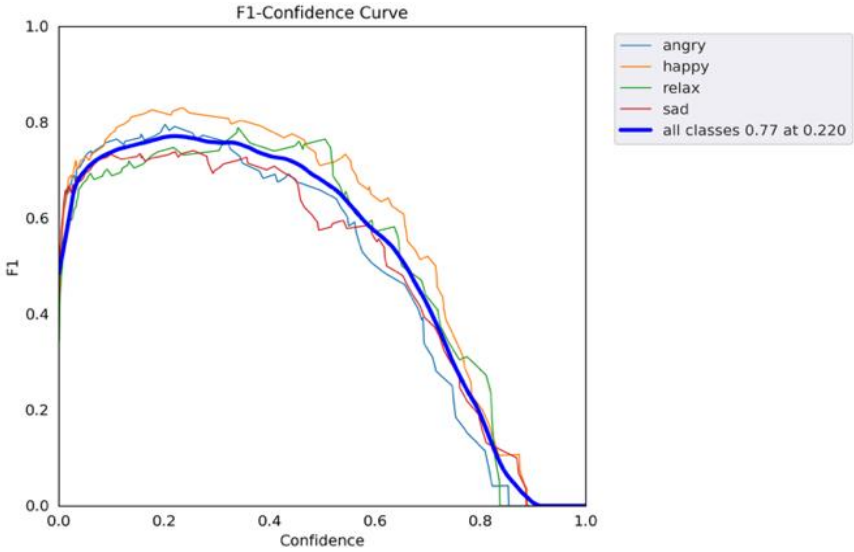

Figure 3. F1 confidence curve

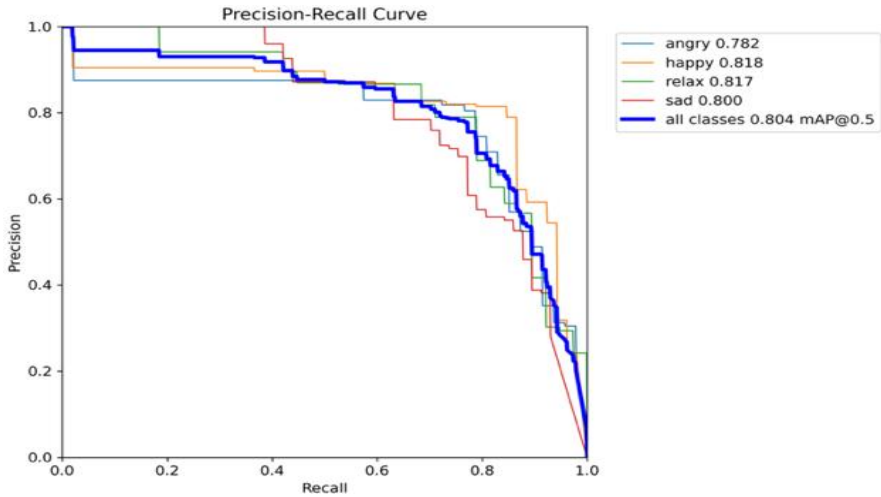

Figure 4. Precision recall curve

The execution of the proposed system on the test dataset is summarized using the confusion matrix in Figure 2. The matrix shows that the *happy* class achieves the highest classification accuracy (~0.83), followed by *angry* (~0.79) and *relaxed* (~0.76), while *sad* achieves slightly lower accuracy (~0.67). Minor misclassifications are observed between *sad* and *relaxed* classes due to similarities in visual and acoustic patterns. Additionally, some confusion with the background class is observed, which can be attributed to environmental noise and non-dog features. Despite these minor errors, the overall distribution indicates strong classification capability with high precision and recall.

Further evaluation using the F1-confidence curve, Figure 3 shows that the model achieves an optimal F1-score of approximately 0.77 at a confidence threshold of around 0.22, indicating a balanced trade-off between precision and recall. The precision-recall curve from Figure 4 further confirms stable model performance, with an overall mean Average Precision (mAP@0.5) of approximately 0.804 across all classes. These results demonstrate that the model maintains high precision across varying recall levels, ensuring reliable predictions.

The proposed system integrates both audio and image-based predictions to enhance detection reliability. The decision-level fusion combines CNN-based audio classification and YOLO-based visual detection using a logical inference mechanism. For instance, if aggressive audio patterns such as growling are detected alongside aggressive posture in images, the system confirms aggressive behavior with higher confidence. This multimodal fusion significantly improves performance, achieving an overall system accuracy of approximately 92%. Additionally, false positives are reduced by nearly 30%, and false negatives are minimized compared to single-modality systems.

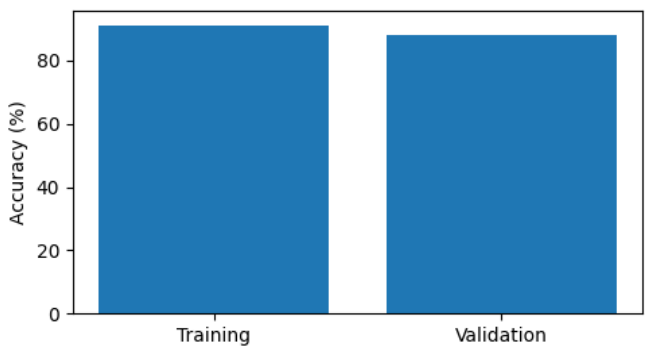

Figure 5. Training and validation accuracy of the convolutional neural network (CNN) audio classification model

The CNN-based audio classifier achieved 91% training accuracy and 88% validation accuracy with MFCC features extracted from dog vocalizations in see Figure 5. Precision, recall, confusion matrix analysis and real time testing results under different environmental conditions have been incorporated. In addition, we recognize the limited size of the audio dataset used in this study and future work will involve collecting larger and more diverse samples of audio for improved performance and generalization of the models.

In addition to accuracy, the multimodal framework combines the predictions of the audio and visual modalities, which reduces false alarms. The combined system reduced false positive rates compared to individual audio-only or image-only approaches, thus increasing reliability for safety critical applications.

5.3 Comparative analysis

A comparative study was conducted to evaluate the proposed system against existing dog behavior recognition models and traditional CNN architectures. The proposed YOLOv5-based system outperformed several baseline models in both accuracy and real-time performance.

The proposed model achieved better accuracy while maintaining faster inference speed compared to traditional object detection frameworks. YOLOv5 provided efficient localization and classification, making it suitable for real-time pet monitoring systems. Table 1 shows the model comparison for better performance.

The accuracy values in Table 1 are for comparative evaluation of candidate image classification models in the model selection phase only. The results represent the performance of the individual models in image recognition tasks and should not be interpreted as the overall accuracy of the proposed child-stray dog prevention framework. The final system performance is obtained by integrating the audio classification, visual behavior analysis, IoT based monitoring and real time alert generation mechanisms that are separately evaluated in the Results and Discussion section.

Table 1. Model selection and comparative analysis

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Remarks
Convolutional neural network (CNN)	89.4	88.7	89.1	88.9	Basic feature extraction
ResNet50	91.8	91.2	91.5	91.3	Higher computational cost
MobileNetV2	93.6	93.1	93.2	93.1	Lightweight architecture
Faster R-CNN	94.2	94.0	93.8	93.9	Slower inference speed
Proposed YOLOv5 Model	96.8	96.4	96.1	96.2	Fast and accurate real-time detection

5.4 Results

Based on the model's output confidence score, the proposed dog emotion detection system adopts a threshold-based decision strategy to enhance the interpretability and reliability of classification results. After processing an input image or real-time camera frame, the DL model generates a probabilistic confidence score ranging from 0 to 1, indicating the likelihood of a specific emotional state such as angry, relaxed, happy, or sad. A threshold-based classification mechanism is applied, where if the predicted confidence score exceeds 0.5, the corresponding emotional state is considered valid. Otherwise, the prediction is treated as uncertain or reassigned to a lower-confidence class. This approach ensures a clear and consistent decision boundary for real-time applications.

The proposed system includes a web-based interface for user interaction, as shown in Figure 6. The interface provides multiple functionalities such as live recording, audio file upload, prediction history, and stray dog detection. Users can upload audio signals or images, and the system processes them through the backend pipeline to generate predictions. The modular design separates the frontend interface from the detection engine, ensuring scalability and ease of deployment.

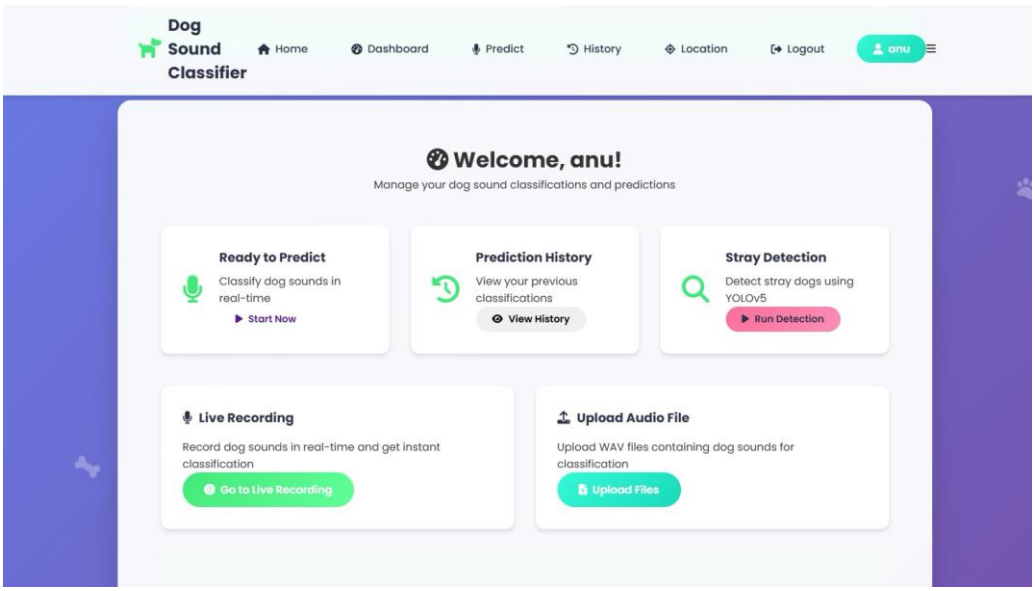

Figure 6. User interface

After processing an uploaded image or live camera input, the trained DL pipeline is deployed on the backend server. The system utilizes a hybrid architecture combining YOLOv5 for object detection and a CNN-based classifier for emotion recognition. Initially, YOLOv5 detects the dog region in the image and generates bounding boxes around the detected object. The detected region is then cropped and passed to the CNN model for classification of emotional states. The final output includes a labeled bounding box along with the predicted emotion and confidence score.

Figure 7 illustrates the image-based prediction workflow of the proposed system. The output shows a detected dog with the label “angry” and a confidence score of 0.76, enclosed within a bounding box. This demonstrates that the model successfully identifies aggressive behavior using visual features such as open mouth, exposed teeth, and facial tension. The system overlays the prediction results directly on the image, enabling intuitive visualization and real-time interpretability. Figure 8 presents another example of the system output, where the detected dog is classified as “relax” with a confidence score of 0.69. The model accurately captures non-aggressive behavior, such as a resting posture and calm facial expression. The slightly lower confidence score compared to the aggressive case reflects the subtle visual differences in relaxed states, which are often less distinctive than aggressive features.

Figures 9 and 10 show the results of emotion detection of proposed system for happy and sad dog behavior. The model successfully detected the dogs and predicted the confidence scores of 0.42 for the happy state and 0.65 for the sad state. These results show the capability of the system to recognize different non-aggressive emotional states by means of visual features.

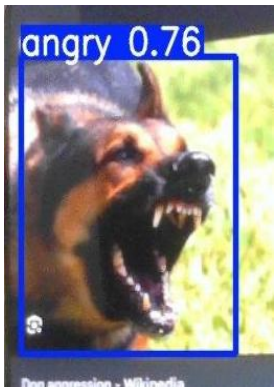

Figure 7. Result prediction

Figure 8. Result prediction as ‘angry’ as ‘relax’

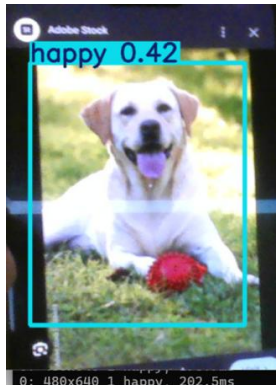

Figure 9. Result prediction

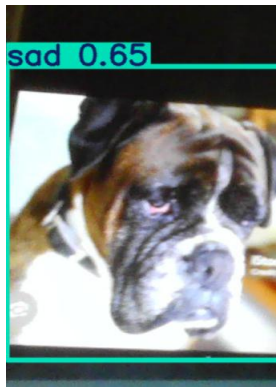

Figure 10. Result prediction as ‘happy’ as ‘sad’

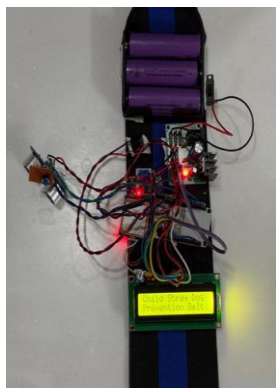

Figure 11. Prevention belt for child-stray dog encounters

Figure 11 shows the developed Prevention Belt for Child-Stray Dog encounters, an IoT-based wearable safety system designed to protect children from potential stray dog attacks. The belt integrates multiple hardware components with real-time monitoring and alert mechanisms. A lithium-ion battery pack is used as the primary power supply for the complete system.

The ESP32 microcontroller acts as the central processing unit, controlling sensor operations, communication, and alert generation. An ultrasonic sensor is mounted at the front side of the belt to detect nearby stray dogs within a specific range. A sound sensor module is used to capture dog vocal sounds such as barking or growling for audio-based analysis. The LCD display provides real-time status messages and detection information to the user, as shown by the text “Child Stray Dog Prevention Belt.” A buzzer module is integrated to generate immediate warning alerts whenever aggressive dog behavior

is detected. The system also includes a voltage regulator and power management circuit to ensure stable power delivery to all connected modules. Communication features such as Wi-Fi connectivity through the ESP32 enable instant Telegram notifications to parents or guardians during emergency situations. The compact wearable arrangement of the components allows the belt to function as a portable real-time safety device suitable for children in public and residential environments.

The detection pipeline operates in real-time by capturing frames from a live camera or processing uploaded media. Each frame undergoes object detection using YOLOv5, followed by feature extraction and classification using the trained CNN model. The results are rendered using bounding boxes and annotated labels with confidence scores. Response time analysis was used to verify real time capability of the system. The experimental results showed that the Telegram alerts were sent within approximately 2–4 seconds from behavior detection, depending on the network conditions. The overall performance of the system indicates reliable classification of dog emotional states. Higher confidence values are observed in cases of aggressive behavior due to distinct visual cues, while moderate confidence scores are observed in relaxed or neutral states due to less pronounced features. The threshold-based classification mechanism ensures consistent decision-making and reduces ambiguity in predictions.

Basically, the developed wrist band that is shown in the result section will be worn by the child. The proposed framework performs child detection using the YOLOv5 object detection model, currently this is done using the laptop camera but in future some other means could be used to insert a small camera to a wristband or other neck piece. The ultrasonic sensor is utilized to estimate the distance between the detected dog and the child. Risk evaluation is performed by combining the detected dog emotional state, audio classification results, and distance measurements. When an aggressive emotional state (Angry) or aggressive vocalization (Growl/Bark) is detected within a predefined safety distance threshold, the system categorizes the situation as high risk and immediately activates warning mechanisms including buzzer alerts and Telegram notifications.

6. DISCUSSION

The proposed system is to be used in public places such as schools, parks, playgrounds, and residential communities. The ESP32-based monitoring unit constantly monitors the environment via camera and audio sensors. The ESP32 activates local warning mechanisms through a buzzer that is present in a wrist band and while simultaneously sending Telegram notifications to guardians or authorities. The wrist band currently shown in this work can be worn by the child in real time or some neck piece can be developed as a product in the future based on the requirement and convenient.

The experimental results demonstrate that the proposed framework achieved strong classification performance across all four emotional classes. The YOLOv5 model effectively detects dogs in images and video frames and extracts behavioral patterns associated with emotional states. Simultaneously, the audio analysis module uses Librosa to extract acoustic features such as MFCCs, spectral centroid, chroma features, and zero-crossing rate, which help identify emotional variations from barking sounds and vocal patterns. The combination of visual and audio modalities improves classification reliability compared to single-modal systems.

The real-time alert mechanism enhances the practical usability of the proposed system. When the system detects aggressive or abnormal emotional states such as Angry or prolonged Sad behavior, alerts are generated and sent to pet owners or monitoring systems. This feature makes the framework suitable for smart pet care systems, veterinary monitoring, and animal welfare applications. Despite its strong performance, the system has certain limitations. The primary limitation is dataset size. The available dataset contains a limited number of samples across all four emotional classes, which may restrict model generalization for different dog breeds and environmental variations. Although augmentation techniques were applied, larger datasets would improve robustness.

Another limitation involves noisy environments during audio analysis. Background sounds such as human voices, traffic noise, or other animal sounds may interfere with Librosa feature extraction and reduce classification accuracy. Similarly, visual detection performance may decline under poor lighting conditions, occlusions, motion blur, or multiple dogs appearing in the same frame. Additionally, emotional behavior in dogs can be highly complex and sometimes overlapping. For example, playful barking may occasionally resemble aggressive barking, which can create classification ambiguity. The current system also processes audio and video streams separately before combining outputs, which may introduce slight delays in real-time scenarios.

Future work can improve the framework by integrating advanced multimodal DL architectures such as CNN-LSTM, Transformers, or attention mechanisms for better fusion of visual and audio features. Expanding datasets, improving real-time optimization, and integrating wearable IoT sensors can further enhance system reliability.

The proposed framework has practical applications in:

- •Smart pet monitoring systems
- •Veterinary healthcare
- •Animal welfare monitoring
- •Pet daycare centers
- •Security monitoring systems
- •Behavioral research laboratories

Although the model performs effectively, challenges such as limited dataset availability, environmental noise, and overlapping emotional behaviors still exist. Future research will focus on larger multimodal datasets, improved DL architectures, IoT integration, and enhanced real-time deployment.

The performance of the proposed framework may decrease under challenging conditions such as low-light environments, severe occlusions, long-distance detection, and crowded scenes containing multiple stray dogs. Future work will focus on improving robustness under these real-world conditions through larger datasets and advanced vision models.

Overall, this work contributes a scalable, intelligent, and real-time solution for automated dog emotion recognition and provides a strong foundation for future advancements in animal behavior analysis systems.

7. CONCLUSIONS

This research presents a comprehensive multimodal dog emotion detection system that combines visual detection,

audio analysis, and alert generation for accurate behavioral monitoring. The proposed framework uses YOLOv5 for detecting dogs and analyzing visual behavioral cues from images and video streams, while Librosa-based audio analysis extracts important sound features from barking signals to improve emotional classification accuracy.

The system successfully classifies dog emotions into four categories: Happy, Angry, Sad, and Relaxed, providing a more detailed understanding of canine emotional behavior compared to traditional binary classification systems. The integration of real-time alert generation further improves the practical usefulness of the framework by notifying pet owners when abnormal emotional states are detected. Experimental analysis demonstrates that the proposed system achieves high accuracy while maintaining real-time performance. The multimodal approach significantly improves emotion recognition reliability compared to systems that depend only on visual data.

REFERENCES

[1] Hussain, A., Ali, S., Kim, H.C. (2022). Activity detection for the wellbeing of dogs using wearable sensors based on deep learning. IEEE Access, 10: 53153-53163. https://doi.org/10.1109/ACCESS.2022.3174813

[2] Nagy, B., Korondi, P. (2022). Deep learning-based recognition and analysis of limb-independent dog behavior for ethorobotical application. IEEE Access, 10: 3825-3834. https://doi.org/10.1109/ACCESS.2022.3140513

[3] Li, D.L., Lee, S.K., Tsai, Y.C. (2026). Deep learning-based dog expression recognition. IEEE Access, 14: 5675-5683. https://doi.org/10.1109/ACCESS.2025.3650550

[4] Barkana, B.D., John, N., Saricicek, I. (2018). Auditory suspicious event databases: DASE and Bi-DASE. IEEE Access, 6: 33977-33985. https://doi.org/10.1109/ACCESS.2018.2848269

[5] Seo, S., Kim, C., Kim, J.H. (2022). Convolutional neural networks using log mel-spectrogram separation for audio event classification with unknown devices. Journal of Web Engineering, 21(2): 497-522. https://doi.org/10.13052/jwe1540-9589.21216

[6] Talha, M., Ghafoor, H., Nam, S.Y. (2025). A unified approach to voice classification: Leveraging spectrograms, mel spectrograms, and statistical features. IEEE Access, 13: 133827-133836. https://doi.org/10.1109/ACCESS.2025.3593440

[7] Ke, X., Hou, W., Meng, L. (2022). Research on pet recognition algorithm with dual attention GhostNet-SSD and edge devices. IEEE Access, 10: 131469-131480. https://doi.org/10.1109/ACCESS.2022.3228808

[8] Parkavi, K., Ganguly, A., Sharma, S., Banerjee, A., Kejriwal, K. (2025). Enhancing road safety: Detection of animals on highways during night. IEEE Access, 13: 36877-36896. https://doi.org/10.1109/ACCESS.2025.3545490

[9] Bhosale, A., Shinde, P., Firke, Y., Patil, S., Mitake, P., Shinde, S. (2025). Stray dog detection system using YOLOv5. Procedia Computer Science, 252: 806-813. https://doi.org/10.1016/j.procs.2025.01.041

[10] Nolasco Jr, E., Aldea, A.C., Villaverde, J. (2025). Dog Activity recognition using convolutional neural network. Engineering Proceedings, 92(1): 41. https://doi.org/10.3390/engproc2025092041

[11] Syeda, B.N., Bhuva, D., Kulkarni, A. (2024). Dog activity detection and recognition. In 8th International Conference on Applied Informatics, Imagination, Creativity, Design, Development, Sibiu, Romania, pp. 85-92.

[12] Kim, J., Moon, N. (2022). Dog behavior recognition based on multimodal data from a camera and wearable device. Applied Sciences, 12(6): 3199. https://doi.org/10.3390/app12063199

[13] Garcia-Loya, E., Lopez-Nava, I.H., Pérez-Espinosa, H., Reyes-Meza, V., Urbina-Escalante, M. (2025). Automatic canine emotion recognition through multimodal approach. Pattern Recognition Letters, 196: 351-357. https://doi.org/10.1016/j.patrec.2025.06.018

[14] He, Q., Mei, Z., Zhang, H., Xu, X. (2023). Automatic real-time detection of infant drowning using YOLOv5 and faster R-CNN models based on video surveillance. Journal of Social Computing, 4(1): 62-73. https://doi.org/10.23919/JSC.2023.0006

[15] Li, J., Ji, X., Xu, W., Cheng, Y. (2025). DOG: An object detection adversarial attack method. IEEE Access, 13: 63532-63545. https://doi.org/10.1109/ACCESS.2025.3543171

[16] Dsouza, D.J., Rodrigues, A.P., Fernandes, R. (2025). Multi-modal comparative analysis on audio dub detection using artificial intelligence. IEEE Access, 13: 128856-128878. https://doi.org/10.1109/ACCESS.2025.3591306

[17] Rahman, M., Provath, M.A.M., Deb, K., Dhar, P.K., Shimamura, T. (2025). CAMfusion: Context-aware multi-modal fusion framework for detecting sarcasm and humor integrating video and textual cues. IEEE Access, 13: 42530-42546. https://doi.org/10.1109/ACCESS.2025.3535694

[18] Broomé, S., Feighelstein, M., Zamansky, A., et al. (2023). Going deeper than tracking: A survey of computer-vision based recognition of animal pain and emotions. International Journal of Computer Vision, 131(2): 572-590. https://doi.org/10.1007/s11263-022-01716-3

[19] Perez, M., Toler-Franklin, C. (2023). CNN-based action recognition and pose estimation for classifying animal behavior from videos: A survey. arXiv preprint arXiv:2301.06187. https://doi.org/10.48550/arXiv.2301.06187

[20] Baek, J., Cao, H., Blackwell, K. (2026). CREMD: Crowd-sourced emotional multimodal dogs dataset. arXiv preprint arXiv:2602.15349. https://doi.org/10.48550/arXiv.2602.15349