IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, June 2025
Impact Factor: 7.67

# Deep Learning Approach for Suspicious Activity Detection from Surveillance Video

**D. B. Mane[1], Jayesh Dhumal[2], Anushka Harle[3], Parth Rananaware[4]**
Professor, Department of Information Technology[1]
UG Students, Department of Information Technology[2-4]
Smt. Kashibai Navale College of Engineering, Pune, India

**Abstract**: *With the proliferation of surveillance systems in smart cities, public spaces, and high-security zones, the demand for intelligent and automated monitoring solutions has become increasingly urgent. Traditional surveillance methods are largely reactive, requiring human intervention after an incident has occurred. Such systems are constrained by human attention span and are overwhelmed by the vast amount of video data generated daily.*

*This study introduces a deep learning-based framework for real-time suspicious activity detection using Convolutional Neural Networks (CNNs). The proposed system analyzes video frames in real-time, learning complex visual patterns to classify behaviors as either normal or suspicious. When an anomaly is detected, the system instantly triggers an alert.*

*The framework is designed for scalability, delivering high accuracy and resilience across a wide range of environmental conditions, including variations in lighting, crowd density, and scene complexity..*

**Keywords:** surveillance systems

## I. INTRODUCTION

The widespread presence of surveillance cameras in urban, commercial, and institutional spaces provides a promising foundation for strengthening security infrastructure. However, the practical utility of this extensive network is often limited by the conventional approach to monitoring, which still depends heavily on human operators. These operators are responsible for either observing live video feeds or conducting time-consuming post-event reviews. This process is vulnerable to human shortcomings such as fatigue, distraction, and insufficient staffing—factors that can critically undermine the effectiveness of surveillance efforts.

As surveillance systems generate a continuously expanding volume of video data, it becomes increasingly impractical to monitor all footage manually. The scale and frequency of data production outpace the capabilities of human analysis, making it difficult to ensure that every important event is noticed and responded to in time. This gap necessitates the introduction of intelligent solutions capable of automating surveillance tasks while maintaining a high level of accuracy and responsiveness.

Convolutional Neural Networks (CNNs), a class of deep learning models particularly well- suited for image and video analysis, offer a promising means to automate suspicious activity detection. These networks excel at learning spatial patterns and structures within visual data, such as identifying abrupt movements, uncharacteristic postures, or interactions that may signal abnormal or suspicious behavior. CNNs operate by passing input frames through multiple convolutional and pooling layers, enabling them to extract meaningful features from complex scenes. As a result, systems equipped with CNN-based capabilities can function with minimal human intervention, autonomously identifying anomalies within surveillance footage.

## II. RELATED WORK

1. Traditional Surveillance Methods: In the past, human operatives have always been involved in watching live footage and verifying incidents using video recordings retrospectively. These methodologies are hindered by human lapses, tiredness, and great volumes of footage making real time response difficult.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 3, June 2025 Impact Factor: 7.67

2. Rule-Based and Motion Detection Approaches: The first systems to come out were first deployed using algorithms that were rule based or only depended on motion detection. These are not effective in intricate settings and do not tailor to different scenarios.
3. Introduction of Deep Learning for Pattern Recognition: A particularly impressive application of deep learning is the use of Convolutional Neural Networks (CNNs) to analyze and classify visualcues, thus making this function useful for the development of video surveillance analytics.
4. Applications of CNNs in Surveillance: CNNs can be used to detect abnormal activities including trends that deviate from standard behaviors in public areas including hostile situations. Such technique encourages security agencies to monitor areas to identify potential threats prior to occurrence of incidents.
5. Studies on Real-Time Threat Detection: Most recent SIC & VDT studies urge the integration of CNNs in the development of systems where human inputs are not needed, eliminating beatings and slow modalities of threat detection. Continuous CNN's model integration deployment in the analysis will significantly reduce the resource dependency on the physical integrity of surveillance points.
6. Current Challenges in Deep Learning-Based Surveillance: Existing deep studying systems for surveillance face challenges such as high fake high quality prices, issue in adapting to various settings, and managing various environmental conditions (e.G., lighting adjustments, crowd density).

## III. LITERATURE SURVEY

Human Activity Recognition (HAR) is a growing field that plays a vital role in enabling intelligent surveillance, healthcare support, behavioral analysis, and automation in various sectors. Traditional surveillance systems are transitioning towards automated solutions, leveraging advancements in machine learning and deep learning to detect human behavior patterns. Early work by Velliangiri Sarveshwaran and Iwin Thankumar Joseph emphasized a foundational perspective on HAR, categorizing activities into economic (e.g., work-related tasks generating measurable output) and non-economic (e.g., activities providing mental satisfaction). Their study also laid out the potential applications of HAR in domains such as workplace evaluation, elderly care, and physical rehabilitation—highlighting its value in both commercial and humanitarian contexts.

In recent years, HAR has become a major application area for machine learning techniques. Pradipti, Shuvojit Das, and Somnath Nath emphasized this by showcasing HAR's role in biomedical engineering, sports analytics, and interactive gaming. Using sensor data from smartphones—such as accelerometers and gyroscopes—they demonstrated how traditional algorithms like Support Vector Machines (SVM), Random Forests, and Decision Trees could be trained to recognize specific human actions. The use of the UCI Machine Learning Repository was central to their study, providing a rich source of labeled sensor data for modeling various human movements. However, these classical approaches were constrained by their reliance on manual feature extraction and lacked the adaptability of modern deep learning models.

The shift from traditional machine learning to deep learning has been pivotal for HAR. Morsheda Akter and Shafew Ansary introduced a deep learning-based methodology that automates the feature extraction process using Convolutional Neural Networks (CNNs). Their research demonstrated how CNNs could leverage mobile sensor data to extract high- level semantic features, significantly improving model accuracy across diverse HAR applications—ranging from elder care to rehabilitation and general activity tracking. Their findings underscore how deep learning has made HAR systems more intelligent, requiring less human tuning while delivering more robust outputs.

Recognizing the importance of demographic-specific HAR systems, Ahatsham Hayat and Fernando Morgado-Dias focused on the elderly population. Their work addressed the physical decline and mobility challenges faced by older adults by designing a HAR system using smartphones equipped with gyroscopes and accelerometers. Their model monitored elderly individuals in both indoor and outdoor settings, aiming to detect subtle changes in activity patterns and trigger interventions when necessary. This application is particularly valuable in healthcare, where continuous monitoring can reduce risks such as falls and medical emergencies.

As the volume of digital sensor data continues to grow, researchers have begun comparing deep learning models for performance and efficiency. Lamiyah Khattar and Chinmay Kapoor conducted a comparative study of two prominent

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, June 2025
Impact Factor: 7.67

deep learning models used in HAR: 2D Convolutional Neural Networks and Long Short-Term Memory (LSTM) networks. By training both models on the same sensor-based dataset, they provided insights into the strengths of each approach—2D CNNs for spatial pattern recognition and LSTMs for capturing temporal dependencies. Such comparisons are crucial for system architects aiming to select the most appropriate models for specific applications.

Further innovations in HAR were introduced by Md Zia Uddin and Ahmet Soylu, who proposed a novel activity modeling system using Neural Structured Learning (NSL). Their approach fused multiple wearable sensor data streams to model the sequential nature of daily human activities. With the focus on eldercare and lifestyle monitoring, their system was designed not just for detection but also for predictive intervention—alerting users or caregivers before a potential hazard occurs. NSL added a new layer of learning that allowed the system to generalize across similar behavioral patterns while still learning individual- specific nuances.

Harpreet Kaur Lohia and Simran Kaur Dari also contributed to the evolving HAR ecosystem by reaffirming the value of supervised machine learning techniques for sensor- based activity recognition. Using data from accelerometers and gyroscopes collected via smartphones, they implemented SVM, Decision Tree, and Random Forest classifiers to recognize real-world human activities. Though based on classical methods, their work stressed the importance of data preprocessing and sensor fusion to boost the performance of non-deep learning techniques in environments where computational resources are limited.

Sensor-based activity recognition has also garnered attention for its cross-device reliability. Bolu Oluwalade and Sunil Neela explored this aspect by comparing data collected from smartphones and smartwatches using the WISDM dataset. Their use of Multivariate Analysis of Covariance (MANCOVA) revealed significant statistical differences in the behavior of sensor data across devices. These insights are vital for creating HAR systems that are device- agnostic and can provide consistent performance regardless of the hardware platform.

Snehal Wankhede and Dr. Sachin Chaudhari brought an interesting perspective by focusing on the visual perception of human activity. Their study emphasized the role of facial and body component detection—such as eyes, mouth, hands, and limbs—in improving HAR accuracy. This approach is particularly relevant for visual-based surveillance systems, where subtle movements or gestures might indicate intent or emotion. The fusion of facial recognition with posture and gesture analysis creates opportunities for highly nuanced behavior recognition in real-time video feeds.

Finally, Kah Sin Low and Swee Kheng Eng provided a comprehensive performance evaluation of HAR systems using deep learning models such as 3D CNN, LSTM, and hybrid combinations like 3D CNN-LSTM. Their experiments concluded that the hybrid model yielded the best accuracy, achieving a peak performance of 86.57%. The comparative analysis they presented not only validated the effectiveness of combining spatial and temporal modeling but also reinforced the value of continuous experimentation in model design.

Collectively, these studies demonstrate the evolution of HAR from sensor-based machine learning systems to sophisticated, multi-modal deep learning architectures. They reflect the versatility of HAR in domains ranging from healthcare and fitness to surveillance and public safety. Most importantly, they highlight the importance of designing systems that are adaptable, context-aware, and capable of functioning across a wide variety of environmental conditions and user needs.

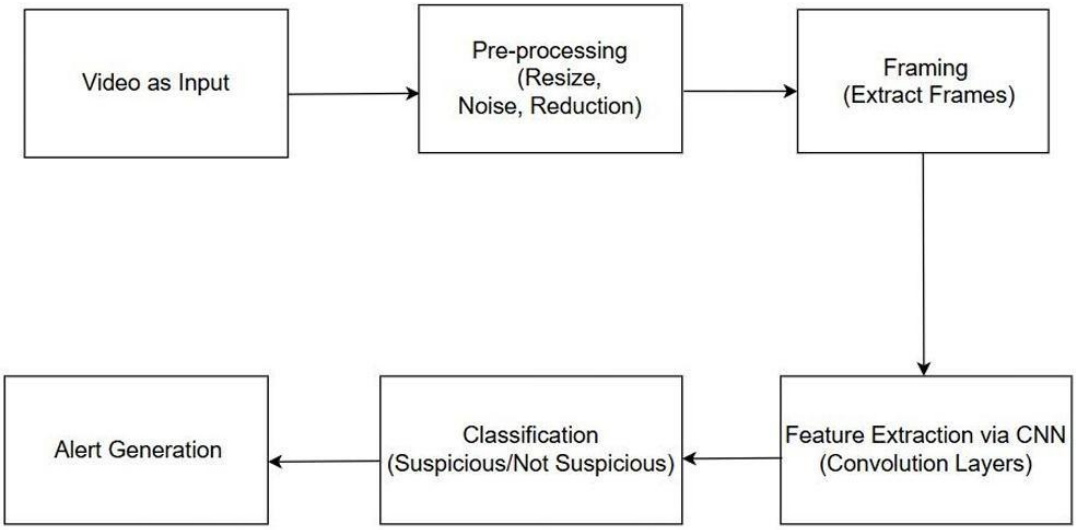
In the context of this research, these prior contributions form a robust foundation. They validate the feasibility of using deep learning—particularly CNNs—for behavior recognition and guide the development of an automated, scalable, and responsive suspicious activity detection system tailored for surveillance environments.

## IV. METHODOLOGY

### 1. Video as Input

The system begins by acquiring video footage from surveillance cameras installed in areas of interest such as public spaces, institutional premises, or commercial establishments. These video streams serve as the raw data for analysis. Given the diversity of surveillance environments, the input can include variations in lighting, crowd density, and scene complexity. The goal of this stage is to provide a continuous and realistic data stream that reflects typical and atypical human behavior.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, June 2025
Impact Factor: 7.67




**2. Pre-processing**

Once the video footage is collected from the surveillance camera, it is prepared for further analysis through a process called pre-processing. This step is important because raw video data can vary greatly in terms of size, quality, and lighting conditions, all of which can affect the accuracy of the system. In pre-processing, each video frame is adjusted to a uniform size so that the system can handle them consistently. Sometimes, the video might appear blurry or grainy due to low light or poor camera quality, so the system tries to improve the clarity by cleaning up the images and reducing unnecessary noise. Additionally, frames might look too dark or too bright, so brightness and contrast are adjusted to help the system clearly recognize what's happening in each scene. These improvements ensure that every frame is clear and standardized before it is passed on for further analysis. By preparing the video in this way, the system becomes better equipped to correctly understand human activities in different environments.

**3. Framing**

Once the video is cleaned and adjusted during the pre-processing stage, the system breaks it down into separate images, also known as frames. Instead of looking at the entire video at once, the system takes snapshots at regular time intervals. These frames give the system a clear look at what is happening at each moment. It's similar to flipping through the pages of a photo album to see how things are changing over time. By focusing on these individual moments, the system can better observe actions, detect changes in behavior, and understand the flow of events more clearly. This step helps the system spot anything unusual as soon as it begins to happen.

**4. Feature Extraction via CNN**

After the video frames are prepared, the system needs to understand what is happening in each frame. This is done through a process called feature extraction. In simple terms, the system looks closely at each frame to find patterns and important visual details that could help it identify human actions. It observes things like shapes, movements, and how different parts of the image relate to each other. To do this, the system uses a type of computer model known as a Convolutional Neural Network (CNN), which is very good at recognizing patterns in pictures and videos. The CNN learns by looking at many examples during its training and uses that learning to understand new frames. It doesn't need a person to tell it exactly what to look for—instead, it teaches itself which parts of the image are important for spotting suspicious behavior. This step is crucial because it turns complex images into meaningful information that the system can use to decide whether something unusual is happening.


DOI: 10.48175/IJARSCT-27588

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, June 2025
Impact Factor: 7.67

**5. Classification**

After the system gathers important information from each frame, it moves on to the task of identifying whether the activity shown is normal or unusual. At this stage, the system looks at the behavior happening in the frame and makes a judgment based on how typical or suspicious it appears. For example, a person walking calmly may be seen as normal, while someone running in a restricted area or acting aggressively may raise concern. The system carefully analyzes the movement and posture of the person to decide if the activity fits into what is generally expected in that environment. If something seems out of the ordinary, the system marks it as suspicious. This process helps the system focus only on actions that might indicate potential trouble, allowing for quicker responses and better use of attention from security personnel.

**6. Alert Generation**

When the system notices any unusual or suspicious activity, it quickly responds by sending out an alert. This alert is a way of informing the responsible security team that something may need their attention. The goal is to make sure that action can be taken immediately without having to wait for someone to notice it manually. These alerts help reduce response time and make it easier to prevent situations from getting worse. Instead of watching hours of video footage, security staff can focus only on moments that truly matter. This real-time warning system makes surveillance more effective and helps ensure that people and places stay safer.

## V. CONCLUSION

This research presents a comprehensive deep learning-based framework aimed at enhancing real-time surveillance through the detection of suspicious human activities. By leveraging the capabilities of Convolutional Neural Networks (CNNs), the proposed system effectively overcomes many limitations associated with traditional surveillance approaches. Conventional systems typically depend on continuous human monitoring, which is susceptible to attentional fatigue, inconsistency, and delayed reactions— particularly when managing multiple video feeds or large-scale environments.

The framework introduced in this study automates the identification of anomalous behavior patterns directly from video feeds. CNNs enable the extraction of meaningful visual features such as movement irregularities, posture changes, and object interactions, which are then classified to determine whether an activity is typical or suspicious. This approach significantly reduces the reliance on manual observation and improves overall surveillance efficiency. Moreover, the integration of real-time alert mechanisms ensures that relevant authorities are notified instantaneously when high-risk events are detected, facilitating swift and targeted responses.

Experimental validation of the model demonstrated strong performance in terms of detection accuracy, precision, and real-time inference speed. These results confirm the system's applicability in various security-sensitive settings, including public transportation hubs, educational institutions, corporate facilities, and high-traffic public spaces. The capacity of the system to operate across varied lighting conditions, environmental contexts, and crowd densities further reinforces its adaptability and robustness.

Despite these promising outcomes, the current implementation represents an initial step toward a fully autonomous surveillance ecosystem. Continued research is necessary to enhance the system's scalability, contextual awareness, and interpretability. Future iterations of the framework can benefit greatly from the integration of multimodal data sources—such as audio signals, thermal imaging, or WiFi-based motion detection—which can provide additional layers of context, especially in low-visibility or occluded scenarios.

Ultimately, the development of such intelligent surveillance systems holds considerable potential for societal impact. As threats to public and private security continue to evolve in complexity and frequency, there is a growing demand for surveillance solutions that are not only reactive but also predictive and preventative. Through ongoing refinement, real-world testing, and ethical deployment, deep learning-powered surveillance frameworks such as the one proposed here can become indispensable tools in ensuring safety, accountability, and preparedness in modern communities.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, June 2025
Impact Factor: 7.67

## VI. FUTURE WORK

To further enhance the capabilities, performance, and deployment potential of the proposed system, several key areas for future research and development are identified:

Spatio-Temporal Modeling: While the current system effectively captures spatial features using 2D CNNs, incorporating temporal dynamics can further improve accuracy in recognizing activity sequences. Future versions can integrate 3D Convolutional Neural Networks (3D CNNs) and Long Short-Term Memory (LSTM) layers to analyze both spatial and temporal characteristics of actions. This spatio-temporal modeling enables the system to understand behavior patterns over time—such as sudden running followed by an aggressive posture—which are essential for recognizing escalating threats or complex suspicious events.

Transfer Learning: To reduce training time and improve accuracy on limited or domain- specific datasets, transfer learning will be applied using pre-trained models such as VGG16, ResNet50, or EfficientNet. These models, trained on large-scale image datasets like ImageNet, provide a strong feature foundation that can be fine-tuned for surveillance scenarios. This approach is particularly beneficial for institutions with limited annotated video data, as it allows the system to inherit knowledge from vast and diverse visual datasets.

Multimodal Fusion: The current system operates on visual data alone, which may not always suffice—especially in low-light or occluded scenarios. To improve anomaly detection accuracy, future enhancements will explore multimodal data fusion by integrating inputs such as:

• Audio (e.g., glass breaking, screams)

• Thermal imaging (for night vision or heat detection)

• WiFi signal disruptions (to detect hidden movement) By combining these diverse modalities, the system can form a more holistic understanding of its environment and detect anomalies that might be visually ambiguous.

Privacy-Preserving Surveillance: Deploying AI surveillance in public and private spaces necessitates strong privacy safeguards. Techniques like federated learning—where the model is trained across decentralized devices without transmitting raw data—can preserve user privacy while enabling continual learning. Homomorphic encryption may also be employed to allow encrypted data processing without exposing sensitive information. These techniques ensure that the system remains ethically and legally compliant, especially in regions with stringent data protection regulations.

Scalability Testing: Finally, the system will be field-tested in high-density environments such as smart cities, transportation hubs, and large events to assess its scalability and reliability under real-world stress conditions. These pilot deployments will evaluate factors like network bandwidth, processing latency, alert accuracy, and user feedback to inform future optimizations. Such testing is crucial for validating the model's performance at scale and identifying infrastructure needs for wide-area rollouts.

These future directions aim to position the system not just as a technical solution, but as a foundation for next-generation, intelligent, ethical, and robust public safety infrastructures.

## REFERENCES

1] Human Activity Recognition Using Attention-Mechanism-Based Deep Learning Feature Combination. Sensors 2023, 23, 5715. https:// doi.org/10.3390/s23125715

2] "Performance evaluation of deep learning techniques" 2641 (2023) 012012 doi:10.1088/1742 6596/2641/1/012012

3] ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue VI June 2022 (Human Activity Recognition using Machine Learning)

4] Velliangiri Sarveshwaran et al. / Procedia Computer Science 204 (2022) 73–80

5] Human Activity Recognition for Elderly People Using Machine and Deep Learning Approaches. Information 2022, 13, 275. https://doi.org/ 10.3390/info13060275

6] "REVIEW ON HUMAN ACTIVITY RECOGNITION USING MACHINE LEARNING
and OPEN-CV PYTHON", Volume:04/Issue:02/February-2022 Impact Factor- 6.752 www.irjmets.com

7] Harpreet kaur lohia, Simran kaur Dari, "Human ActivityRecognition using Machine Learning Techniques", 2022 JETIR June 2022, Volume 9, Issue 6

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, June 2025
Impact Factor: 7.67

8] Md Zia Uddin1* and Ahmet Soylu, ”Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning” (2021) 11:16455

9] Lamiyah Khattar, Chinmay Kapoor, ”Analysis of Human Activity Recognition using Deep Learning”, DOI:10.1109/Confluence51648.2021.937711

10] Bolu Oluwalade1, Sunil Neela1, Judy Wawira, "Human Activity Recognition using Deep Learning Models on Smartphones and Smartwatches Sensor Data”,HEALTHINF 2021 - 14th International Conference on Health Informatics