WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Detection of Suspicious Human Behavior Using Deep Learning and Machine Learning leveraging IoT and Sensors: A Review

Shivansh*, Pradeep Chouksey, Parveen Sadotra, Mayank Chopra, Abhishek Bhardwaj

Department of Computer Science & Informatics, Central University of Himachal Pradesh, Kangra,176206, India

A B S T R A C T

In this study we explored the potential of Machine Learning and Deep Learning and their culmination with IoT and sensors to address the rising need for intelligent surveillance in sensitive and variable environments. Suspicion detection has become critical for public safety, especially in areas like defense zones, smart cities, crowded public places, schools, and transportation hubs. Traditional surveillance setups still mainly depend on people watching feeds, which can lead to mistakes, take up a lot of time, don't easily scale, and often aren't as efficient as they could be. With threats and crime on the rise, there's a real pressing need for smart, automated systems that can monitor in real-time with accuracy. This paper dives into how Machine Learning, Deep Learning, the Internet of Things, and sensors are transforming activity detection. ML and DL help spot suspicious behaviors by analyzing massive amounts of video footage and sensor data. Devices like cameras and motion sensors play a big part too, by gathering and sending real-time info from different spots. When these techs work together, they can totally change how we think about surveillance — making systems more adaptable, scalable, and reliable. The paper also points out what's new in the field, what challenges are still out there, the datasets researchers are using, and gaps that need filling. It focuses on blending ML, sensors, and IoT for real-world use cases. Looking ahead, possibilities like edge computing, combining data from multiple sensors, and lightweight models could make real-time human activity recognition even better. Down the road, this research could lay the groundwork for smarter surveillance in cities, defense, and other sensitive areas.

Keywords: Suspicious, Anomaly, Machine Learning, Deep Learning, Internet of Things (IoT), Video Surveillance, Real-Time Monitoring, Behaviour Analysis

1. Introduction

Machine Learning and Deep Learning have revolutionized the detection of suspicious activity where IoTs plays a significant role in data collection. Surveillance systems have become very popular but are still limited with data collection and traditional manual monitoring of video footage. IoT-enabled detection systems with multiple data sources like cameras, drones and sensors can help in taking more accurate and advanced decisions as just opposed to just relying on images and video frames. However, the final decision is still left to human expertise. Great work in machine learning and Internet of Things (IoT) technologies and demand for automated and real time surveillance has made the way for more intelligent and scalable solutions for activity recognition and anomaly detection. These systems take benefits from computational models to analyze human behavior, identify unusual patterns, and provide timely alerts for possible threats by integrating the information from multiple sources not being limited to single image or frame. This review focuses on the works and techniques used in suspicious activity detection and aims to provide a holistic understanding of the field by addressing various approaches and state-of-the-art technologies in suspicious activity detection. Machine Learning and Deep Learning leveraging IoT and sensors data will make the surrounding environment more understandable to the employed system and thus the decision-making authority multi factor eye keeping can make a more secure environment.

2. Key Aspects

2.1 Machine Learning and Deep Learning in suspicious activity detection

Machine Learning and Deep Learning are crucial in modern suspicious activity detection as these can process large amounts of data to find patterns suggesting suspicious activity. These solutions are fast, scalable and adapting for surveillance and improving upon existing methods

Various Machine Learning (ML) methods used for identifying presence, counts, location, tracking, activity, identity and particularly Deep Learning (DL) models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid models, are employed to analyze human behavior, identifying unusual objects and facial expressions. Deep learning models like CNNs works excellent at detecting spatial features in images, whereas RNNs and LSTMs work excellently at examining sequential data and identifying temporal patterns in activities. The integration of IoT devices and machine learning/deep learning models having heterogeneous data collected from various sources and supports swift, intelligent analysis.

2.2 Use of IOTs and Sensors

Various IOT devices are used to collect environmental data where the most common is videos from cameras like from public places, living environments, campuses and surveillance areas and then using these sequences of images fed into DL models to detect unusual behaviors like fighting, unauthorized objects and abandoned objects [1]. Some Systems even analyses real-time video to send immediate results [10]. Drones equipped with cameras are also used for aerial surveillance [12]. These act as mobile IoT devices capturing video and images from above for tasks like object detection. Sensors also play a crucial role in gathering different types of data just beyond video, like motion sensors for facial expressions recognition using dynamic geometric image (DGIN) [9]. Which are also efficient in computation and data storage.

There are Dynamic Vision Sensors (DVS) that detect human activity based on motion rather than full frames [4]. Ultrasonic sound sensors can be used for human presence detection [15,16]. LiDAR can be used as source of 3D spatial information, and their point clouds are processed to extract deep features [17]. PIR sensors detect changes in infrared radiation emitted by living beings, such as body heat thus can be used for motion detection and determining the occupant's direction of movement [15,18].

2.3 Data collection and use

Video frames are extracted and labeled for training DL models [2]. Kernel density can be used for clustering and predicting abnormal activities using datasets [3]. Further using the DL models can be used to judge human activities that can be ambiguous in certain environments [4].Motion dynamics such as stride length and speed, tailored for activity distinction (e.g., walking vs. running) from key body points (e.g., head, shoulders, wrists) can be captured for motion patterns across frames, providing a dynamic activity representation.[5] From 3D video data face scans can be taken and dynamic evolution of facial deformations can be captured for Expression Recognition[9] [10].

Framework that combines object detection and face detection from the image data helps in getting more accurate situation understanding. Ariel images are used particularly from drones for object detection [12.] Unsupervised learning can be employed to uncover hidden patterns in the sensor data and then find activity probabilities based on latent patterns.

Arranged in a 360-degree configuration kept activated can provide coverage to detect occupant motion and direction using machine learning algorithms. The sensor signal like ultrasonic sensor can also capture the characteristic movement signature of moving objects, caused by their separately moving parts called Spectrogram. LiDAR data can be used for detecting, tracking, and classifying humans [17].

Sensors and camera data can be fused together for 3D object detection this deep feature fusion strategy leads to improved detection accuracy, particularly for long-range objects, and demonstrates enhanced robustness against data corruption and out-of-distribution data, achieving state-of-the-art performance. Sensors being employed with systems that first identify regions of interest based on movement to process data efficiently and find the presence of humans [15,17,18,19].

Data collected from sensors being employed to machine learning algorithms can help in collecting occupancy information like presence, counting, location, tracking, activity, identity [13, 18]. Thus, enhancing the improve accuracy and broaden capabilities. Motion maps created from data from dynamic vision sensors' binary data can be used.

3. Methodology

This review paper produces a systematic review of state-of-the-art technologies used and being used in detection of suspicious activities in different environments. Which employs Machine Learning and Deep Learning for data analysis while taking data from various sources which includes IOT including sensors. This will help us to understand what the existing methods are, technologies and their fusions being used and how these can be enhanced further. Also, it will make us visit the gaps still there and can be worked upon further or improvised.

3.1 Literature Search and Selection criteria

Initially search for literature was conducted for the use of machine learning and deep learning for suspicious detection via object and human behavior detection. Several key terms used were "Suspicious Activity Detection using ML", "Suspicious Activity using Dl", "Object detection". Later search for literature was conducted for use of sensors and IOTs for the detection of human intervention and presence "Use of IOT in suspicious activity detection". This search was conducted on platforms like IEEE Xplore, Scopus, Research Gate and Google Scholar.

Papers relevant with use of technologies for detection of suspicious human activities or can help in detection of unusual human activities or their presence were chosen for study. During search papers were found which also discuss the fusion of different technologies and methods. Only the papers found on reputable journals, peer reviewed articles and surveys were used. Papers that focus on Methods and technologies which can contribute, improvise or expand the current systems and finding the gaps.

3.2 Data Extraction and Analysis

Data was extracted from each paper for technology, devices and methods used; method for data collection; how the collected data was employed of certain environmental surveillance. It is made sure that that study source's work leads to same aim however it may use different methods, sources, and data types keeping the study standardized for our purpose.

Extracted data includes key findings, work area, methods, future recommendations and technologies used to find the research gaps, possibilities of integration of technologies to improvise or create the system for future and advanced environments. However, this is a very dynamic and possibilities rich discipline but concluding the data from various work areas and using it for specific purposes still becomes difficult.

4. Objective and Significance of this review

4.1 Objectives

The idea is to implement multi source data for determining suspicious activity possibilities which can be done by adding multiple IOT and sensors' data while still using image processing which still have many limitations. Integration of IOT and further sensors can help produce more accurate results and surpass the limitations of object detection and human intervention. Such a system can even work better in variable light conditions, extreme weather conditions and unclear environments.

Ultra sonic sensors can help to detect human presence detection in smoke filled spaces and the distance estimation [16]. LiDAR sensors are efficient in detection, tracking and classification of people. These provide low- resolution shape and depth information which is considered critical which complement the high-resolution information form cameras and can improve accuracy and use of cameras [17]. This can help in maintaining computation like increasing data processing at detection of sensors. PIR sensors which take heat change data in bodies used for localization and triggering signal, activity detection, and motion detection can be used for people counting [15].

Several works directly address suspicious activity detection using deep learning techniques such as CNNs, RNNs, or hybrid CNN-LSTM models [2, 3, 7]. These models were trained on custom surveillance datasets involving activities like punching, snatching, or fighting, allowing for precise classification of abnormal behavior. Some studies also integrated deep learning into IoT environments for real-time surveillance and security, including smart homes and monitoring of children's behavior [3, 6, 14]. A few approaches, like in [13], explored semantic rule-based systems to detect suspicious behavior with lower computational requirements.

Human Activity Recognition (HAR) underpins many suspicious activity detection models by providing foundational classification of human motion patterns. There are HAR systems developed to distinguish between normal and abnormal actions in different contexts. These models often serve as the first stage in broader surveillance systems, offering crucial behavior understanding for downstream suspicion detection [4, 5, 8].

Some studies contribute indirectly by detecting emotional or expressive cues that may signal suspicious intent. Using facial expression recognition [9] and applied aerial object detection from broader perspectives [11]. Although not designed specifically for suspicion detection, such methods enhance the contextual understanding of human actions and can improve the results if implemented.

4.2 Significance

Integrating IoT sensors with image processing can significantly improve suspicious activity detection systems. While image processing alone has limitations, combining data from sensors like ultrasonic, LiDAR, and PIR enhances accuracy by providing critical contextual information in various environments, including low visibility or extreme conditions. This integration can lead to more reliable human presence detection, motion tracking, and behaviour classification, surpassing the limitations of traditional camera-based systems.

Deep learning techniques, particularly CNNs, RNNs, and hybrid CNN-LSTM models, have proven effective in classifying abnormal human behaviour. These models, trained on custom datasets, offer precise detection of suspicious activities like fighting or snatching, making them highly valuable for security and surveillance in smart environments. Real-time surveillance, especially in smart homes or childcare monitoring, benefits from deep learning's ability to analyse complex data efficiently.

HAR plays a foundational role in suspicious activity detection by helping to classify human motion patterns and distinguish between normal and abnormal actions. By accurately understanding human behaviour, HAR models serve as the first stage in broader surveillance systems, enhancing the identification of potential threats and suspicious behaviours.

Though not directly related to suspicious activity detection, systems that analyse facial expressions and emotional cues provide valuable insights into human intent. Recognizing negative emotions or stress signals can serve as an additional layer of context, improving the accuracy of suspicion detection systems when implemented alongside other methods.

5. Literature Review

The studies reviewed focus on building automated systems that detect suspicious human behavior in surveillance setups. Traditional monitoring methods often fail due to human error or attention fatigue. Making real-time, AI-powered and sensors leveraged methods for solutions are practical need.

At the core of these systems are deep learning models like CNNs, RNNs, and LSTMs, which process large volumes of video and sensor data to recognize patterns that signal abnormal or risky actions. These models are often supported by IoT devices and various sensors — including CCTV, drones, ultrasonic sensors, PIR, and LiDAR — to collect and analyze data in real-time.

Applications include detecting suspicious actions like fights or theft, recognizing human activities (HAR), locating people indoors, and spotting small or hidden objects — even in complex drone footage. A lot of research talks about using info from different sensors together, blending data to get better results, especially when visibility is poor or it's super crowded. Though we see real progress in detection and real-time responses, there are still challenges. Dealing with objects blocking the view, need for heavy computing power, and making sense of behavior in tricky or unfamiliar situations, weather obstructions are the major ones. Many researches discuss about using info from different sensors together, blending data to get better results, especially when visibility is poor and it's super crowded. But overall, it's clear that the push to develop smarter, more responsive security and surveillance systems isn't slowing down anytime soon. But overall, it's clear that the push to develop smarter, more responsive security and surveillance systems isn't slowing down anytime soon.

Table 1 – Literature Analysis

Title	Methodology	Conclusion	Future Work
"Deep Learning Approach for Suspicious Activity Detection from Surveillance Video" [1]	A pre-trained CNN model (VGG-16) extracts high-level features from each individual frame. The sequence of features from the frames is fed into an LSTM model, which classifies human activity as either suspicious (e.g., using mobile phone, fighting, fainting) or normal (e.g., walking, running). Based on the classification, the system predicts behavior. If a suspicious activity is detected, the system sends an SMS alert to the corresponding authority.	The system classifies activities as suspicious (like fighting or fainting) or normal. It achieved 87.15% test accuracy after 10 epochs using a combination of datasets. The goal is to automatically alert authorities.	Additional detection of activities such as objects, vandalism and trespassing. The system can be enhanced to identify individuals involved in suspicious activity. Advanced DL Features can make computation and accuracy better.
"A novel approach for suspicious activity detection with deep learning" [2]	Using Inception V3 (a CNN) for video frames and then passing them through an LSTM to catch pattern. It classifies actions as non-suspicious, semi-suspicious, and suspicious. For live video system is employed with asynchronous threads to handle video reading, frame sequencing, and recognition all at once.	Most actions were correctly identified, with only slight confusion between visually similar ones. The system handled live video streams effectively with a 6-second delay. Performed well to that of others. Thus, combining Inception V3 and LSTM proved well to detect suspicious human activities on various conditions.	Activity categories can be expanded for more real world activities and other hybrid combinations can be tried for better feature extraction Work can be done on reducing live stream delay.
"Suspicious activity detection using deep learning in secure assisted living IoT environment" [3]	A multi-classifier processes the sequence of video frames, and a deep neural network trains the data. Kernel density is utilized for clustering and data prediction. This research developed a system to detect suspicious activities, including both static (like prolonged standing or falling) and dynamic (like strangers approaching or touching children) and other events (fire, blood), in assisted living environments such as daycare centers.	It uses a deep neural network trained on video frames, a multi-classifier, and kernel density for clustering and prediction. The RKFD method achieved 98.88% accuracy on HHAR and AAR datasets.	The system needs further refinement and adjustment for the limited activities currently Broder range of activities can be in corporatized to improve the handling of more complex situations.

"Dynamic Vision Sensors for Human Activity Recognition" [4]	Method uses motion maps—2D projections like x-y, x-t, and y-t—of DVS video data as features. Convert event streams into 30fps videos. These maps show average pose, horizontal, and vertical motion over time. Then extract SURF features from the motion maps using Bag of Features (BoF), cluster them with k-means to create a visual vocabulary, and encode them as histograms. To capture appearance, include Motion Boundary Histogram (MBH) from optical flow. Combining motion maps with MBH gave the best results. While BoF histograms may lose spatial info, so combine these features and classify with a linear SVM. Tests were done on the UCF11 dataset, its DVS version, and a custom DVS gesture dataset.	The research emphasizes the promising application of Energetic Vision Sensors (DVS) in recognizing human activities. By efficiently detecting pixel intensity changes, DVS offers notable benefits such as reduced power consumption. The methodology involves generating motion maps from DVS data and applying advanced techniques like Bag of Features (BoF) and Motion Boundary Histograms (MBH). Results demonstrate performance levels comparable to traditional video-based systems when evaluated on datasets such as UCF11 and DVS Gesture.	Feature extraction methods that suit DVS's sparse high-resolution data and refinement methods to retain only meaningful movements and to eliminate noise. One way to achieve this is to use spatial correlograms to preserve spatial relations and improve the accuracy and detail. Another direction is to explore deep learning techniques (CNNs, RNNs) for spatiotemporal pattern discovery. For higher accuracy, we suggest using both BoF and MBH features. Dealing with the limitations regarding DVS data characteristics, and the computational Bof MBH extraction.
"Human activity recognition using combinatorial Deep Belief Networks" [5]	A dual deep learning framework with HAR that combines feature extraction with deep learning. It integrates both motion and static image features to improve accuracy. The system uses Modified Weber Local Descriptors (WLD) for motion features and Modified Local Binary Pattern (LBP) for static image details, fusing them with CNN before classification. It shows high accuracy on benchmark datasets. It tracks key body points for dynamic activity representation.	Address dimensionality problems by using two separate networks for classifying motion-related and static features. The system achieved high accuracy on benchmark datasets like HMDB51 and Hollywood, performing comparably to or better than existing methods.	Computational complexity is a general concern, particularly for real-time applications. More varying datasets need to be employed for more complex environments.
"An IoT Platform Based Deep Learning System for Human Behaviour Recognition in Smart City Monitoring Using the Berkeley MHAD Datasets" [6]	Development and evaluation of a deep learning system for recognizing human behavior in a smart city context. It uses the Berkeley MHAD dataset and CNN. The system is designed to operate on an IoT platform and was optimized by tuning parameters like batch size and epochs. The system involves testing and training phases. The testing phase includes video capture, normalization (using OpenPose for viewpoint correction), object recognition, data compression, and pattern extraction. It uses CNN and RNN to train the model. The CNN architecture includes 2 convolutional, 2 max pooling, and 2 full connection layers.	System developed for human behavior recognition in smart cities, achieving approximately 98% precision in classifying six human activities on the Berkeley MHAD dataset..	System can be improvised to identify more complex and potentially harmful behaviors, such as aggressive gestures, threatening actions, and climbing. Optimizing algorithms and exploring efficient deployment strategies will be crucial for real-time applications in integration of multiple sensors is a future step.
"A Deep Autoencoder-Based Approach for Suspicious Action Recognition in	The system frames of 112×112 from video and grouping them into 16-frame clips. 3D CNN (C3D) to capture both visual and motion data, giving a 4096-dimensional feature	Combination approach is used combining C3D for feature extraction which works very well and autoencoders for action detection and GAN for classification. A semi supervised	Guided research on implementing action detection approaches in a real-time environment is an important direction.

Surveillance Videos" [7]	vector. To detect unusual behavior, these features go through a CNN-based autoencoder trained only on normal actions—so when it sees something suspicious, it struggles to reconstruct it, and the high reconstruction loss flags it as abnormal. Finally, for classifying the type of suspicious activity, a GAN-based semi-supervised classifier uses compressed or latent features. Thus, GAN generates synthetic examples, and its discriminator learns to classify real vs. fake frames and recognize the action class using the labeled data.	approach is employed proved beneficial but with challenges due to limited labeled data of activities. Outperformed recent models across datasets with 97.5% accuracy on UT interaction, 89.6% on HCA, and a 47.34% on UCF crime (multi-class). The UCF crime AUC reached 80.6% for detection, showing reliable abnormal event spotting. Keeps processing time under 3 seconds per 16-frame making it useful for near real-time surveillance.	Considering motion information in future improvement can make system more efficient.
"Combining Public Machine Learning Models by Using Word Embedding for Human Activity Recognition" [8]	This unique method enhances HAR by combining pre-trained ML models without additional training. It uses word embedding techniques like GloVe to convert activity labels into vectors, averages them, and converts the result back to a word for a consensus label. This integrates predictions from multiple models effectively, showing better classification accuracy than individual models It can combine models with different label sets and generate new labels.	This approach, particularly the Weighted Mean Method, outperformed individual models and demonstrated the ability to combine models with different labels and infer new ones.	Work can be expanded with the use of sensors. Using embeddings trained specifically on HAR data could improve activity label representations and support for multi-label outputs is suggested to better capture complex activities. There is a need of specifically trained models for better alignment and sensitivity to the selection of embedding methods and dimensionality reduction techniques.
"Automatic 4D Facial Expression Recognition using Dynamic Geometrical Image Network" [9]	Geometrical Image Generation extracts 3D face scans and estimates differential geometry features like surface normal and curvatures. It creates geometrical images (DPI, NCI, SII). Dynamic Geometrical Image (DGI) Creation captures the dynamic evolution of facial surface deformations using rank pooling.	The Dynamic Geometrical Image Network (DGIN) for 4D facial expression recognition (FER) from 3D video data. DGIN creates geometrical images (Depth, Normal Component, Shape Index) using differential geometry to capture dynamic shape details and analyzes them with a multi-stage network structure and combined loss function. It achieved state-of-the-art performance on the BU-4DFE database.	Future work could combine geometry with things like textures or body signals and use model compression or hardware boosts to make DGIN run in real time. Handling issues like head movements or occlusion can make it more reliable. It could also be expanded to areas like emotion detection or mental health. Key challenges to address include data limitations, high computational needs, and sensitivity to facial changes.
"Real Time Video Analytics for Object Detection and Face Identification using Deep Learning" [10]	Face identification employs the dlib_facail_recognition library to generate 128-dimensional encoding vectors for images. The system gets the nearest image by calculating the difference between vectors and assigning label. The object detection module uses a "divide and conquer" approach, dividing the image into a grid. A vector 'Y' is generated for each grid, containing information about object presence, bounding box coordinates, and classes. Intersection over the Union (IoU) measures the accuracy of the predicted bounding box.	This work presents a real-time video analytics framework that combines object detection (using a grid-based method with IoU) and face identification (using dlib for encoding and comparison). It enables automated video analysis.	Using edge computing to cut down delay and bandwidth use. There's also a need to handle different data types, heavy processing loads, and make things work accurately even with limited resources. Cloud integration for deeper analysis is another area to explore.

"Enhancing object detection in aerial images" [11]	Enhancing object detection in aerial images via drones. It addresses challenges like small/dense objects and class imbalances using a three-step pipeline on the VisDrone-2019 dataset. The method involved data augmentation to handle class imbalance, an improved multi-column CNN with dilated convolutions to create density maps for focusing detection on probable object regions, and RetinaNet with adjusted anchors for better detection of small objects.	Small objects, dense overlaps, and class imbalance, a three-step methodology is proposed. The pipeline also involves image decomposition to minimize background noise.	Combining different object detection models for improved performance. Overlapping and extremely thin objects need improvement. Minimizing the impact of perspective distortion and occlusion. Improving the detection of certain classes for people bicycles is needed.
"Fusion of multi-scale attention for aerial images small-target detection model based on PARE-YOLO" [12]	PARE-YOLO is a modified version of YOLOv8, created to detect small objects in aerial drone images. It adds several enhancements. Restructured neck network which Enhances feature extraction and multi-scale fusion. Lightweight detection head optimized for small objects using a new P2 layer and ideas from RT-DETR.C2f-PPA structure which improves multi-scale feature representation. EMA-GIoU loss function Boosts model stability and handles class imbalance in complex scenarios. These changes help the model handle the common challenges of aerial imagery, such as: Complex environments, small size object s, Occlusions, lesser clear object features, cluttered background..	On the VisDrone2019 dataset, PARE-YOLO showed a 5.9% increase in mAP@0.5 over YOLOv8 and on the HIT-UAV dataset, it outperformed YOLOv8, YOLOv10, and RT-DETR in both robustness and accuracy. The model maintains a good balance of accuracy, speed, and resource efficiency, making it suitable for real-time detection of dense small objects.	Despite advancements, challenges like occasional omissions of small objects and misclassifications still exist. Future research focuses on optimizing PARE-YOLO's performance in lightweight configurations.
"Suspicious Activity Recognition in Video Surveillance System" [13]	The system uses a semantics-based approach with rules and conditions to understand and identify suspicious activities. It gets background image for reference, which is updated dynamically . Objects are detected using background subtraction. A relationship tracking algorithm tracks detected objects using colour histograms. By analysing motion features and spatial relationships, the system determines if an object is abandoned or if someone is loitering. The approach is computationally efficient because it avoids the training phase required by machine learning methods.	The approach identifies activities such as loitering, unauthorized entry, and abandoned objects by employing semantic rules based on human understanding. It was found to be computationally efficient and reliable for real-time surveillance, achieving high accuracy on standard datasets, especially for loitering and abandoned luggage detection.	The sources point out the need for better object detection and tracking. Blurring objects could also be explored. Since suspicious activity depends on context, customizable rule-based systems make sense. There's also a gap in standard, challenging datasets for testing these systems.
"Using Latent Knowledge to Improve Real-Time Activity Recognition for Smart IoT" [14]	Approach to enhance real-time activity recognition (AR) in smart IoT systems. Uses event-count sliding windows. It combines offline and online learning to improve feature representation. Sliding windows are based on a fixed number of sensor events. The system uses Temporal	Generating high-level feature representations from latent knowledge. This approach combines basic hand-crafted features with these learned high-level features, offering a balance of performance and computational complexity compared to some deep learning methods.	These networks can be used to spot suspicious activities. There's also a need for smarter algorithms that can tell the difference between normal behavior and real threats.

	attributes, sensor IDs, Previous activity, and Sensor mutual information as features. It employs Topic-aware Bayesian Prediction (TB) and HMM-based Prediction (HM) to predict activity probabilities based on latent patterns and transitions. These features are combined with training classifiers like Random Forest.		
"A Multi-Sensor Fusion Approach Based on PIR and Ultrasonic Sensors Installed on a Robot to Localise People in Indoor Environments" [15]	This has applied PIR sensors for motion detection by sensing infrared radiation from moving objects, people, under range of 7 meters with 120-degree detection angle. It also uses ultrasonic sensors, which measure distance by emitting 40 kHz sound wave and listening for the echo, having range of 43 mm to 11 meters. When motion detected by PIR, activates ultrasonic sensor for distance measurement from person. The system then processes this data using machine learning. A Decision Tree algorithm analyzes the PIR data to get direction of movement, while a K-Means clustering algorithm uses the ultrasonic data to analyze the distance and differentiate between people and stationary objects. This combination of sensors and algorithms helps determine person's exact location indoors.	Concludes that their multi-sensor system, using four PIR sensors and four ultrasonic sensors installed on a mobile social robot, effectively localizes an occupant in an indoor environment. The system is highlighted as non-invasive, low-cost, and privacy-respecting, requiring no infrastructural changes. Through tests, the work demonstrated the possibility of reconstructing occupant movement using the multiple-sensor system and proposed sensor placement	Future works include conducting a real-time application to test the system in a real-life scenario where the robot and occupant can move freely within a home environment. Further investigation can be done on the DT (Decision Tree) and K-Means algorithms to potentially increase accuracy in localizing occupants.
"An Ultrasonic Sensor For Human Presence Detection In Large Buildings" [16]	A sensor built with basic electronic parts that uses ultrasonic waves to detect if a person is present and moving, focusing on its ability to work even in dense smoke, making it useful for firefighters. The sensor is accurate enough to detect movement and recognize the unique movement pattern (signature) of someone walking by analyzing the signal's spectrogram. It uses a particle filter algorithm to detect when a person is moving through an area, like a hallway.	Sensors constructed from simple electrical components can measure movement with sufficient accuracy using ultrasonic measurements and spectrogram analysis to distinguish the characteristic movement signature of a walking person. This sensor was also tested and found to work in dense smoke conditions, suggesting its suitability for disaster management systems for the fire brigade, potentially extended to remember the locations of people who have fainted.	Significant work remains to extend the current system into an environmentally extended positioning solution to assist human detection.
"DeepFusion; Lidar-Camera Deep Fusion for Multi-Modal 3D Object Detection" [17]	DeepFusion uses two key techniques for fusion at deep feature level rather than input level: InverseAug: This technique reverses geometric changes (like scene rotation) made during data prep, ensuring that LiDAR points and camera pixels match up correctly. LearnableAlign: This uses a machine learning method cross-attention to figure out which camera features are most relevant for each part of the	LiDAR sensors provide shape and depth information, while cameras give high-resolution details about texture and shape. Combining both offers complementary insights for better 3D object detection. Effectively fusing lidar and camera data for multi-modal 3D object detection is possible. Well performed on the Waymo Open Dataset and this generic fusion method	While primarily focusing on lidar and camera fusion, the proposed method can be extended to other modalities such as range image, radar, and high-definition map. Performance might be further improved by adopting stronger baselines.

	LiDAR data, when one LiDAR area corresponds to multiple pixels in the camera image. This alignment improves 3D object detection.	applicable to different 3D detection frameworks.	
"Passive Infrared Sensor-Based Occupancy Monitoring in Smart Buildings: A Review of Methodologies and Machine Learning Approaches" [18]	This review focuses on methods for counting, localization (including direction and tracking), and activity detection using PIR sensors that provide Binary data to find person's location and further can help in knowing motion direction and Signal Based data that provides a more detailed analog signal that changes with the amount of infrared energy detected, which can offer information about an object's size, speed, and direction	PIR sensors are more suitable for localization tasks. In contrast, binary PIR sensors are mostly employed for activity monitoring in buildings and for identifying complex or unusual behaviors. Both versions are useful in the counting of individuals. The review highlights the growing importance of machine learning, and particularly deep learning, in developing additional features for PIR sensors concerning counting and localization, and further emphasizes strategic sensor placement and the number of sensors used. A multi-agent system where agents with different classifiers observe the sensor data for local predictions and collaborate for overall activity recognition. An improved variant, DCR-OL, adds online learning within interactions to enhance performance.	Investigating non-invasive sensors beyond PIR for tracking occupancy. Combining these with PIR in a hybrid system is worth exploring. Sensor fusion—using PIR with things like ultrasonic or environmental sensors—might boost accuracy and expand what can be detected. It's also important to improve PIR sensitivity and compare it with other infrared-based options.
"Fusing Dynamic Images and Depth Motion Maps for Action Recognition in Surveillance Systems" [19]	Utilizes RGB-D sensor data for human action recognition. It constructs Dynamic Images from RGB frames. and Depth Motion Maps (DMMs) from depth data, generating three different views: front, side, and top for a four-stream model. Each of these four streams is individually trained using a pre-trained VGG-F (an eight layer Deep Convolutional Neural Network Model). The VGG-F architecture is adapted for this purpose. Finally, the scores from the four trained streams are fused, and a SoftMax classifier is used to perform the final action classification.	Dynamic images generated from depth data for top, bottom, side are trained individually on a pre-trained model, and their scores are combined for final classification. Experimenting this approach on the UTD MHAD and MSR Daily Activity datasets achieved state-of-the-art results. They suggest the approach can be implemented in real-time surveillance environments and ATMs.	The work can be further extended by using other modalities to achieve better results. Approach is good for ATMs and real-time systems.
"A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras" [20]	Addressing human mistakes in surveillance and monitoring for irregular activities, particularly for violence. It utilizes ResNet50, which is a Convolutional Neural Network (CNN), apace with Recurrent Neural Network (RNN). As with CNNs, video frames created are provided as input to ResNet50 with frame differences to extract important features. These features are captured by a ConvLSTM network which learns spatial (CNN) and temporal (across frames) relationships to aid detection of abnormal events. To determine whether the event is normal or abnormal, max-pooling and fully	The proposed model, abnormal behavior in the UCF-Crime dataset. The results show that this method outperforms existing approaches on this dataset, despite challenges like variations in illumination, speed, subjects, and the short duration of some abnormal events. This demonstrates the model's effectiveness and its potential for improving automatic surveillance systems in real-world applications. Here a combined approach is implemented as a main tool for its potential to increase accuracy and reduce response time.	Improved work on classification of more anomalies. More intense work in variations in illumination, speed, and subjects in the dataset, as well as anomalies happening quickly within videos that mostly show normal behavior

	connected layers process the ConvLSTM output.		
"A Real-time Automated System for Object Detection and Facial Recognition" [21]	The suggested system uses a three-stage architecture. In the first stage, the EfficientDet model is applied for person detection in the given image. In the second stage, MTCNN (Multi-task Cascaded Convolutional Networks) Model is utilized to detect faces within the bounding boxes of the individuals marked in the first stage. The last stage is concerned with performing facial recognition using the FaceNet model by learning and comparing embeddings of the faces. This allows the system to recognize and name individuals.	After implementing the model on a limited dataset of images, the work resulted favorable outcomes. Works well in detecting and recognizing people in images.	Dealing with occlusion, which is a big challenge for detection and recognition. Network designs can help. Improving model performance through ensemble and transfer learning. Training on larger datasets is important, since the current one is too small for efficient working.
"Suspicious Actions Detection System Using Enhanced CNN and Surveillance Video" [22]	System for detecting suspicious actions, specifically shooting and stealing, in surveillance videos to provide on time warnings. introduces an Enhanced Convolutional Neural Network (ECNN) algorithm specifically for this purpose. The idea is to create a proactive system that can provide warnings and preventing incident	The experiment demonstrated that the ECNN algorithm's performance measures used like accuracy, precision, false-positive rate, and false-negative rate are comparatively better to that of standard CNN algorithm, thus achieving novelty for the proposed method	It should be improvised with diverse data sets to evaluate performance. Detecting untrained human actions in crowded scenes and the enhancing complexity of used ECNN compared to standard models.

6. Key Findings

Detection of suspicious in any environment with great results is a crucial task, especially in security and public safety maintaining surveillance systems. Thus, there raises a need for automated and lesser prone to errors. Development of such systems relies heavily on ML and DL
Data from IoT devices like cameras and other sensors is analyzed using machine learning (ML) and deep learning (DL) algorithms to find patterns that may differ from usual or expected behavior in any or certain environment and suggest possibly of suspicion and future suspicious activity. Deep learning models surpasses other techniques for their ability to automatically extract potent features from raw data, overcoming the traditional methods that rely on handcrafted features.

6.1 Use of Deep Learning:

Models like CNNs, RNNs, LSTMs, ConvLSTMs, Autoencoders, GANs, and DBNs are used for suspicion analysis and thus detection. One of their biggest strengths is that they learn features on their own from raw data whereas ML features are to be handcrafted. CNNs work great for spatial from image pixels, LSTMs and RNNs handle sequences like videos, and autoencoders flag anomalies through reconstruction loss. Various CNNs, like AlexNet, ResNets, VGG, and Inceptions, have different structures and layer counts. There is ResNets skip connections to jump over and pass its output to later layers while adding outputs of skipped layers too and creating alternate path, doesn't require recreate information within that block and focus is switched to learn difference and desired output, helping with parameter memorization and reducing vanishing gradients [20]. These models require a lot of data feeding but with enough examples, they perform well and thus there is a scope of enhancement for different scenarios.

Deep Learning stands out with ability to automatically extract features from raw data while traditional machine learning (ML) methods that rely on handcrafted features. DL models like ECNN and DCNN generally outperform classical classifiers used for categorizing or assigning labels to data instances [2] such as Naive Bayes, KNN, DT, SVC, and Random Forest in terms of accuracy and lower error rates. They're also more effective at handling high-dimensional data like images and videos, though they require larger datasets and higher computational resources. Traditional ML methods, while less data requirements and computationally lighter, may not match DL's performance on complex tasks. But hybrid approaches combining handcrafted and learned features show promise in balancing efficiency and accuracy.

6.2 Use of Machine Leaning:

Traditional methods like KNN, SVMs, Decision Trees, Random Forests, Gradient Boosting, Naive Bayes, and SVC have also been used in this area. They work on manually picked patterns from the data. Some lightweight motion-based methods don't even need ML training and still give real-time results. These models work great with simplicity and speed when resources or data are limited

Old-school ML methods remain highly relevant and effective in specific cases, particularly when applied to sensor-based systems for tasks like occupant localization or activity recognition. These are especially well-suited for scenarios where infrastructure is minimal. These can deliver robust results when features are well-defined. Moreover, traditional ML is often integrated into hybrid or novel systems, contributing meaningfully to state-of-the-art performance. It also plays a crucial role in intermediate processing steps, enabling the learning of latent knowledge to support downstream tasks. Even as deep learning gains prominence, traditional ML continues to serve as a baseline or comparative standard, emphasizing its enduring importance and practical utility, especially in applications requiring explicit training on structured data.

6.3 Contribution of edge computing

Edge computing devices are used alongside fog and cloud computing, offering flexibility in managing onsite collected information and analysis. It enables localized data processing, particularly basic image processing and sensors data processing, thus does not require to send data to other computation hubs and still decision can be made [10][18]. Edge computing is seen as having the potential to enhance real-time data processing and privacy by processing sensitive data locally rather than transferring the data [10]. Real-time video analytics, including applications like crowd counting, is a simpler example of that. It shows great potential direction with the idea of processing the data at site for normal decisions and sending the data to cloud or main computation hub for more precise and critical decisions [18].

6.4 Role Of data

Image Data (Video): Video surveillance cameras are a primary source of data. Image data provides rich visual information about scenes, objects, and human actions. CNNs are particularly effective at extracting spatial and visual features from images. Analyzing sequences of images over time allows for the detection of motion and temporal patterns, which is crucial for understanding activities and detecting anomalies. Video data is used for object detection, face detection and recognition, tracking, and recognizing human activities and behaviors.

Sensor Data (Non-Image): Data from sensors like PIR (Passive Infrared) and ultrasonic sensors, or sensors on smartphones and wearables, provide information about presence, motion, proximity, and activity state. PIR sensors can be used for occupancy monitoring, localization, and activity detection, especially in indoor and smart home environments. This data, often in the form of time series or binary signals, can be processed using ML/DL (like CNNs, LSTMs, GM-HMMs, Random Forests) to infer activities or detect deviations. Sensor data can be less intrusive and provide privacy-respecting monitoring. It helps in understanding human behavior and movement patterns.

Benchmark datasets play a crucial role in developing and evaluating ML and DL models for suspicious activity detection and human activity recognition (HAR), providing essential data and labels for training and comparison. Standard datasets like UCF-Crime, UT Interaction, HCA, and others offer a consistent basis to test model performance and compare results against existing methods, especially under real-world conditions. Like, UCF-Crime dataset includes real-time, untrimmed surveillance camera data having abnormal, illegal, and violent behaviors taken in public places, with 13 categories of anomalous events and a normal events category [20]. Such datasets are valuable because have real-time scenes from CCTV and very closely relatable cameras with varying circumstances, in short, such datasets are essential for exceptional decision making [20]. Deep learning models, being data-driven, benefit from the size and diversity of these datasets, which improve generalization across scenarios. However, challenges like varied camera angles, low-resolution footage, and limited labeled samples for rare events highlight the need for robust models and methods like semi-supervised learning, transfer learning, and data augmentation.

6.5 Data fusion

Fusion techniques can involve processing data streams separately and then combining results or processing features together. Multi-sensor fusion based on multiple classifier systems is also explored for HAR. Combining different data sources increases recognition accuracy compared to using single sources alone. Directs exploring integration of PIR sensors with other non-invasive sensors like ultrasonic and environmental sensors to develop more comprehensive systems. Combination of such data can also help in taking real time decisions and sending alerts

Integrating data from multiple sensors or modalities, known as sensor fusion or multi-modal approaches, is highlighted to achieve higher accuracy and more robust detection. Different data types can provide complementary information. For example, image data provides detailed visual context and object identity, while PIR sensors can provide information about presence, movement direction, and patterns less affected by lighting conditions.

Fusion of data of variable data types can be crucial for further developments of more accurate and robust detection systems like fusion of LiDAR and camera data. Fusion of Dynamic Images and Depth Motion Maps i.e. both colored and depth data when processed their final scores can prove to be as state-of-art work. Fusion of data from multiple sources like in areas like cities for videos is common for monitoring and can be extended with broader network. Sensor monitoring can be used for occupancy information and can help in complex situation analysis. A system using only IoT and sensor data but taking different features and from different types of sensors can enhance real time recognition.

6.6 Method fusion

Several approaches have been developed with mix of deep learning models and sensor data sometimes integrated with machine learning. Combining multiple models and Deep Learning models for object detection, face detection & alignment model, and face recognition model to detect people, align

faces, and then recognize them accurately. Image Feature Extractor with Sequence Processing Model is also used to classify human activities into suspicious, semi-suspicious, and normal classes while supporting real-time alerts, by extracting features and learning temporal patterns. Enhanced CNN models are also proposed to boost detection accuracy compared to other ML and DL methods. This proves out that combining multiple model and integration approaches have a great scope of further improvement discoveries.

Beyond deep learning, semantic-based methods depend upon motion patterns and object relations to define suspicious activity without heavy training. Some work uses deep belief networks with new descriptors to improve recognition across static and motion features. There are also models like DeepFusion that combine Lidar and camera data at the feature level, and RGB-D fusion techniques that use dynamic images and depth maps from different angles for better activity recognition.

For aerial surveillance, a multi-step pipeline improves small object detection using data augmentation, modified models, and specialized model like RetinaNet which helps to solve class imbalance problem . In smart home environments, models use sensor data with a mix of offline and online learning to detect activities efficiently. Others rely on multi-agent systems (DCR) that learn and adapt collaboratively or analyse sensor data using scan path trends for elder care monitoring.

One-Class Neural Networks offer a direct way to detect anomalies by learning from normal behaviour alone and still able to detect suspicious activities, making them suitable for situations where abnormal data is rare or hard to label and at controlled environments . All these approaches aim to make surveillance smarter, faster, and more reliable in real-world conditions.

6.7 Major factors for suspicion detection

Object detection is a fundamental in computer vision and can also be in suspicion detection, especially because objects do affect the security of interested environment. It is a crucial component in various applications like surveillance, security, and autonomous vehicles. Automated object detection systems can identify both living and non-living objects. Identifying Items and their presence in certain environment in accordance with time and later subsequent steps like tracking and classifying objects. But these require large, reliable, and standard datasets which are usually hard to obtain. In ariel images due to smaller, complex backgrounds, higher density and occlusion object detection becomes deficient but can be worked.

Human behaviour is a central theme analysing visuals and classify human actions in suspicious actions. Aiming to provide timely alerts for violence, loitering, fainting or unauthorised entry. It works well but still have challenges because such actions are context dependent and finding data of such rarely happening abnormal events is difficult. Both behaviour detection and object detection lack in complex weather conditions but sensors has shown the path to resolve such problems.

Face recognition is a crucial task along side object and human behaviour analysis . It has potential workspace for human mental state, face identification.

Presence and Occupancy details is fundamental in living environments, other features like speed, position, distance, optical flow, distance between multiple objects also play an important role in multi factor decision making and sometimes as complementary data

7. Conclusion

ML and DL are indispensable in automated suspicious activity detection while Benchmark and good datasets are important for developing and validating these models. Research is advancing through unique methods, hybrid models combining different ML/DL techniques and integrating DL with traditional methods, and through the fusion of data from multiple sensors or modalities (like video, depth, Lidar, PIR) to gain a more detailed understanding of scenarios and behaviors, leading to enhanced detection accuracy and robustness. Also edge devices don't lose space in making real time decisions and give new directions to send data to major computation hubs for better analysis of situation. In this direction there is still a lot of potential for future work. One may say that integrated work on technologies with improvement still have to be achieved.

The review and findings direct the work essentially towards integration of sensors primary and IoT devices with Deep Learning and Machine Learning where Machine Learning is still relevant since it is a great tool where the availability of computation is less and where we want to give a helping hand in analysis and decision making for better results. Use of such methodology will help in creating systems which have multi factor decision making, using lesser computation and using lesser resources. It also gives an idea of primary use sensors for normal decisions and secondary use more detailed data capturing or analysis on some threshold that may require higher computation power. There has been great work done on object detection but in weather conditions like foggy environments they lack, sensors can help in filling that gap. If object and behavior detection are integrated very early signs of suspicion can be found. Employing human and object detection by night vision cameras still have area to be explored. Object detection, depth imaging and behavior analysis in parallel can take detection far because both have the potential to find suspicion and together decisions can be more robust and accurate. In this area, depth imaging has been explored very little. Further understanding human expressions through temporal face data can help in understanding human mental state.

References

C. V. Amrutha, C. Jyotsna, and J. Amudha (2020). Deep Learning Approach for Suspicious Activity Detection from Surveillance Video, in 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Mar. 2020, pp. 335–339. doi: 10.1109/ICIMIA48430.2020.9074920.

N. Dwivedi, D. K. Singh, and D. S. Kushwaha (2023). A novel approach for suspicious activity detection with deep learning, Multimed. Tools Appl., vol. 82, no. 21, pp. 32397–32420, Sep. 2023, doi: 10.1007/s11042-023-14445-7.

G. Vallathan, A. John, C. Thirumalai, S. Mohan, G. Srivastava, and J. C.-W. Lin (2021). Suspicious activity detection using deep learning in secure assisted living IoT environments, J. Supercomput., vol. 77, no. 4, pp. 3242–3260, Apr. 2021, doi: 10.1007/s11227-020-03387-8.

S. A. Baby, B. Vinod, C. Chinni, and K. Mitra (2017). Dynamic Vision Sensors for Human Activity Recognition, in 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR), Nov. 2017, pp. 316–321. doi: 10.1109/ACPR.2017.136.

S. N. Gowda(2017). Human Activity Recognition Using Combinatorial Deep Belief Networks (2017) presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Computer Society, Jul. 2017, pp. 1589–1594. doi: 10.1109/CVPRW.2017.203.

O. O. Khalifa et al. (2022). An IoT-Platform-Based Deep Learning System for Human Behavior Recognition in Smart City Monitoring Using the Berkeley MHAD Datasets, Systems, vol. 10, no. 5, Art. no. 5, Oct. 2022, doi: 10.3390/systems10050177.

W. Ahmed and M. H. Yousaf (2024). A Deep Autoencoder-Based Approach for Suspicious Action Recognition in Surveillance Videos, Arab. J. Sci. Eng., vol. 49, no. 3, pp. 3517–3532, Mar. 2024, doi: 10.1007/s13369-023-08038-7.

K. Shimoda, A. Taya, and Y. Tobe (2021). Combining Public Machine Learning Models by Using Word Embedding for Human Activity Recognition, in 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), Mar. 2021, pp. 2–7. doi: 10.1109/PerComWorkshops51409.2021.9431141.

W. Li, D. Huang, H. Li, and Y. Wang (2018). Automatic 4D Facial Expression Recognition Using Dynamic Geometrical Image Network, in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), May 2018, pp. 24–30. doi: 10.1109/FG.2018.00014.

S. J. Patro (2019). Real Time Video Analytics for Object Detection and Face Identification using Deep Learning, Int. J. Eng. Res., vol. 8, no. 05.

V. Pandey, K. Anand, A. Kalra, A. Gupta, P. P. Roy, and B.-G. Kim (2022). Enhancing object detection in aerial images, Math. Biosci. Eng. MBE, vol. 19, no. 8, pp. 7920–7932, May 2022, doi: 10.3934/mbe.2022370.

H. Zhang, P. Xiao, F. Yao, Q. Zhang, and Y. Gong (2025). Fusion of multi-scale attention for aerial images small-target detection model based on PARE-YOLO, Sci. Rep., vol. 15, no. 1, p. 4753, Feb. 2025, doi: 10.1038/s41598-025-88857-w.

U. M. Kamthe and C. G. Patil (2018). Suspicious Activity Recognition in Video Surveillance System, in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Aug. 2018, pp. 1–6. doi: 10.1109/ICCUBEA.2018.8697408.

S. Yan, K.-J. Lin, X. Zheng, and W. Zhang (2020). Using Latent Knowledge to Improve Real-Time Activity Recognition for Smart IoT," IEEE Trans. Knowl. Data Eng., vol. 32, no. 3, pp. 574–587, Mar. 2020, doi: 10.1109/TKDE.2019.2891659.

L. Zhang, X. Wu, R. Gao, L. Pan, and Q. Zhang (2023). A multi-sensor fusion positioning approach for indoor mobile robot using factor graph, Measurement, vol. 216, p. 112926, 2023, doi: https://doi.org/10.1016/j.measurement.2023.112926.

T. van Groeningen, H. Driessen, J. Söhl, and R. Vôute (2018). An Ultrasonic Sensor for Human Presence Detection to Assist Rescue Work in Large Buildings, ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci., vol. 44W7, pp. 135–140, Sep. 2018, doi: 10.5194/isprs-annals-IV-4-W7-135-2018.

Y. Li et al. (2022). DeepFusion: Lidar-Camera Deep Fusion for Multi-Modal 3D Object Detection, in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2022, pp. 17161–17170. doi: 10.1109/CVPR52688.2022.01667.

A. Shokrollahi, J. A. Persson, R. Malekian, A. Sarkheyli-Hägele, and F. Karlsson (2024). Passive Infrared Sensor-Based Occupancy Monitoring in Smart Buildings: A Review of Methodologies and Machine Learning Approaches, Sensors, vol. 24, no. 5, Art. no. 5, Jan. 2024, doi: 10.3390/s24051533.

R. Khurana and A. K. S. Kushwaha (2021). Fusing Dynamic Images and Depth Motion Maps for Action Recognition in Surveillance Systems, Int. J. Sens. Wirel. Commun. Control, vol. 11, no. 1, pp. 107–113, doi: 10.2174/2210327909666191209155141.

S. Vosta and K.-C. Yow (2022). A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras, Appl. Sci., vol. 12, no. 3, Art. no. 3, Jan. 2022, doi: 10.3390/app12031021.

K. S. S. Reddy, G. Ramesh, J. Praveen, P. Surekha, and A. Sharma (2023). A Real-time Automated System for Object Detection and Facial Recognition, E3S Web Conf., vol. 430, p. 01076, 2023, doi: 10.1051/e3sconf/202343001076.

E. Selvi et al. (2022). Suspicious Actions Detection System Using Enhanced CNN and Surveillance Video, Electronics, vol. 11, no. 24, Art. no. 24, Jan. 2022, doi: 10.3390/electronics11244210.