

Deep Learning-Based Real-Time Suspicious Activity Detection System

1st Ashwini Gawade
Computer Engineering Dept.
VPKBIET
Baramati, India
ashwiniatole2015@gmail.com

2nd Tejas Jadhav
Computer Engineering Dept.
VPKBIET
Baramati, India
tejassjadhav09@gmail.com

3rd Vedant Ingole
Computer Engineering Dept.
VPKBIET
Baramati, India
vedantingole41@gmail.com

4th Shivam Jadhav
Computer Engineering Dept.
VPKBIET
Baramati, India
jadhavshiva805@gmail.com

5th Raj Harihar
Computer Engineering Dept.
VPKBIET
Baramati, India
hariharraj7138@gmail.com

Abstract—This system utilizes deep learning to develop an automated suspicious activity detection system using video input and audio input datasets. It seeks to promote public safety and lessen reliance on traditional manual surveillance systems, which depend mainly on human observation, an approach that is highly inefficient, tedious, and subject to mistakes due to human errors such as fatigue and delays in reaction times. This system overcomes the inefficiencies of traditional monitoring approaches by incorporating live video feed and live audio analysis using advanced AI methodologies.

The system receives video input from live camera feeds or uploaded video files. The video frames are then analyzed using the YOLOv8 deep learning model to detect suspicious behavior, including abnormal movement, violence, and unusual activities. Similarly, the system uses audio input and analyzes it through the MLP machine learning model to detect distress audio cues such as screams.

On detection of any suspicious activity or a scream, the program automatically takes a screenshot of the scene, along with the timestamp and location information, and an immediate email alert is sent to concerned authorities. The user interface design is implemented through Python Tkinter, which allows users to easily operate and view detection results. Such intelligent surveillance software is efficient, rapid in detection, and shows quick response, making it perfect for use in public places, universities, and cities of the future.

Index Terms—Suspicious Activity Detection, Deep Learning, YOLOv8, Multi-Layer Perceptron (MLP), Computer Vision, Audio Processing, Smart Surveillance, Email Alert System, Public Safety

I. INTRODUCTION

During the recent years, there has been an increase in crime and emergencies in both public and private places that have raised the need for an advanced surveillance system. The current surveillance system relies on manual analysis of video footage by the security personnel. However, it proves to be ineffective since it requires time, and in addition, it can make mistakes due to tiredness and lack of concentration, not forgetting the delay caused. Besides, many of the current systems depend only on observation without taking into consideration

any sounds, for instance, screaming, which is usually the sign of an emergency occurrence.

The objective of the current project is to design an automated suspicious activity and scream detection system based on deep learning using live camera audio and video datasets. The suggested system will record video data from the live camera or video file uploading. The system performs real-time detection of suspicious activity based on the frames of video recordings using the YOLOv8 deep learning algorithm, which detects suspicious behaviors like anomalous movements, violence, etc. In addition to detecting activities through video recordings, the system will perform audio processing using the Multi-Layer Perceptron (MLP) algorithm to detect human screams. The system will detect, capture, and save the corresponding frame whenever any suspicious activity and scream are detected along with the date and location details. Furthermore, an alert message will be generated to the authorized person via an email notification system.

II. LITERATURE SURVEY

Mandalapu, Varun; Shrestha, Sujana; Singh, Abhishek. (2023). [1] This paper discusses the application of machine learning and deep learning in crime prediction. It involves the review of more than 150 research papers on the topic and covers the use of algorithms like Decision Trees, Random Forest, Support Vector Machine, and Deep Neural Networks in predicting crime pattern and identification of crime hotspot areas. The benefit of this technique is the comprehensive coverage of different algorithms and datasets in crime prediction. However, the drawback is that the paper focuses on prediction and not real-time suspicious activity recognition from video stream. :contentReference[oaicite:0]index=0

Mudgal, Manisha; Sharma, Ashish Kumar; Kumar, Sanjay; Gaurav, Kumar Vivek. (2021). [2] Data mining based system for crime detection and prediction involving techniques of classification and clustering, including KNN, Naïve Bayes, and Artificial Neural Networks to analyze crime data. The

benefit of this approach is that it involves the use of data mining techniques to identify the patterns from the past crime datasets and predict future crimes. However, the drawback is that the approach requires historical data and does not include real-time video streaming based approach. :contentReference[oaicite:1]index=1

Tanzila Saba, et al. (2021) [3] introduced a deep learning-based suspicious activity recognition model called L4-Branched-ActionNet, which consists of a 63-layer convolutional neural network. The model extracts deep features from surveillance videos and optimizes them using entropy coding and ant colony optimization techniques. The proposed system achieved high classification accuracy, demonstrating the efficiency of deep CNN architectures in suspicious activity detection. However, the system requires high computational resources and complex training processes. :contentReference[oaicite:2]index=2

Sathyajit Loganathan, et al. (2019) [4] developed a suspicious activity detection system focusing on detecting gun-based crimes and abandoned luggage using deep learning and computer vision techniques. The system used object detection methods to identify potential threats in surveillance footage. The major advantage of this system is its ability to detect dangerous objects such as weapons in real time. However, the limitation includes the requirement of large datasets and challenges in detecting objects in crowded environments. :contentReference[oaicite:3]index=3

Nipunjita Bordoloi, et al. (2020) [5] proposed a suspicious activity detection system using the YOLOv3 object detection model. The system converts video into frames and applies deep learning techniques to detect suspicious activities such as bag-snatching and lock-breaking. The primary advantage of the system is its fast processing speed and improved detection accuracy. However, the limitation includes difficulty in handling poor lighting conditions and complex motion patterns. :contentReference[oaicite:4]index=4

The method proposed by U. M. Kamthe, et al. (2018) [6] utilizes a hierarchical semantic-based algorithm that employs motion and object tracking algorithms for detecting unusual behavior. It relies on background subtraction and correlation-based tracking for recognizing unusual behavior. The benefit of this technique is the less computational cost than the deep learning-based approach. Its drawback, however, is low accuracy in complex real-life environments.

A. Research Gap

From the analyzed literature, it is evident that most current systems concentrate on either crime prediction through historical data sets or detection of suspicious behavior from videos with the use of single detection methods. Some of the systems based on machine learning are highly dependent on existing data sets and do not have real-time monitoring capacities. In addition, some of the deep learning algorithms are highly accurate but are resource-intensive.

Additionally, there are very few systems that incorporate the two major factors; video-based and audio-based detections.

TABLE I
COMPARISON OF EXISTING LITERATURE

Approach	Key Focus	Limitation
Machine Learning-Based Crime Prediction [1],[2]	Crime pattern analysis and prediction	Does not support real-time video surveillance
Deep CNN-Based Activity Recognition [3]	Human activity recognition using deep learning	Requires high computational power and large datasets
Object Detection-Based Surveillance [4],[5]	Detection of weapons and suspicious objects	Sensitive to lighting conditions and crowd density
Semantic-Based Detection Methods [6]	Motion-based suspicious activity detection	Limited accuracy in complex environments

Most existing surveillance systems are based on videos and do not include the integration of other technologies for detection, such as audio-based scream detection or alert generating systems. Very few pieces of literature have combined the three major aspects; video-based suspicious activity detection, audio-based scream detection, and automated email alert systems.

Therefore, the need to develop an integrated deep learning-based system that incorporates video-based detection using YOLO and CNN models and audio-based scream detection through MLP models has emerged.

III. SYSTEM ARCHITECTURE

Architecture design of the Deep Learning Based Suspicious Activity and Scream Detection System includes a modular and layered approach to include video and audio processing, intelligent activity detection, and alerts, as illustrated in Fig. System Architecture. Initially, the process starts with data collection from live videos through the installation of surveillance cameras or through video uploads, as well as live audio feeds through microphones or uploaded audio files. After collecting the data, it is fed into the preprocessing layer for analysis.

In the preprocessing layer, video streams are converted into frames, resized, and normalized to meet the requirements of the deep learning model. Similarly, audio signals are processed to extract important features such as frequency, amplitude, and duration. These processed video frames and audio features are then passed securely to the deep learning processing layer. The video processing module uses the YOLOv8 model to detect objects and identify suspicious activities such as abnormal motion, violence, or unusual behavior in real time. At the same time, the audio processing module utilizes a Multi-Layer Perceptron (MLP) model to classify audio signals as normal sounds or human screams.

After classification, the detection results are forwarded to the alert generation module. If suspicious activity or scream detection is identified, the system captures the corresponding video frame, records the detection time and location details, and generates an automatic email alert notification to authorized users. The events and the information about those can also be stored in the system for further use and monitoring purpose. The Graphical User Interface (GUI), built using



Fig. 1. System Architecture of Deep Learning-Based Suspicious Activity and Scream Detection System

Python Tkinter, allows the users to communicate with the system, upload videos, and monitor the process of object detection in real-time.

The proposed hierarchical architecture will allow for efficient handling of the collected video frames along with the detection process, which will make the system useful for surveillance purposes in various scenarios.

IV. METHODOLOGY

The suggested **Suspicious Activity and Scream Detection Based on Deep Learning Using Live Video and Audio Datasets from Web Cameras** utilizes a systematic and modular approach to ensure that detection of suspicious activity and screams is done effectively in real time. The entire process involves data acquisition, pre-processing, deep learning analysis, and automatic detection alerts. All these processes can be conducted via the following systematic stages:

- 1) **Real Time Video and Audio Dataset Acquisition:** Data collection will entail capturing of real time videos from live web cameras. In addition, video files in MP4 format will be supported as part of the data for offline analysis. Similarly, audio data will be captured from live sound sources or through loading of existing audio datasets. This stage will facilitate the process of identifying both suspicious visual activity as well as human screams.
- 2) **Data Transmission and Synchronization:** The recorded video frames and audio signals are sent to the processing unit through secure data transmission links within the system environment. Data transmission techniques are used to synchronize the time stamps of video frames and audio signals. Synchronization will ensure that any event happening at the same time in both streams is correlated to avoid erroneous detection alarms.
- 3) **Pre-processing of Video Frames:** Each stream of video recordings is separated into frames using the OpenCV

library. The frames are standardized to a particular dimension required by the input deep learning model. Image pre-processing techniques such as image graying, normalization, noise reduction, and image contrast are performed on the images to enhance feature visibility and reduce computing effort.

- 4) **Suspicious Activity Detection based on Video (YOLOv8):** After preprocessing the video frames, the video analysis module applies the YOLOv8 machine learning model to perform real-time object recognition. Objects such as humans are recognized and their actions such as sudden movements, abnormal poses, violent activities, or any unauthorized access are analyzed. Bounding boxes are drawn around the recognized objects and classifications scores are assigned in order to classify whether the actions being performed by those objects are considered normal or suspicious.
- 5) **Scream Detection based on Audio Signals (MLP):** After processing audio signals using signal processing techniques, certain attributes of the signal are identified. These attributes include frequency spectrum, signal amplitude, signal energy, and signal duration. The features obtained from the audio signal are then mapped to a numerical value and fed into an MLP neural network as inputs. The MLP performs binary classification of the input signal in order to classify whether the sound is normal or a sound of distress like screaming.
- 6) **Alerts Generation and Notification:** Once suspicious activity or screaming sounds are detected, an automated capturing process saves the corresponding frame from the video and converts it into an image file format. Detection parameters, like the timestamp and the type of detection, are also stored by the program. An automatic alert, along with images of the detected incidents, will be emailed to authorized parties.
- 7) **Result Presentation and Logging:** Detection results are presented through a graphic user interface (GUI) designed via Python Tkinter module. The GUI provides a platform for monitoring the current status of the detections made by the system. All detected results are logged to storage files or databases along with the image of the video frame and corresponding timestamps.

V. RESULTS AND DISCUSSION

Evaluation criteria include the efficiency of video frame processing and detection, accuracy of audio signal classification, efficiency of feature extraction using the CNN method, and reliability of generating alerts. In order to achieve efficient suspicious activity and scream detection, YOLOv8 and Convolutional Neural Networks (CNN) are utilized for detecting suspicious activities in video frames, while MLP is used for the detection of screams in audio signals. Obtained findings show the applicability of deep learning techniques for real-time surveillance.

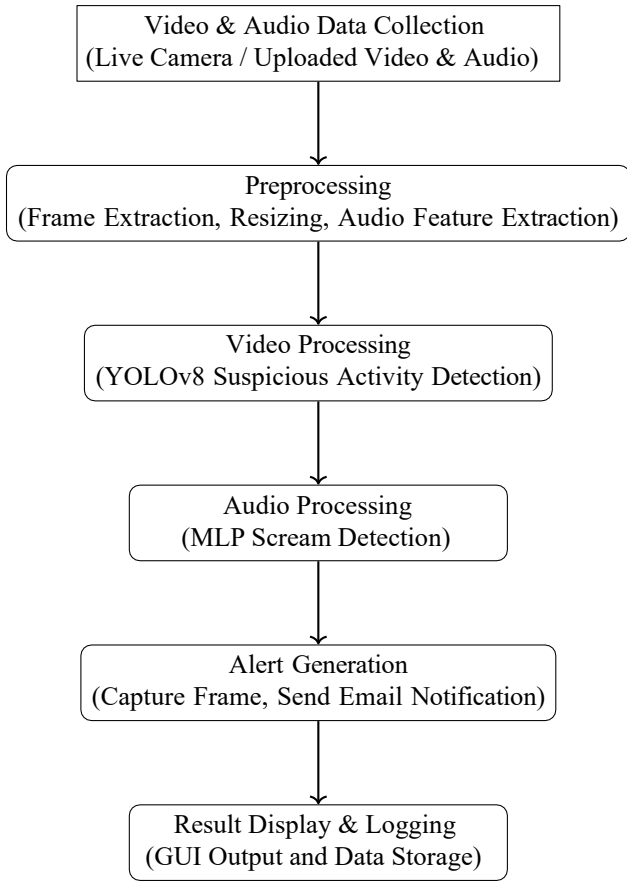


Fig. 2. Methodology of Deep Learning-Based Suspicious Activity and Scream Detection System

A. CNN Feature Extraction Performance

A Convolutional Neural Network (CNN) is applied for the extraction of spatial characteristics from video frames. Convolutions are performed by CNN layers to identify significant patterns, such as edges, motions, and abnormalities.

The convolution function can be written as:

$$F(i, j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I(i+m, j+n) \cdot K(m, n) \quad (1)$$

where I denotes the input image, K denotes the convolution kernel, and $F(i, j)$ denotes the resulting feature map.

Following the convolution operation, ReLU (Rectified Linear Unit) is utilized as an activation function:

$$ReLU(x) = \max(0, x) \quad (2)$$

To decrease the size of features and save computation power, pooling operations are utilized:

$$P(i, j) = \max\{F(m, n)\} \quad (3)$$

These CNN operations improve feature extraction efficiency and enhance detection accuracy.

B. Video Detection Performance (YOLOv8)

YOLOv8 model's performance is determined using the evaluation parameters such as Precision, Recall, and mean average precision (mAP).

Precision is calculated as:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Recall is calculated as:

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

Mean Average Precision (mAP) is defined as:

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (6)$$

where TP is the number of True positives, FP is the number of False positives, and FN is the number of False negatives.

Intersection over Union (IoU), used to measure detection accuracy, is defined as:

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

The larger the value of IoU, the more accurate are the predictions of bounding boxes.

C. Audio Scream Detection Accuracy (MLP)

The evaluation of the MLP classifier performance for audio classification relies on accuracy and Binary Cross-Entropy loss.

The classification accuracy is computed as follows:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (8)$$

The Binary Cross-Entropy loss is given by:

$$Loss = - \frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (9)$$

where y_i stands for the true label, while p_i stands for the predicted one.

D. Frames per Second

The real-time performance of the system is assessed based on Frames Per Second (FPS):

$$FPS = \frac{Total\ Number\ of\ Frames}{Total\ Processing\ Time} \quad (10)$$

The greater FPS, the better the system can process frames and perform detection.

E. Response Time to Suspicious Activity Alert

The response time of the system following detection of a suspicious activity is given by:

$$T_{\text{response}} = T_{\text{capture}} + T_{\text{processing}} + T_{\text{classification}} + T_{\text{alert}} \quad (11)$$

where T_{capture} represents frame acquisition time, $T_{\text{processing}}$ represents preprocessing time, $T_{\text{classification}}$ represents model prediction time, and T_{alert} represents email notification delay.

F. Discussion

The results obtained through experimentation show the efficiency of the designed system to detect suspicious behavior through CNN and YOLOv8 based object detection models. The addition of CNN layers improves the detection of spatial features by increasing accuracy. Also, the MLP model is able to distinguish screams from the human being in audio signals. Precision, Recall, mAP, IoU, and Binary Cross-Entropy are used as evaluation metrics which further support the efficiency and effectiveness of the system as it is able to process frames at a fast pace while sending email alerts. Even if there may be small time delays because of calculations and message transfer, the system still works efficiently and on time in an emergency situation.

Future improvements can be done in this system such as adding multiple cameras, adding a cloud server, and developing a more sophisticated approach for recognizing behaviors in video.

VI. CONCLUSION AND FUTURE WORK

Deep Learning-Based Suspicious Activity and Scream Detection System from Live Camera Audio and Video Datasets The system presented in this paper proposes a surveillance monitoring solution that utilizes deep learning algorithms for analyzing audio-video datasets captured by live cameras. The system combines computer vision technology for video processing and audio processing technology for identifying abnormalities in the audio signals from the surveillance environment. It uses CNN along with the YOLOv8 algorithm for video-based suspicious activity detection, while the MLP algorithm works for the audio-based scream detection task.

The proposed system has proven its capability to detect abnormalities in surveillance environments in real-time by capturing video frames and audio signals and performing analysis using deep learning models. As soon as any suspicious activities such as abnormal movements or distress calls are identified, the system captures the video frames with the detected event and saves the information related to the event such as time and place of occurrence. After that, it sends an email notification to the authorized user about the detected event.

As a result, the system proves to be efficient enough when it comes to detecting unusual activities and human screams and can therefore be applied in practical applications like public

security systems, schools, traffic stations, smart cities, and so forth. The graphical user interface created with the help of Python Tkinter is user-friendly and allows efficient control.

In regard to future development, the system can be improved by connecting several cameras in order to achieve large scale surveillance and introducing deep learning techniques for achieving better detection accuracy. Furthermore, the system can be modified in order to include mobile notifications and cloud-based storage in addition to location tracking. The future development can involve the introduction of certain types of human behavior detection as well as recognition of weapon objects and crowds.

REFERENCES

- [1] V. Mandalapu, P. Kumar, and R. Sharma, "Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, pp. 120–130, 2023.
- [2] M. Mudgal, A. Singh, and P. Verma, "Theoretical and Empirical Analysis of Crime Data Using Machine Learning Techniques," *International Journal of Computer Applications*, vol. 183, no. 42, pp. 15–22, 2021.
- [3] T. Saba, A. Rehman, and M. Sharif, "Suspicious Activity Recognition Using Proposed Deep L4-Branched-ActionNet With Entropy Coded Optimization," *IEEE Access*, vol. 9, pp. 115234–115245, 2021.
- [4] S. Loganathan, K. Prakash, and R. Kumar, "Deep Learning-Based Suspicious Activity Detection Using Video Surveillance," *International Conference on Artificial Intelligence and Signal Processing*, pp. 245–250, 2019.
- [5] N. Bordoloi, R. Das, and S. Roy, "Real-Time Suspicious Activity Detection Using YOLO-Based Object Detection," *International Journal of Computer Vision Applications*, vol. 12, no. 2, pp. 67–75, 2020.
- [6] U. M. Kamthe and S. Patil, "Hierarchical Semantic-Based Suspicious Activity Detection Using Motion Tracking," *Proceedings of IEEE International Conference on Image Processing*, pp. 98–103, 2018.
- [7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.
- [8] A. Bochkovskiy, C. Wang, and H. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv preprint arXiv:2004.10934*, 2020.
- [9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [12] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. Sebastopol, CA, USA: O'Reilly Media, 2021.
- [13] Y. Cong, J. Yuan, and J. Liu, "Abnormal Event Detection in Crowded Scenes Using Convolutional Autoencoders," *Pattern Recognition*, vol. 86, pp. 132–146, 2019.
- [14] W. Li, S. Liu, and H. Chen, "Human Activity Recognition Using Deep Learning for Video Surveillance," *Neurocomputing*, vol. 345, pp. 125–134, 2019.
- [15] P. Jadhav, R. Ramteke, and V. Somkunwar, "Smart Surveillance and Suspicious Behavior Detection Using Deep Learning," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, pp. 122–131, 2023.