A Framework for the Video Surveillance Suspicious Activity Detection

K. Rohitaksha
Department of Computer Science and Engineering, JSS Academy of Technical Education, Bengaluru, India
rohithaksha.k@gmail.com

Annapurna L. Pujari
Master of Computer Applications, JSS Academy of Technical Education, Bengaluru, India
annapurnalplj09@gmail.com

Shashank Dhananjaya
Department of Information Science and Engineering, The National Institute of Engineering, Mysuru, India
shashank@nie.ac.in, (corresponding author)

M. Narender
Department of Computer Science and Engineering, The National Institute of Engineering, Mysuru, India
narender@nie.ac.in

Received: 12 April 2025 | Revised: 17 May 2025 and 28 May 2025 | Accepted: 7 June 2025

 DOI: https://doi.org/10.48084/etasr.11437

ABSTRACT

Video surveillance is globally considered to be of considerable importance. Recent advances have resulted in notable improvements in the incorporation of artificial intelligence, machine learning, and deep learning techniques into video surveillance devices. The utilization of combinations and distinct frameworks facilitates the differentiation of various questionable behaviors through real-time image analysis. Human behavior is inherently unpredictable, making it difficult to determine whether it is suspicious or typical. This study characterized human actions into two categories: normal and suspicious. Normal actions include sitting, strolling, running, waving hands, etc., while arrest, abuse, shoplifting, etc., are examples of suspicious actions. This study used a convolutional neural network, achieving 97.96% accuracy on the CIFAR-100 dataset, demonstrating its effectiveness in recognizing and categorizing various activities, and paving the way for improved surveillance and security applications. Future work will focus on further refining the model and expanding its capabilities to include real-time video analysis, allowing more dynamic responses to potential threats and enabling faster decision-making in critical situations. Additionally, the integration of advanced algorithms for behavior prediction could further enhance the model's performance in complex environments.

Keywords-suspicious activity detection; profound learning; convolutional neural networks

I. INTRODUCTION

With the increase in video surveillance, manual supervision becomes ineffective. Detecting inconsistencies helps by filtering normal activity and notifying only suspicious events [1]. Social media has also seen a surge in claims of anomalies in surveillance footage, mostly technical faults or misinterpretations, but some may require investigation [2]. High-traffic public places, such as malls and airports, see increasing crowds due to urbanization, increasing safety risks [3]. Monitoring such areas is crucial, but individual operators often struggle to keep up, prompting the need for automated detection of unusual behavior [4, 5]. Each flagged event should be analyzed independently. Although some may seem extraordinary, conclusions should be based on data, not assumptions [6]. This work aimed to enhance surveillance systems by classifying 13 specific suspicious activities, going beyond traditional binary detection. The addition of real-time alerts and facial analysis adds practical value. Compared to previous works, such as AMDN [6] and optical flow-based methods [2], this approach offers improved precision and real-time multiclass recognition. The objectives of the proposed work are as follows:

- Identify inappropriate movements within a video by using their feedback.
- Promote efficient and reliable methods for perceiving suspicious behaviors, simplifying distortion localization, and increasing speed.
- Detect any sort of peculiarity in a reconnaissance camera.
- Raise an alarm when a strange action is recognized.

The problem statements are identified as follows:

- Utilize CNN for image recognition, Haar cascade for face detection, and the Slow-Fast CNN algorithm for movement detection. The model is trained on the DCSASS dataset [7], which comprises 16,853 videos (9,676 normal, 7,177 abnormal) with a 73-27% split between train-test data.
- The system classifies input or live-stream videos as normal or abnormal, identifying specific classes of abnormal activity, such as abuse, assault, robbery, and vandalism. It also detects the age, gender, and mask status of individuals in real time, helping to reduce false positives and enhance security response. Finally, abnormal activity alerts are sent through a mobile application.

II. LITERATURE REVIEW

AMDN [6] is a deep learning model that combines appearance and motion characteristics through stacked denoising autoencoders and one-class SVMs for anomaly detection in surveillance videos, outperforming existing methods on datasets such as the UCSD Pedestrian. In [2], anomalies were detected by capturing short-term motions using optical flow, modeling similarities among KNN with a Gaussian model to identify uncommon events. In [5], spatial-temporal deviations were analyzed in motion by tracking points of interest across video cubes to detect irregularities. In [4], background subtraction and morphological processing were used to detect suspicious movements by analyzing the displacement and size changes of segmented regions in indoor videos. Existing Suspicious Human Activity (SHA) detection methods combine background subtraction, feature extraction (motion, shape), object classification (SVM, Bayesian, Haar, KNN, face/skin detection), and threshold-based anomaly detection [8]. The HADE framework was developed for robust 3D human action recognition [9], recommending a stronger integration between HAR and computer vision [10]. Other related works include AI-based insider threat detection [11], ResDLCNN-GRU attention network for violence detection [12], and ASRNet, a 63-layer CNN model for anomaly classification [13]. In [14], the inadequate use of movement patterns and the inconsistencies of various datasets were analyzed.

III. PROPOSED FRAMEWORK

Many anomaly detection models are based on statistical or AI techniques [15]. Common types include threshold models, distribution models, and behavior models [16]. Existing systems classify videos as binary, normal or abnormal, whereas the proposed model uses 13 specific classes, such as abuse, arrest, arson, assault, and vandalism. For live streaming, if it detects weapons, it can send alerts through a mobile application, identifying gender, estimating age, and detecting mask usage. In videos, frames consist of two sections: static and dynamic. As frames pass, the static area remains absolute, whereas the dynamic area changes the object, background, and other elements. For instance, during a meeting, handshakes between two individuals are dynamic and fast-paced, while the background and other objects remain unchanged. This study designs a path to capture static information from videos with low frame rates and slow refresh rates. The fast path captures all dynamic data at high frame rates and a fast refresh speed. The formal path is very lightweight. Parallel connections merge both paths. For both paths, the Slow-Fast network uses the ResNet model and runs 3D convolution operations on it. The slow path uses large strides. Stride is defined as the number of frames skipped per second. The slow path typically allows two sampled frames per second, while the fast path uses a small stride to allow 15 frames per second.

A. Video Tests

Tests were carried out with the help of images taken by a camera.

$$X \rightarrow Input\ Image\ Set$$

During the feature extraction phase, the grayscaled image frame set is defined as $X = X_i$ for $i = 0, 1, 2, \ldots, n$. X is the input image set, and X_i is the i^{th} image with n frames.

B. Video Preprocessing

Video processing has several subcomponents to correct the input image in a variety of ways, including removing noise, outliers, dimensionality, and data, among others.

C. Training on Images

The image dataset is split into training and testing subsets to validate the proposed model, evaluate its accuracy, and determine whether a particular image is typical or unusual. To remove noise and improve the image, it is pre-processed by using explicit procedures, such as histogram balance and the median filter.

$$X \rightarrow Preprocessed\ Image\ Set$$

Figure 1 shows the anomaly detection and classification procedure. The anomaly detection algorithm receives the preprocessed data to detect any uncommon data patterns. The distinguished oddities are then provided to the classification module. The user is then presented with the results of the anomaly detection and identification algorithms. Figure 2 illustrates how data are processed as they pass through the system through several transformations.

IV. METHOD

The median filter is used as a noise-removal technique, which is a non-linear digital filtering technique often used to remove noise from an image or signal. Zeros are appended at the edges and corners of the matrix that represents the grayscale image. Then, for every 3×3 matrix, elements are arranged in ascending order to find the median/middle element of those nine elements.

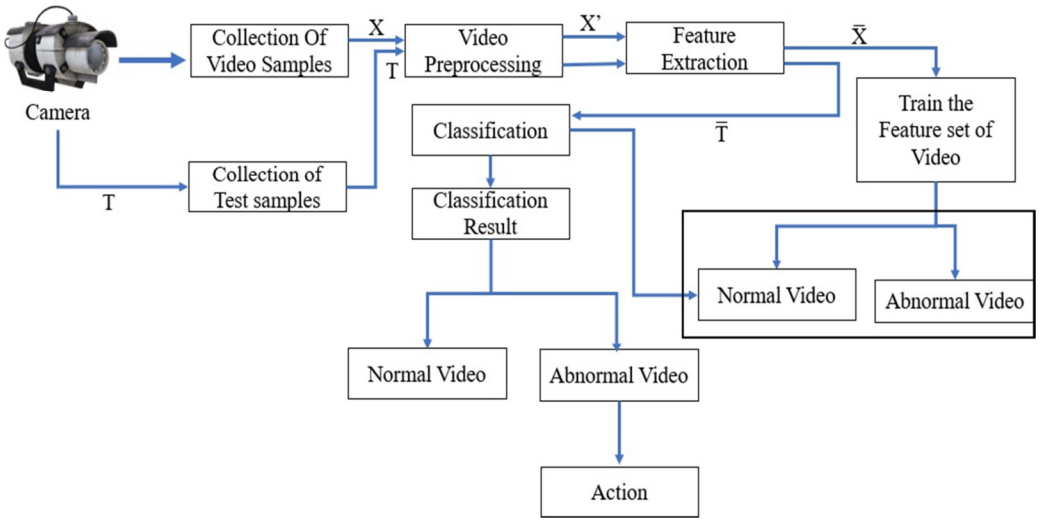

Fig. 1. System architecture.

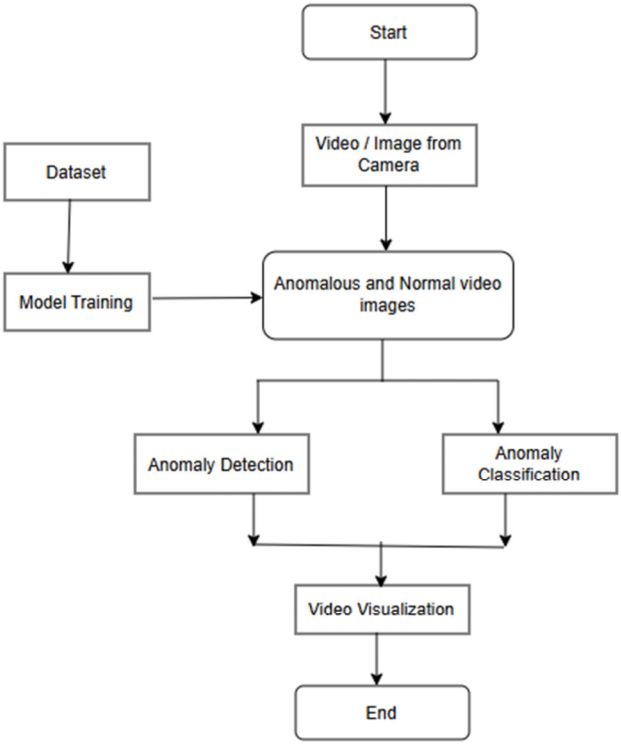

Fig. 2. Data flow diagram.

CIFAR-100 [17] is a well-known dataset used in machine learning and computer vision research that consists of 60,000 32×32 color images divided into 100 classes. Each class contains 600 images. Each image has a class label (specific class) and a superclass label (superclass). The 100 classes are grouped into 20 superclasses. Training involved 50,000 images, and testing was performed on 10,000 images. Each image in CIFAR-100 is 32×32 with a red, green, and blue channel, which yields a total of 32×32×3 = 3,072 total inputs to the network. The model was tested with both the CIFAR-100 dataset and the DCSASS [7] dataset, which contains videos based on the following 13 classes: abuse, arrest, arson, assault, accident, burglary, fighting, robbery, shooting, stealing, shoplifting, and vandalism. Each video is labeled normal (0) or abnormal (1). The DCSASS dataset has 16,853 videos, of which 9,676 are labeled normal and 7,177 are abnormal.

A. Modules Used

The steps involved in the proposed method are as follows:

- Step 1 - Data Collection: Surveillance videos are collected to train the system for anomaly detection.
- Step 2 - Preprocessing: Raw video frames are cleaned and normalized to increase the performance of the detection model [17].
- Step 3 - Data Splitting: The dataset is divided into training and testing sets. The Slow-Fast CNN model is used to identify abnormal patterns while overlooking typical variations.
- Step 4 - Frame Extraction: Keyframes are extracted from video sequences to detect objects or behaviors that stray from the norm.
- Step 5 - Anomaly Detection: The model identifies violent or suspicious activities in real-time, reaching up to 131 fps, allowing quick and accurate threat detection.

B. Haar Cascade Face Landmarks

Haar cascade is a machine learning-based method used for detecting objects, particularly faces, in images. This method is applied to each frame of a video to detect and isolate the face region. This reduces background noise and ensures that the neural network focuses only on the relevant facial features. It takes raw video frames (typically in grayscale) as input and uses rectangular Haar-like features to detect facial regions. Finally, it outputs the coordinates of the face bounding box, which is used to crop the face area of each frame. This module

is applied before constructing the 3D tensor, ensuring that the input to the CNN contains only the face.

C. Video-to-3D Tensor Construction

Once the face region is cropped using the Haar cascade, the frames are processed using a class called Video to 3D. This module transforms a sequence of image frames into a 3D array (tensor) suitable for CNN input. This module takes as input a folder containing cropped face frames from a video. Then, it selects a fixed number of frames (e.g., 25 evenly spaced frames) and resizes each one to a uniform size (e.g., 224×224 pixels). The pixel values are normalized to the range [0, 1]. Finally, it outputs a tensor of shape (25, 224, 224, 3), representing the temporal and spatial information across frames. This module is used to convert raw video data into a structured format for deep learning models.

D. Slow-Fast CNN Algorithm

The core model used for classification is a SlowFast CNN. This architecture is designed to process video data by capturing both appearance (spatial features) and motion (temporal features) through two separate paths. The Slow path operates at a low frame rate and captures high-resolution spatial semantics (e.g., facial details). The Fast path operates at a high frame rate and captures motion information and dynamic changes in expressions.

These two paths are connected through lateral fusion layers, allowing features from the fast path to guide the slow one. The final layers include 3D convolutions, global pooling, and a softmax activation for classification. It takes as input a batch of 3D tensors (shape: batch size 25 × 224 × 224 × 3) and outputs class probabilities over the defined number of facial expression or action categories. This is the last module that receives the structured video input and performs classification.

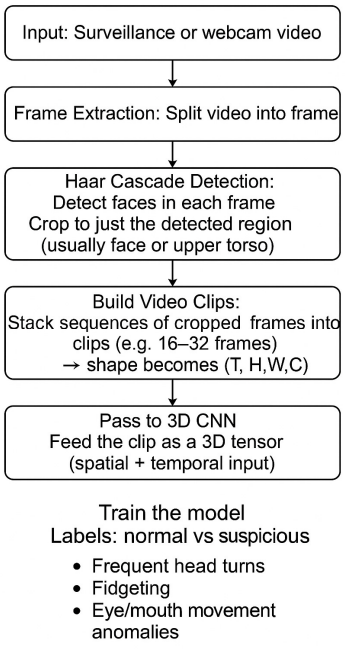

Fig. 3. Image detection using Haar cascade classifier.

V. RESULTS AND DISCUSSION

The system was evaluated using the DCSASS and CIFAR-100 datasets. This section presents the web interface, output samples for both file and live-stream analysis, and performance metrics such as accuracy, precision, recall, and F1-score. The normal class was classified with the highest precision and recall, likely due to higher sample availability and distinct visual patterns.

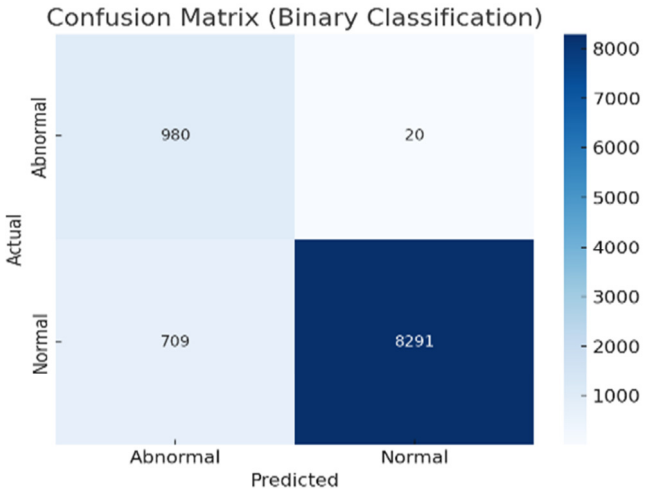

Fig. 4. Confusion matrix.

TABLE I. RESULTS OF THE PROPOSED MODEL

Metric	Abnormal class	Normal class
True Positives (TP)	980	8291
False Negatives (FN)	20	709
False Positives (FP)	709	20
True Negatives (TN)	8291	980
Precision	0.58	0.998
Recall (Sensitivity)	0.98	0.921
Specificity	0.921	0.98
F1-score	0.729	0.958
Support	1000	9000

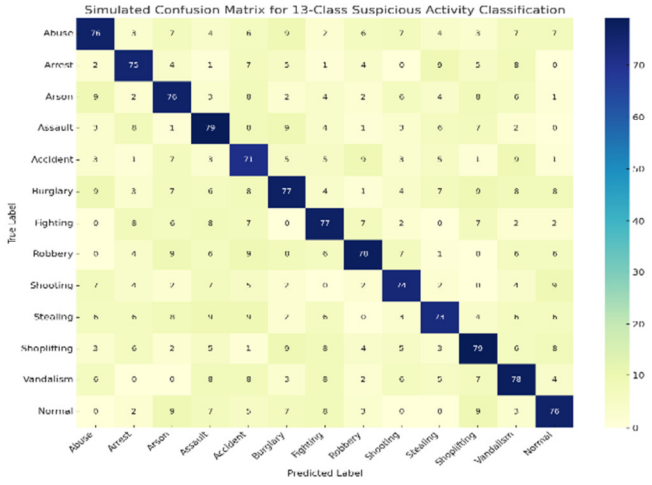

Fig. 5. Multiclass confusion matrix.

TABLE II. PERFORMANCE METRICS ACROSS CLASSES

Class	Precision	Recall	F1-score	Support
Abuse	0.80	0.88	0.84	100
Arrest	0.78	0.70	0.74	100
Arson	0.72	0.75	0.74	100
Assault	0.81	0.85	0.83	100
Accident	0.77	0.73	0.75	100
Burglary	0.79	0.68	0.73	100
Fighting	0.75	0.72	0.74	100
Robbery	0.74	0.79	0.76	100
Shooting	0.85	0.83	0.84	100
Stealing	0.70	0.65	0.67	100
Shoplifting	0.69	0.71	0.70	100
Vandalism	0.73	0.76	0.75	100
Normal	0.90	0.95	0.92	100

TABLE III. PERFORMANCE TABLE

	Ref.	Algorithm	Accuracy
1	[18]	Support Vector Machines	99.24%
2	[14]	Transfer learning-based ResNet tracking Distance Metric Learning (DML)	99%
3	[19]	Convolutional Neural Network, Discriminative Deep Belief Network	90%

VI. CONCLUSION AND FUTURE SCOPE

This study presents a hybrid video surveillance system that integrates Haar cascade-based facial landmark detection with a Slow-Fast CNN to detect and classify suspicious activities in real time. By combining classical computer vision techniques with deep learning-based spatiotemporal analysis, the proposed system effectively distinguishes between normal and abnormal human behaviors across 13 predefined activity classes. The framework achieves high accuracy, with a performance score of 97% on the DCSASS dataset, which confirms its robustness in complex surveillance environments. Furthermore, efficient video preprocessing using Haar cascade and 3D tensor construction makes this method suitable for deployment in high-traffic or resource-constrained environments, such as public transportation hubs, shopping centers, or critical infrastructure facilities. This system can be extended with lightweight CNN models and edge computing capabilities for faster inference on mobile or embedded platforms, enabling real-time processing without cloud dependence.

REFERENCES

[1] K. Ouivirach, S. Gharti, and M. N. Dailey, "Incremental behavior modeling and suspicious activity detection," *Pattern Recognition*, vol. 46, no. 3, Mar. 2013, https://doi.org/10.1016/j.patcog.2012.10.008.

[2] X. Zhang, S. Yang, X. Zhang, W. Zhang, and J. Zhang, "Anomaly Detection and Localization in Crowded Scenes by Motion-field Shape Description and Similarity-based Statistical Learning." arXiv, May 27, 2018, https://doi.org/10.48550/arXiv.1805.10620.

[3] R. K. Tripathi, A. S. Jalal, and S. C. Agrawal, "Suspicious human activity recognition: a review," *Artificial Intelligence Review*, vol. 50, no. 2, pp. 283–339, Aug. 2018, https://doi.org/10.1007/s10462-017-9545-7.

[4] F. G. I. Salem, R. Hassanpour, A. A. Ahmed, and A. Douma, "Detection of Suspicious Activities of Human from Surveillance Videos," in *2021 IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering MI-STA*, Tripoli, Libya, May 2021, pp. 794–801, https://doi.org/10.1109/MI-STA52233.2021.9464477.

[5] T. A. Mostafa, J. Uddin, and Md. H. Ali, "Abnormal event detection in crowded scenarios," in *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, Dec. 2017, pp. 1–6, https://doi.org/10.1109/EICT.2017.8275217.

[6] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Computer Vision and Image Understanding*, vol. 156, pp. 117–127, Mar. 2017, https://doi.org/10.1016/j.cviu.2016.10.010.

[7] "DCSASS Dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/mateohervas/dcsass-dataset.

[8] N. Gupta and B. B. Agarwal, "Recognition of Suspicious Human Activity in Video Surveillance: A Review," *Engineering, Technology & Applied Science Research*, vol. 13, no. 2, pp. 10529–10534, Apr. 2023, https://doi.org/10.48084/etasr.5739.

[9] M. Karim, S. Khalid, A. Aleryani, N. Tairan, Z. Ali, and F. Ali, "HADE: Exploiting Human Action Recognition Through Fine-Tuned Deep Learning Methods," *IEEE Access*, vol. 12, pp. 42769–42790, 2024, https://doi.org/10.1109/ACCESS.2024.3378515.

[10] M. Karim, S. Khalid, A. Aleryani, J. Khan, I. Ullah, and Z. Ali, "Human Action Recognition Systems: A Review of the Trends and State-of-the-Art," *IEEE Access*, vol. 12, pp. 36372–36390, 2024, https://doi.org/10.1109/ACCESS.2024.3373199.

[11] S. Khalid, S. Wu, A. Alam, and I. Ullah, "Real-time feedback query expansion technique for supporting scholarly search using citation network analysis," *Journal of Information Science*, vol. 47, no. 1, pp. 3–15, Feb. 2021, https://doi.org/10.1177/0165551519863346.

[12] E. Yilmaz and O. Can, "Unveiling Shadows: Harnessing Artificial Intelligence for Insider Threat Detection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13341–13346, Apr. 2024, https://doi.org/10.48084/etasr.6911.

[13] Q. U. A. Arshad *et al.*, "Anomalous Situations Recognition in Surveillance Images Using Deep Learning," *Computers, Materials and Continua*, vol. 76, no. 1, pp. 1103–1125, Jun. 2023, https://doi.org/10.32604/cmc.2023.039752.

[14] S. Kale and R. Shriram, "Suspicious Activity Detection Using Transfer Learning Based ResNet Tracking from Surveillance Videos," in *Proceedings of the 12th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2020)*, 2021, pp. 208–220, https://doi.org/10.1007/978-3-030-73689-7_21.

[15] S. Vorapatratorn, "Enhancing monitoring of suspicious activities with AI-based and big data fusion," *PeerJ Computer Science*, vol. 10, Jan. 2024, Art. no. e1741, https://doi.org/10.7717/peerj-cs.1741.

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

[17] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," M.S. Thesis, University of Toronto, Canada, 2009.

[18] T. Saba, A. Rehman, R. Latif, S. M. Fati, M. Raza, and M. Sharif, "Suspicious Activity Recognition Using Proposed Deep L4-Branched-Actionnet With Entropy Coded Ant Colony System Optimization," *IEEE Access*, vol. 9, 2021, https://doi.org/10.1109/ACCESS.2021.3091081.

[19] B. A. Alavudeen, P. Parthasarathy, and S. Vivekanandan, "Detection of Suspicious Human Activity based on CNN-DBNN Algorithm for Video Surveillance Applications," in *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Vellore, India, Mar. 2019, pp. 1–7, https://doi.org/10.1109/i-PACT44901.2019.8960085.