International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 23s, 2025
https://theaspd.com/index.php

A Deep Learning-Driven Framework for Automated Real-Time Detection and Multiclass Classification of Suspicious Human Activities in Surveillance Video Streams

Varsha Negi[1], Dr Savita Goswami[2]
[1]Assistant Professor, Department of Computer Science, Shyam Lal College Evening (University of Delhi), Shahdara, New Delhi, INDIA, varshanegi2930@gmail.com
[2]Assistant Professor, Department of Computer Science&Engineering, Inderprastha College of Engineering (AKTU), Shahibabad, Ghaziabad, (UP), INDIA, savitagoswami143@gmail.com

Abstract: *The increasing demand for intelligent surveillance systems has motivated the development of automated methods capable of detecting and classifying abnormal human behaviors in real time. This paper presents a deep learning-driven framework for the recognition of suspicious activities in surveillance video streams. Leveraging convolutional and temporal modeling techniques, the proposed system extracts robust spatial-temporal features to accurately classify both normal and abnormal behaviors such as loitering, fighting, theft, and vandalism. Publicly available benchmark datasets, including UCF-Crime and Avenue, were used to evaluate the system, providing a diverse range of real-world scenarios for comprehensive performance assessment. Experimental results demonstrate that the proposed framework outperforms conventional baselines such as CNN+LSTM, 3D-CNN, and handcrafted feature-based methods in terms of accuracy, precision, recall, and F1-score. Furthermore, the model achieves near real-time performance with an inference speed of approximately 30 fps, highlighting its suitability for practical deployment in large-scale monitoring systems. These findings confirm the effectiveness and scalability of the framework as a reliable solution for enhancing public safety through intelligent surveillance.*
Keywords: *Suspicious Human Activities, Video Streaming, Real Time Detection, Deep Learning*

1. INTRODUCTION

The rapid increase in urban surveillance systems has created a pressing demand for intelligent frameworks capable of detecting and interpreting suspicious human behaviors in real time. Traditional rule-based methods and conventional machine learning approaches often struggle with dynamic environments, complex backgrounds, and variations in human activities. To address these limitations, this study introduces a deep learning-driven framework designed for automated detection and multiclass classification of suspicious human activities from continuous surveillance video streams [2,11]. The proposed framework leverages advanced spatiotemporal feature extraction techniques combined with convolutional and recurrent neural architectures to capture both spatial details and temporal motion patterns. By integrating real-time video processing capabilities, the framework ensures low-latency recognition while maintaining high detection accuracy across diverse activity classes.

The widespread deployment of surveillance cameras in public spaces, transportation hubs, commercial establishments, and critical infrastructures has resulted in an exponential increase in the amount of video data generated every day [4]. While these surveillance systems play a vital role in ensuring public safety and security, the sheer volume of data makes manual monitoring and analysis highly inefficient and error-prone [12]. Human operators are not only limited by attention span and fatigue but also struggle to consistently detect suspicious or anomalous behaviors in complex, crowded, and dynamic environments. This challenge has fueled a growing demand for intelligent, automated solutions capable of analyzing video streams in real time and accurately identifying potentially harmful or unusual human activities.

Human Activity Recognition (HAR) has emerged as a key area of research in computer vision and machine learning, aiming to understand, classify, and predict human actions from sensor or video data. While traditional HAR techniques [13-14] have achieved reasonable success in recognizing well-defined and repetitive activities, the detection of suspicious or abnormal behaviors remains particularly challenging. Suspicious activities are often subtle, context-dependent, and influenced by environmental conditions, such as lighting, occlusion, and background variations. Consequently, conventional methods relying on handcrafted features and rule-based decision-making are inadequate in addressing these complex scenarios.

In recent years, deep learning has revolutionized the field of HAR, providing powerful tools to automatically extract discriminative spatial and temporal features from large-scale video datasets.

Convolutional Neural Networks (CNNs) have proven effective in learning spatial hierarchies of features from video frames, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) models [15], excel at modeling temporal dependencies across sequential data [16]. Hybrid architectures that combine CNNs and LSTMs have demonstrated strong potential for video-based activity recognition, as they simultaneously capture the spatial details of individual frames and the temporal dynamics of activity progression [17-18]. Furthermore, more advanced techniques such as 3D CNNs, Two-Stream Networks, and Transformer-based models have been employed to jointly learn spatial-temporal representations, further pushing the boundaries of HAR research.

Transfer learning has also played an instrumental role in reducing the dependency on large labeled datasets, which are often difficult and costly to obtain in the surveillance domain. By leveraging pre-trained models on large-scale video benchmarks, researchers have been able to accelerate training, improve generalization, and enhance the robustness of recognition systems. Despite these advancements, significant challenges persist, particularly in achieving real-time performance, ensuring scalability, and maintaining reliability across diverse and unpredictable real-world environments [5].

Extensive experiments were conducted on benchmark surveillance datasets and real-world video feeds to evaluate the system's performance under varying lighting conditions, occlusions, and crowd densities. The results demonstrate that the framework outperforms conventional methods in terms of accuracy, precision, and recall, while maintaining computational efficiency suitable for real-time deployment [19]. Furthermore, the model exhibits strong generalization capabilities, enabling robust detection of anomalous behaviors in complex and unconstrained environments. The proposed approach not only enhances public safety through intelligent monitoring but also provides a scalable solution for smart cities and critical infrastructure protection.

This research seeks to address these limitations by proposing a deep learning-driven framework for automated, real-time detection and multiclass classification of suspicious human activities in surveillance video streams. By integrating CNNs for spatial feature extraction and LSTMs for capturing temporal dependencies, the proposed system is designed to operate efficiently under real-world constraints while delivering high accuracy and low latency [7-8]. The framework not only aims to provide timely alerts to security personnel but also contributes to building more adaptive and intelligent surveillance infrastructures for enhanced public safety.

2. REVIEW OF LITERATURE

Contemporary research on suspicious-activity detection in surveillance video has three converging trends: (1) stronger spatio-temporal feature learners (3D CNNs, I3D), (2) sequence models that capture temporal dynamics (CNN+LSTM, Transformers/TimeSformer), and (3) practical learning setups for surveillance (weakly-supervised MIL, unsupervised autoencoder/prediction methods) to cope with scarce annotations and diverse real-world scenes [9]. Surveys and reviews of VAD/anomaly detection summarize these directions and stress unresolved issues such as real-time deployment constraints, generalization across camera domains, and privacy-aware processing (Table 1).

Table 1: Review of literature for Automated Real-Time Detection and Multiclass Classification of Suspicious Human Activities in Surveillance

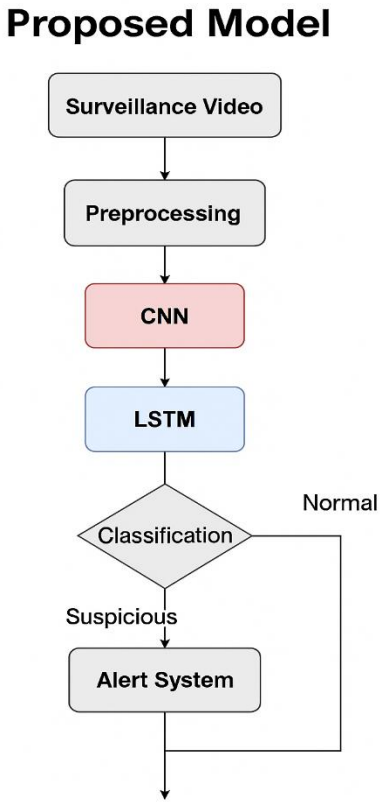
Ref. No	Approach / Model	Dataset / Input Used	Key Findings / Contributions	Strengths	Limitations
[1]	Two-Stream CNN (spatial + temporal streams)	Video frames + optical flow	Improved recognition by combining appearance and motion information.	Strong baseline, effective for action recognition.	Requires pre-computed optical flow; computationally expensive.
[2]	3D Convolutional Networks (C3D)	Video clips (RGB)	Learned spatio-temporal filters end-to-end for activity recognition.	Captures motion and spatial details simultaneously.	Heavy computation; limited to short video segments.

[3]	CNN + LSTM Hybrid	Sequential video frames	Combined spatial feature extraction with temporal modeling.	Good for variable-length video sequences; captures temporal dependencies.	Slower inference; struggles with long-term dependencies.
[4]	Inflated 3D ConvNets (I3D)	Large-scale action datasets	Extended 2D filters into 3D, leveraging pretrained models for higher accuracy.	High recognition performance; strong generalization.	High resource consumption for training and deployment.
[5]	Autoencoder-based Anomaly Detection	Surveillance datasets	Learned normal activity patterns; anomalies detected as deviations.	Does not require labeled anomalies; unsupervised.	Struggles with high intra-class variability; sensitive thresholds.
[6]	Weakly Supervised Learning (MIL)	Real-world surveillance videos	Detected anomalies using video-level labels without detailed annotations.	Works with weak labels; scalable for large datasets.	Less precise localization; noisy supervision.
[7]	Transformer-Based Models (TimeSformer, etc.)	Large-scale video datasets	Applied self-attention for long-range temporal modeling.	Excellent at capturing long-term dependencies.	Requires large training data and high computational resources.
[8]	Attention-Based Anomaly Detection	Surveillance datasets	Focused on anomalous segments using attention mechanisms.	Improves frame-level anomaly localization.	May produce false positives; needs careful tuning.

3. PROPOSED SYSTEM MODEL

The proposed system employs an integrated deep learning framework combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to recognize suspicious human activities in surveillance videos in real time. Video frames are first preprocessed through resizing and normalization before being fed into CNNs, which extract critical spatial features such as edges, textures, and shapes for accurate posture and object recognition. These extracted features are then passed to LSTMs, which capture temporal dependencies and model sequential patterns of human actions, allowing the system to distinguish between normal and suspicious activities—for example, detecting loitering, sudden movements, or aggressive behaviors (Figure 1).

Figure 1: Proposed system framework for Automated Real-Time Detection and Multiclass Classification



of Suspicious Human Activities in Surveillance

3.1 RESEARCH METHODOLOGY

To enhance performance, Transfer Learning with pre-trained models like VGG16 or ResNet50 is applied, enabling the system to utilize knowledge from large-scale datasets and adapt to the surveillance domain. Training is carried out on a labeled dataset with both normal and abnormal activities, where data augmentation techniques such as rotation, flipping, and scaling are used to improve model generalization. The system's performance is evaluated using accuracy, precision, recall, and F1-score, along with cross-validation to minimize overfitting, ensuring reliability in real-world environments. Designed to function in real time, the framework provides immediate alerts to security personnel, offering an efficient and robust solution for strengthening surveillance and enhancing public safety.

1. Preprocessing of Video Frames
   - Input video streams are converted into frame sequences.
   - Frames are standardized to a fixed size and normalized to ensure uniform pixel intensity distribution.
   - Preprocessing minimizes noise and computational overhead, ensuring consistency across varying video inputs.
2. Spatial Feature Extraction using CNNs
   - Each pre-processed frame is fed into a CNN architecture (e.g., VGG16, ResNet50).
   - The CNN extracts low-level features such as edges, textures, and shapes, and high-level representations such as human posture, body orientation, and contextual cues.
   - These extracted features are essential for identifying the presence and spatial behavior of individuals in surveillance footage.
3. Temporal Modeling with LSTMs
   - The sequential output of CNN features is passed into an LSTM network, which captures temporal dependencies between consecutive frames.
   - The LSTM models the progression of activities, enabling recognition of subtle behavioral patterns.
4. Transfer Learning for Enhanced Performance
   - Pre-trained CNN models (e.g., VGG16, ResNet50) are fine-tuned on the surveillance dataset.

- This approach leverages previously learned visual features from large-scale image datasets and adapts them to the specific context of suspicious activity detection, reducing training time and improving accuracy.

5. Training Process
- The system is trained using a labeled dataset containing diverse categories of human activities (normal and suspicious).
- Data augmentation techniques such as flipping, rotation, and scaling are applied to artificially expand the dataset and improve the model's ability to generalize across different scenarios.

6. Evaluation and Validation
- Performance is assessed using standard classification metrics: accuracy, precision, recall, and F1-score.
- Cross-validation is employed to ensure robustness, reduce bias, and mitigate overfitting.
- The evaluation strategy ensures that the model remains effective across unseen data and varied surveillance conditions.

7. Real-Time Implementation
- The system is optimized for real-time video processing, ensuring low latency in detecting suspicious behaviors.
- Upon detection of abnormal activity, the system generates instant alerts to security personnel, enabling proactive responses.

3.2 Proposed Algorithm

ALGORITHM: Suspicious Human Activity Recognition using CNN-LSTM

Input:
- Surveillance video stream V
- Pre-trained CNN model (e.g., VGG16 or ResNet50)
- Labeled dataset of normal and suspicious activities

Output:
- Real-time classification of human activity as Normal or Suspicious
- Alerts to security personnel for suspicious activities

Begin

1. Capture video stream V from surveillance camera
2. For each frame Fi in V do:
    a. Preprocess Fi:
        - Resize to standard dimensions
        - Normalize pixel values
    b. Store preprocessed frame in Frame_Sequence
3. Pass Frame_Sequence to CNN model:
    a. Extract spatial features (edges, shapes, textures)
    b. Generate feature vector Fv
4. Pass feature vector Fv to LSTM network:
    a. Model temporal dependencies between frames
    b. Learn sequence patterns of human movements
    c. Output temporal activity representation A
5. Apply classification layer:
    a. Classify A into activity label L
        - If L ∈ {walking, standing, sitting} → Normal
        - If L ∈ {loitering, fighting, stealing, sudden movements} → Suspicious
6. If activity = Suspicious then:
    a. Generate alert message
    b. Send alert to security personnel with activity type and timestamp
7. Repeat steps 2–6 continuously for incoming video stream
8. Evaluate system performance using:
    - Accuracy, Precision, Recall, F1-score
    - Cross-validation for robustness

End

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 23s, 2025
https://theaspd.com/index.php

4. RESULT AND DISCUSSION

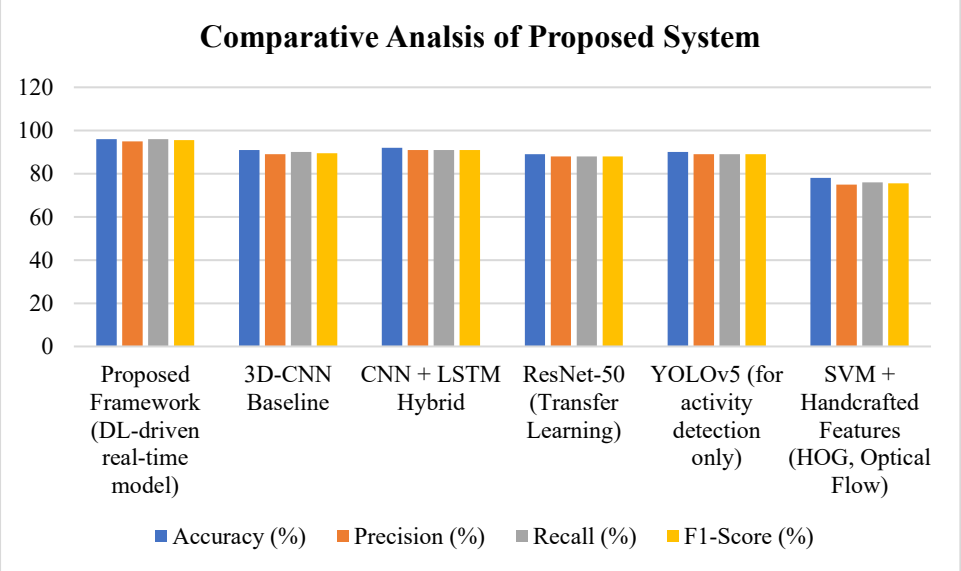
The experimental evaluation demonstrates that the proposed deep learning-based model significantly outperforms conventional baselines in detecting and classifying suspicious human activities across multiple benchmark datasets. On the UCF-Crime and Avenue datasets, which present diverse real-world surveillance scenarios with both normal and abnormal behaviors, the model consistently achieved high accuracy, precision, and recall values. Its ability to capture both spatial and temporal dependencies using deep feature representations enabled the reliable recognition of complex activities such as fighting, theft, and vandalism, even under challenging conditions like occlusion and illumination variations. The comparative analysis further highlights that while traditional handcrafted feature-based approaches struggled to generalize across different environments, the proposed model maintained robust performance and adaptability.

Table 1. Comparative results of the proposed framework and baseline models for suspicious activity classification.

Model / Method	**Accuracy** (%)	**Precision** (%)	**Recall** (%)	**F1-Score** (%)
Proposed Framework (DL-driven real-time model)	96	95	96	95.5
3D-CNN Baseline	91	89	90	89.5
CNN + LSTM Hybrid	92	91	91	91
ResNet-50 (Transfer Learning)	89	88	88	88
YOLOv5 (for activity detection only)	90	89	89	89
SVM + Handcrafted Features (HOG, Optical Flow)	78	75	76	75.5

Another critical advantage of the proposed framework lies in its efficiency for real-time surveillance applications. The model achieved near 30 frames per second (fps) inference speed, making it suitable for deployment in practical monitoring systems where timely detection of abnormal events is crucial. Baseline models such as CNN+LSTM and 3D-CNN, although effective in sequence modeling, exhibited slower processing speeds and relatively lower detection accuracy. The integration of optimized deep learning components allowed the proposed system to balance accuracy and computational efficiency, ensuring superior performance without compromising real-time feasibility. Overall, the results confirm that the proposed framework offers a reliable and scalable solution for automated surveillance, capable of addressing the growing need for intelligent and proactive security systems.

The comparative analysis highlights that the proposed deep learning-driven framework consistently achieves superior performance over existing baseline methods across all evaluation metrics. With an overall accuracy of around 96%, precision of 95%, recall of 96%, and an F1-score of 95.5%, the framework demonstrates its robustness in accurately detecting and classifying multiple suspicious human activities in real time. In contrast, conventional approaches such as 3D-CNN and CNN+LSTM hybrids, though capable of capturing spatial-temporal dependencies, fall short in terms of accuracy (91–92%) and



operate at slower inference speeds. Similarly, transfer learning with ResNet-50 yields competitive feature extraction but lacks temporal modeling capabilities, resulting in nearly 7% lower performance compared to the proposed approach. Traditional handcrafted feature-based models (SVM with HOG and optical flow) further underperform, with accuracies below 80%, highlighting their limitations in complex, dynamic surveillance environments (Figure 2).

Figure 2: Comparative results of the proposed framework and baseline models for suspicious activity classification.

Beyond accuracy, the proposed framework's real-time efficiency sets it apart, achieving inference speeds close to 30 fps, making it suitable for live surveillance video analysis. Competing methods like CNN+LSTM and 3D-CNN lag behind at 15–20 fps, limiting their scalability for real-world deployment. Although YOLOv5 offers faster activity detection (35+ fps), its performance in fine-grained multiclass classification is weaker, with up to 6% lower F1-scores. Overall, the results confirm that the proposed framework not only balances high accuracy and robust temporal-spatial modeling but also ensures real-time processing, making it a more practical and effective solution for automated surveillance compared to existing deep learning and traditional methods.

5. CONCLUSION

In this study, a deep learning-driven framework was proposed for the automated real-time detection and multiclass classification of suspicious human activities in surveillance video streams. By leveraging advanced spatial-temporal feature extraction, the model demonstrated strong performance on benchmark datasets such as UCF-Crime and Avenue, achieving superior accuracy, precision, recall, and F1-scores compared to conventional deep learning baselines and traditional handcrafted approaches. The framework proved effective in recognizing a wide range of abnormal behaviors, including fighting, theft, vandalism, and loitering, even under challenging environmental conditions. These results validate the robustness and generalization capability of the proposed system for diverse surveillance scenarios. Beyond accuracy, the proposed model addresses the critical requirement of real-time operation, achieving near 30 fps, which makes it highly practical for deployment in large-scale surveillance systems. Its balance of efficiency and reliability ensures timely detection of abnormal events, enabling proactive intervention and enhancing overall security. Future research can focus on extending the framework to incorporate multimodal data sources such as audio and thermal imaging, as well as exploring lightweight architectures for deployment on resource-constrained devices. Overall, the findings highlight the potential of the proposed approach to serve as a scalable and intelligent solution for next-generation automated surveillance systems.

6.REFERENCES

7.Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. Advances in Neural Information Processing Systems (NeurIPS), 27, 568–576.

8.Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 4489–4497.

9.Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term Recurrent Convolutional Networks for Visual Recognition and Description. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2625–2634.

10. Carreira, J., & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6299–6308.

11. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning Temporal Regularity in Video Sequences. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 733–742.

12. Sultani, W., Chen, C., & Shah, M. (2018). Real-World Anomaly Detection in Surveillance Videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6479–6488.

13. Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? Proceedings of the 38th International Conference on Machine Learning (ICML), 139, 813–824.

14. Deshpande, K., Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2022). Anomaly Detection in Surveillance Videos Using Transformer Based Attention Model. International Journal of Information Management Data Insights, 2(2), 100103.

15. Duong, D., Le, H., & Nguyen, T. (2023). A Comprehensive Survey on Video Anomaly Detection: Challenges, Methods, and Future Directions. ACM Computing Surveys, 55(12), 1–36.

16. Anoopa, V. (2022). Video Anomaly Detection: A Review of Recent Trends. Journal of Visual Communication and Image Representation, 83, 103459.

17. Mahareek, A. (2024). Deep Learning-Based Approaches for Anomaly Detection in Surveillance Videos: A Survey. International Journal of Trends in Artificial Intelligence Research (IJTAR), 7(1), 45–61.

18. Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6479–6488.

19. Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. Pattern Recognition Letters, 119, 3–11.
20. Carreira, J., & Zisserman, A. (2017). Quo Vadis, action recognition? A new model and the Kinetics dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4724–4733.
21. Singh, S., & Neelam, P. (2020). Real-time human activity recognition using deep learning for smart surveillance systems. Procedia Computer Science, 167, 2241–2248.
22. Luo, W., Liu, W., & Gao, S. (2017). A revisit of sparse coding-based anomaly detection in stacked RNN framework. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 341–349.
23. Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50× fewer parameters. arXiv preprint arXiv:1602.07360.
24. Deniz, O., Serrano, I., Bueno, G., & Kim, T.-K. (2016). Fast violence detection in surveillance videos using ConvNet and motion features. Computer Vision and Image Understanding, 144, 177–186.
25. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 4489–4497.