ijamrsd.com
ISSN: 2583-6404
Journal of Advanced Multidisciplinary Research Studies and Development
Peer Reviewed Referred Journal | Bi-monthly | Open Access

A REAL-TIME SPATIO-TEMPORAL CRIMINAL ACTIVITY DETECTION SYSTEM USING DEEP LEARNING

Prajwal Temak[1], Vicky Raut[2], Smita Kadam[3]
[1,2,3] Department of Computer Science, Savitribai Phule Pune University.
Email: Ptemak9763@gmail.com[1], vraut4143@gmail.com[2], smitakadam2004@gmail.com[3]

Abstract
The significant increase in surveillance cameras has made continuous manual monitoring inefficient and impractical. Therefore, to meet this challenge, the following has been proposed: a real-time spatio-temporal criminal activity detection system using deep learning within intelligent video surveillance. A combination of CNN for the spatial feature extraction process in single frames and LSTM for the model of temporal relationships between consecutive frames was considered. Compared to frame-based methods, this method is much better at identifying dynamically developed criminal activities such as assault, burglary, and fighting because it analyzes visual content and motion patterns over time. Besides the learning model, a system-integrated architecture provides support for real-time inference, automated event logging, and selective clip extraction, all of which lead to a massive reduction in storage requirements and manual inspection effort. From the results of the experiments, stable training was indicated with approximately 90% validation accuracy, which reflected good generalization. This provides a meaningful and scalable application of spatio-temporal deep learning for modern video surveillance.

Keywords: Crime Detection, Spatio-Temporal Learning, Deep Learning, Video Surveillance, CNN, LSTM.

► *Corresponding Author: Prajwal Temak*

1. Introduction

The widespread adoption of surveillance cameras in public spaces, transportation systems, commercial establishments, and residential areas has significantly increased the volume of recorded video data. While these systems are intended to enhance safety and security, continuous manual monitoring of video streams is inefficient and prone to human fatigue, delayed response, and oversight. As urban environments become more complex, there is an increasing need for intelligent systems capable of automatically detecting suspicious or criminal activities in real time. Despite advances in computer vision, accurate crime detection in surveillance videos remains a challenging problem. Many traditional approaches rely on frame-based image classification or simple motion detection techniques, which are insufficient for recognizing complex human behaviors. Criminal activities such as assault, burglary, or fighting occur as sequences of actions over time rather than isolated visual events. Therefore, systems that analyze only spatial information within individual frames often fail to capture the temporal dynamics necessary for reliable activity recognition.

The primary aim of this research is to develop a real-time spatio-temporal criminal activity detection system using deep learning. The proposed approach integrates spatial feature extraction

with temporal sequence modeling to enable robust recognition of dynamic behaviors in video streams. By combining deep learning-based action recognition with a real-time processing pipeline, the system seeks to provide automated event detection, efficient surveillance monitoring, and intelligent clip extraction.

The rationale behind this work lies in bridging the gap between theoretical action recognition models and practical surveillance deployment. While numerous studies focus on improving classification accuracy in controlled datasets, fewer works emphasize real-time applicability and system-level integration. This research contributes by presenting a deployable and scalable solution that enhances surveillance efficiency while maintaining strong generalization performance.

2. Literature Review

Recent advancements in artificial intelligence and deep learning have significantly improved automated surveillance systems for crime detection and anomaly recognition. Several researchers have explored different machine learning and deep learning techniques to analyze crime patterns, detect abnormal activities, and enhance public safety.

Manal Mostafa Ali (2022) proposed a real-time video anomaly detection system for smart surveillance environments [1]. The study combines a spatial feature extractor with a temporal autoencoder to capture temporal patterns in video frames. Background Subtraction (BS) was used to focus on Regions of Interest (ROI), while YOLOv5 was utilized for fast and accurate object detection in dynamic scenes. The model was evaluated using multiple anomaly detection datasets, including the UCSD anomaly dataset, Campus dataset, and CUHK abnormal event dataset. The proposed approach achieved an accuracy of 97.5%, precision of 98.0%, recall of 95.2%, and an F1-score of 96.57%, demonstrating its effectiveness in real-time anomaly detection.

Varun Mandalapu and Lavanya Elluri (2023) presented a systematic review on crime prediction using machine learning and deep learning techniques[2]. Their study explored various classification algorithms such as Support Vector Machines (SVM) and Naive Bayes, along with regression and time-series forecasting models. The researchers analyzed crime datasets from government sources such as NCRB and data.gov.in to identify crime trends and patterns. Their findings indicated that machine learning models can achieve high prediction accuracy, with reported results showing an accuracy of 97.5%, RMSE of 0.0023, MAE of 0.0017, and an R^2 value close to 0.9.

Elmetwally, Eldeeb, and Elmougy (2024) proposed an optimized machine learning and big data approach for crime detection[3]. Their framework integrates deep learning models such as VGGNet-19 for violent frame detection along with motion vector extraction and foreground segmentation techniques. The study also incorporates SoftMax regression, CNN, SVM, and ANN models in a hybrid architecture. The approach was tested using crime datasets from Los Angeles and San Francisco containing attributes such as crime type, date, location, and time. The model achieved an accuracy of 97.5%, precision of 98%, recall of 94%, and an F1-score of 94.7%.

Shanthi P and Manjula V (2025) focused on weapon detection using deep learning models such as YOLOv8 and Faster R-CNN to improve crime prevention and security in surveillance systems[4]. Their research highlighted the importance of real-time detection and computational efficiency in CCTV-based monitoring systems. The proposed model was trained using a gun detection dataset from Kaggle and demonstrated strong performance with an accuracy of 98.7%, precision of 97.2%, recall of 95%, and an F1-score of 96.7%.

Eugenio Cesario et al. (2024) introduced the Multi-Density Crime Predictor (MD-Crime Predictor), which uses clustering techniques to detect crime hotspots in urban areas[5]. Their approach integrates forecasting models such as SARIMA and Long Short-Term Memory (LSTM) networks to predict future crime occurrences. The model was evaluated using crime datasets from Indian government portals and NCRB. The proposed system achieved an accuracy of 94.58%, precision of 97.4%, recall of 93.8%, and an RMSE of 0.0056, demonstrating its ability to predict crime patterns effectively.

Gupta and Mehta (2024) proposed a hybrid CNN–LSTM framework for spatio-temporal crime forecasting[6]. The model combines convolutional neural networks for spatial feature extraction with LSTM networks for temporal sequence modeling. The system was evaluated using the Chicago Crime Dataset and New York City Open Crime Data. Their results showed an accuracy of 96.8%, precision of 95.4%, recall of 94.8%, and an RMSE of 0.0042, highlighting the effectiveness of hybrid deep learning models in crime prediction tasks.

Kumar and Sinha (2023) introduced a Graph Neural Network (GNN) based approach for modeling spatial relationships between crime hotspots[7]. By capturing spatial dependencies between different geographical regions, their model improved the prediction of crime-prone areas. The system was tested using San Francisco and Los Angeles crime datasets and achieved an F1-score of 0.94, Mean Absolute Error (MAE) of 0.005, and clustering purity of 0.91.

Wang and Zhao (2022) proposed a deep learning framework based on 3D Convolutional Neural Networks (3D-CNN) with an attention mechanism for real-time violent activity detection in CCTV footage[8]. The model captures both spatial and temporal features of video sequences, enabling accurate recognition of violent events. The system was evaluated on benchmark datasets such as UCF-Crime, Avenue, and ShanghaiTech, achieving an AUC of 97.2%, precision of 96%, and recall of 95.8%.

Overall, the existing literature demonstrates that deep learning techniques such as CNNs, LSTMs, 3D-CNNs, and hybrid architectures have significantly improved the accuracy and efficiency of crime detection and prediction systems. However, many existing approaches focus either on crime prediction using structured datasets or anomaly detection in surveillance videos. There remains a need for integrated systems capable of performing real-time crime detection in CCTV streams while analyzing spatio-temporal patterns of human behavior. The proposed system addresses this gap by leveraging deep learning techniques for automated detection of suspicious activities in real-time surveillance environments.

3. Proposed Methodology

3.1 System Architecture Overview

The proposed system is designed to detect criminal activities in real-time surveillance video streams using a spatio-temporal deep learning framework. The architecture integrates video acquisition, sequential data processing, deep feature extraction, temporal modeling, and intelligent event prediction into a unified pipeline. The primary objective of the system is to identify suspicious behavioral patterns rather than relying solely on static frame-level object detection.

The system operates in four major stages. First, continuous surveillance video streams are captured and segmented into sequential frames. The extracted frames are resized to 224 × 224 pixels and organized into fixed-length sequences in order to preserve motion continuity between consecutive frames. Second, a Convolutional Neural Network (CNN) is employed to extract high-level spatial features from each frame, capturing appearance-based information such as human posture, scene context, and object presence.

In the third stage, the extracted spatial features are passed to a Long Short-Term Memory (LSTM) network, which models temporal dependencies across sequential frames. This temporal modeling allows the system to learn evolving activity patterns and distinguish between normal and suspicious behaviors in surveillance footage.

Finally, a fully connected classification layer produces probability scores for the predefined activity categories, enabling the system to classify input video sequences into one of the ten activity classes: Abuse, Arson, Assault, Burglary, Explosion, Fighting, Normal, Road Accidents, Robbery, and Shooting.

To ensure real-world applicability, the architecture is integrated with a backend API framework that enables real-time inference and event logging. When abnormal activity is detected, the system can trigger automated alerts and extract relevant video clips, supporting efficient monitoring and rapid response in intelligent surveillance environments.

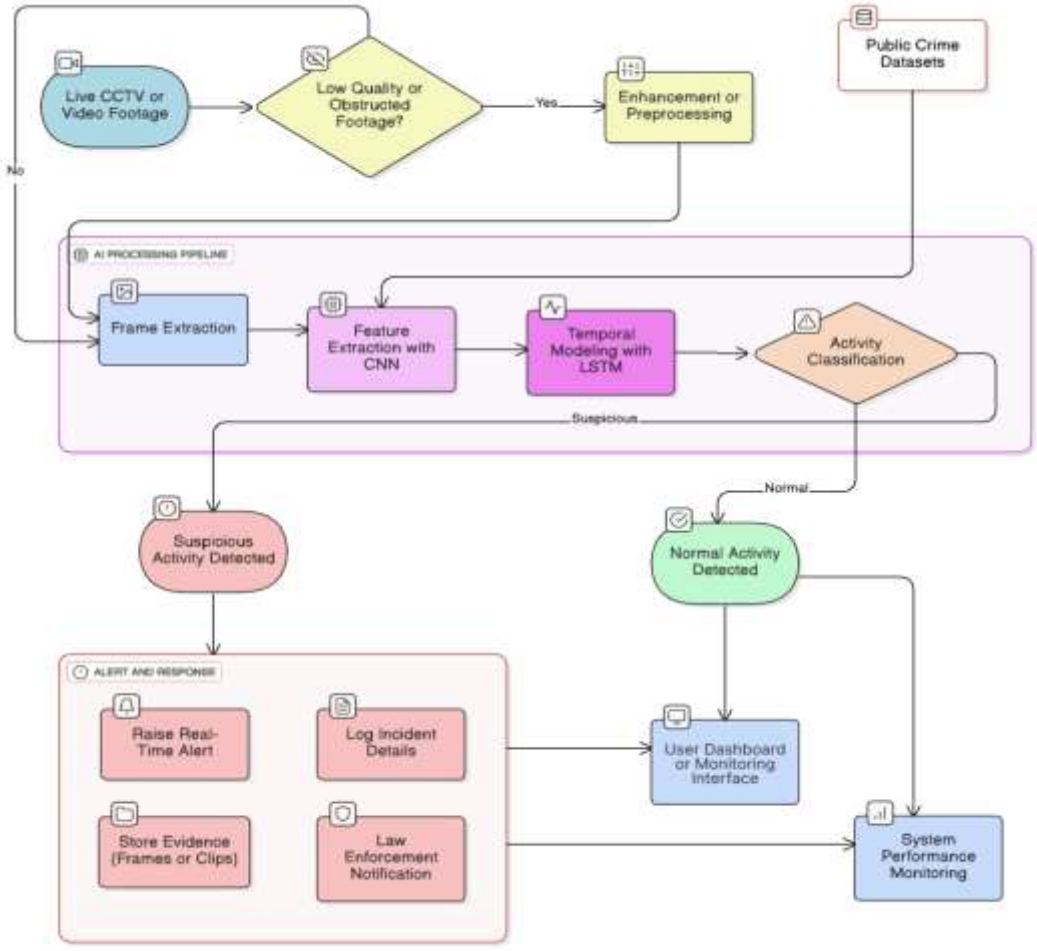

Figure 1. System Architecture

3.2 Dataset Seletion

UCF-Crime:

For the development and evaluation of the proposed crime detection system, the UCF-Crime Dataset was utilized. This dataset is widely adopted in research for real-world anomaly detection in surveillance videos, as it contains long, untrimmed videos collected from CCTV cameras that simulate realistic monitoring environments.

The dataset was obtained from Kaggle and contains 1900 surveillance videos with a total duration of approximately 128 hours, occupying nearly 82.9 GB of storage. Unlike trimmed action-recognition datasets, these videos contain long durations of normal activity before and after anomalous events, making them highly suitable for evaluating real-world intelligent surveillance systems.

The original dataset includes 13 anomaly categories, representing different criminal or suspicious activities such as Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism, along with a large number of normal activity videos.

However, for the purpose of this study, ten activity classes were selected to train and evaluate the proposed model. These classes include Abuse, Arson, Assault, Burglary, Explosion, Fighting, Normal, Road Accidents, Robbery, and Shooting.

Class Distribution

The dataset contains an imbalanced distribution of videos across different anomaly categories. As illustrated in Figure 2, the Normal activity class contains the highest number of videos, while several anomaly classes have fewer samples. The approximate distribution used in this study includes:

- Burglary – 100 videos
- Fighting – 50 videos
- Road Accidents – 150 videos
- Robbery – 150 videos
- Shooting – 50 videos
- Shoplifting – 50 videos
- Stealing – 100 videos
- Abuse – 50 videos
- Arrest – 50 videos
- Arson – 50 videos
- Assault – 50 videos
- Explosion – 50 videos
- Vandalism – 50 videos
- Normal – 950 videos

This imbalance reflects real-world surveillance scenarios where abnormal events occur significantly less frequently than normal activities.

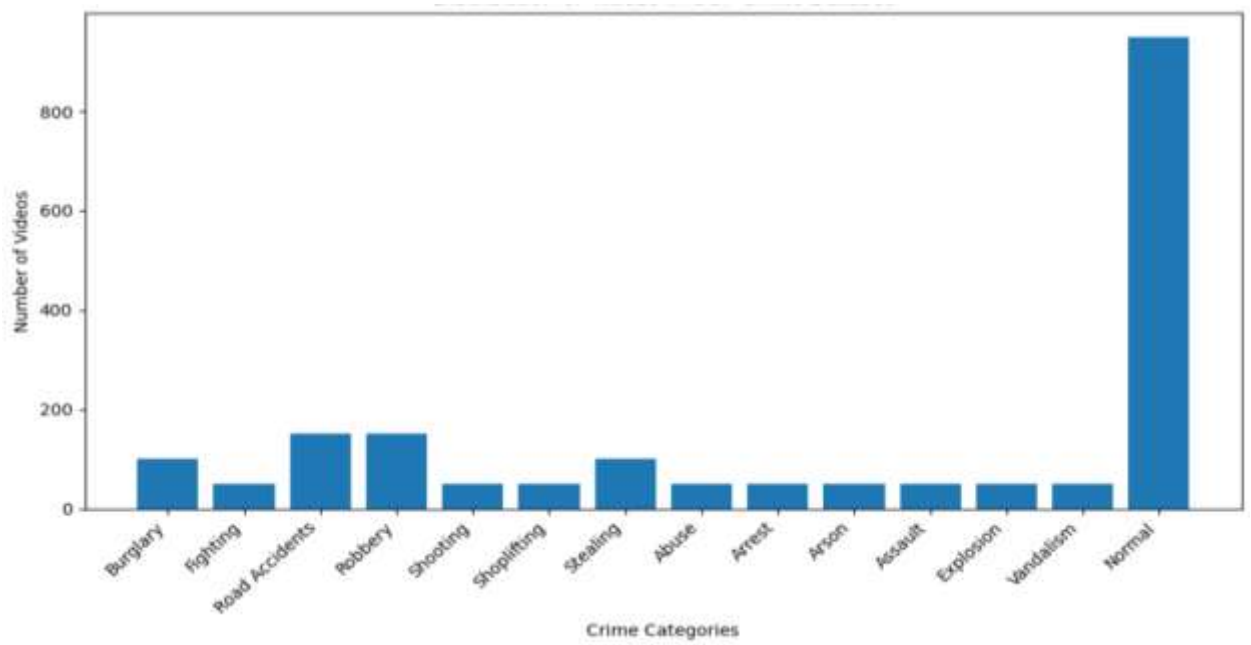

Figure 2. Distribution of Videos in UCF Crime Dataset

3.3 Data Preprocessing and Sequence Formation

The proposed system is trained and evaluated using the UCF-Crime dataset, which consists of real-world surveillance videos representing various criminal and normal activities. Each video is treated as a sequence of ordered frames containing both spatial and temporal information.

All video frames are extracted and resized to a fixed resolution of 244 × 244 pixels to ensure uniformity and computational efficiency. To preserve motion continuity, the video stream is segmented into fixed-length sequences of 16 consecutive frames. Each sequence represents a temporal clip that captures short-term activity evolution, enabling the model to learn dynamic behavioral patterns rather than isolated frame-level features.

The dataset is divided into training and validation sets using an 80/20 split to evaluate generalization performance. During preprocessing, frame normalization and sequence structuring are applied to convert raw video input into model-ready tensors. This structured representation ensures compatibility with the hybrid deep learning architecture used for spatio-temporal learning.

3.4 Spatial Feature Extraction Using CNN

To capture spatial information from individual video frames, a Convolutional Neural Network (CNN) is employed as the feature encoder in the proposed architecture. Each frame within a 16-frame sequence is independently passed through the CNN to extract high-level spatial representations. These representations encode important visual cues such as human posture, object presence, scene context, and interaction patterns that are essential for identifying suspicious activities.

Rather than performing direct frame-level classification, the CNN transforms each input frame of size 224× 224 into a compact feature vector. This encoded feature embedding reduces dimensional complexity while preserving discriminative spatial characteristics. The use of convolutional layers enables hierarchical feature learning, allowing the model to detect low-level patterns such as edges and textures in earlier layers and more abstract semantic features in deeper layers.

By acting as an encoder, the CNN converts raw visual data into structured feature sequences, which are then forwarded to the temporal modeling component. This separation of spatial encoding and

temporal reasoning improves learning efficiency and enhances the model's ability to recognize complex criminal behaviors in dynamic surveillance environments.

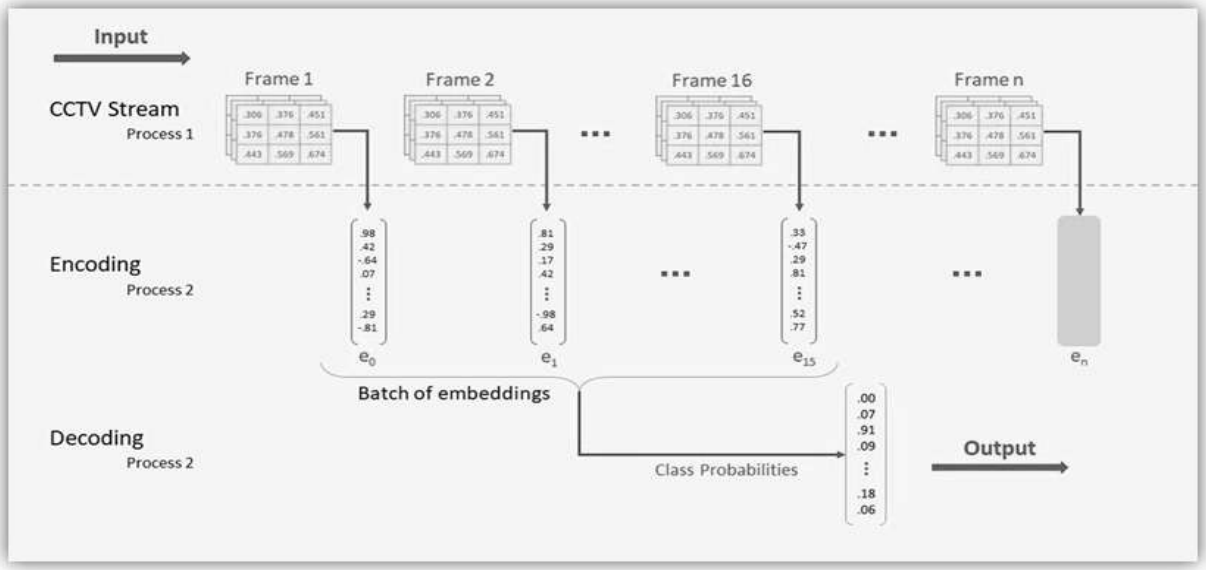

Figure 3. Feature Extraction

3.5 Temporal Modeling Using LSTM

While the CNN extracts spatial features from individual frames, recognizing criminal activities also requires understanding how actions evolve over time. To capture these temporal patterns, an LSTM network is used as the decoder. The CNN generates feature vectors from 16 consecutive frames, which are then passed to the LSTM.

Because LSTMs retain information across time steps, they recognize sequential behaviors such as fights, sudden movements, or unusual crowd activity. By learning relationships between frames, the LSTM produces a final representation that summarizes the overall behavior, which is then classified into activity categories.

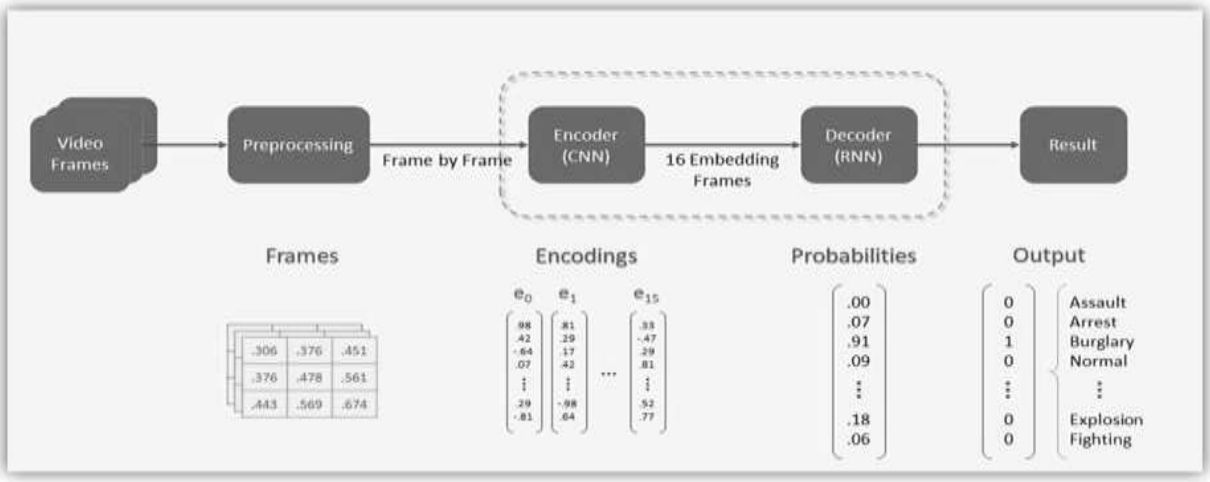

Figure 4. Video Processing Pipeline

3.6 Classification Strategy

The final activity prediction is generated by a fully connected classification layer that receives the last hidden output from the LSTM. This output acts as a compact spatio-temporal descriptor containing both spatial features from the CNN and temporal motion patterns learned across frames. The fully connected layer maps this representation to the predefined activity classes, and a Softmax function converts the output into probability scores. The class with the highest probability is selected as the predicted activity.

During training, categorical cross-entropy loss is used to measure the difference between predicted probabilities and the true labels, ensuring stable learning for multi-class classification. The probabilistic output also provides confidence levels for each prediction, making it useful for real-time surveillance systems where thresholds can trigger alerts for suspicious activities.

3.7 Real-Time Deployment and Video Processing Pipeline

The proposed system is designed for real-world surveillance, where continuous CCTV streams generate large amounts of data. To achieve low-latency performance, we use a multiprocessing pipeline that separates frame capture and model inference into two parallel processes. This ensures that video capture never blocks prediction, allowing the system to process frames smoothly and deliver near real-time activity classification.

Captured frames are grouped into fixed-length sequences and analyzed by the deep learning model. When a suspicious event is detected, the system immediately triggers alerts and saves only the relevant video clips, reducing storage requirements. A backend API handles communication with dashboards and databases, making the system scalable, efficient, and ready for deployment in smart surveillance environments.

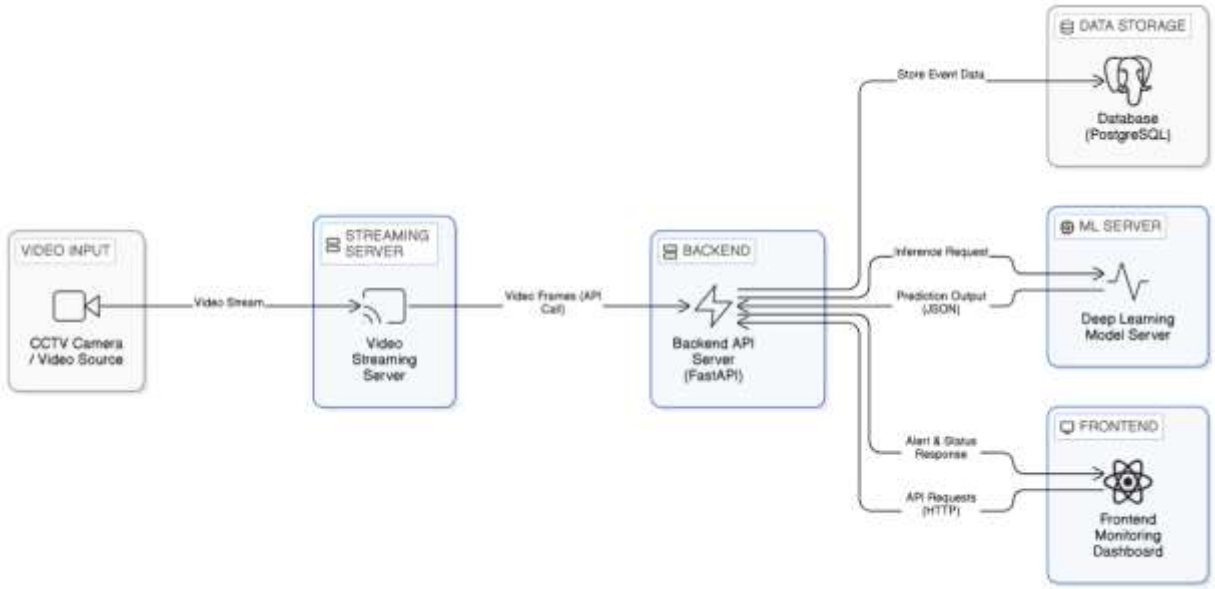

Figure 5. Model Architecture

4. Results

The performance of the proposed CNN–LSTM based crime detection system was evaluated using a surveillance video dataset containing 500 videos distributed across 10 activity classes, namely Abuse, Arson, Assault, Burglary, Explosion, Fighting, Normal, Road Accidents, Robbery, and Shooting.

The dataset was divided into training and validation sets using an 80:20 ratio. The training set was used to learn spatial and temporal features, while the validation set was used to evaluate model performance and monitor overfitting during the training process.

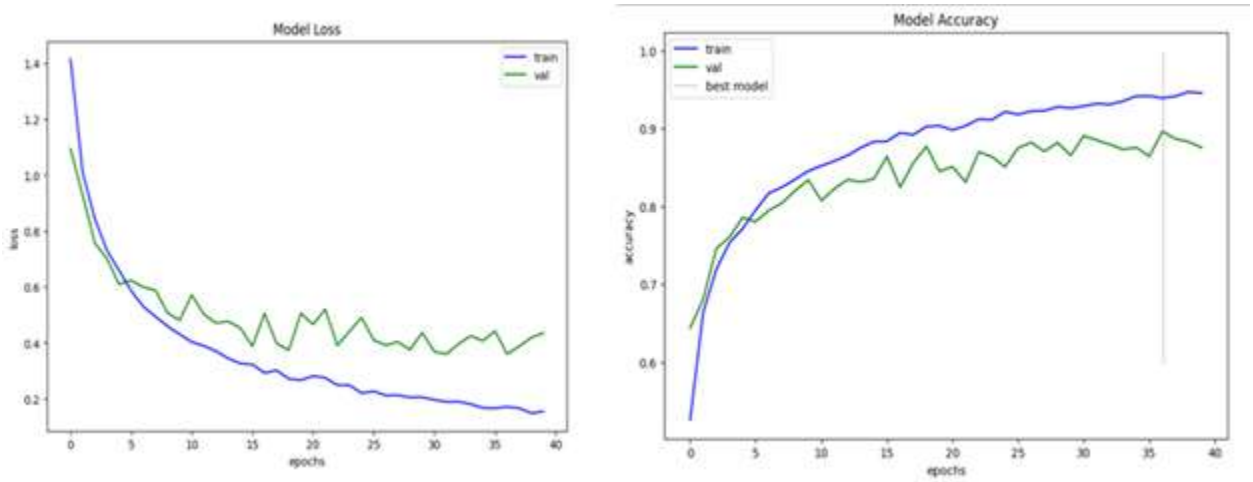

Figure 6. Training and validation accuracy across epochs

For training, video frames were extracted and resized to 224 × 224 pixels, and sequences of frames were used as input to the CNN–LSTM
network. The model was trained for 40 epochs.

Fig. 6 illustrates the training and validation accuracy obtained during the learning process. The training accuracy gradually increased throughout the training process and reached approximately 95%, while the validation accuracy stabilized around 89%, indicating good generalization capability of the proposed model.

Fig. 6 shows the training and validation loss curves. The training loss decreased steadily and converged near 0.15, while the validation loss stabilized around 0.42. The relatively small gap between training and validation curves indicates that the model does not suffer from severe overfitting and is capable of learning meaningful spatio-temporal representations from surveillance videos.

To further evaluate the model performance, Precision, Recall, and F1-Score metrics were computed for each activity class. The detailed classification performance is presented in Table I.

Table 1. Performance metrics

Class	Precision	Recall	F1-Score	Support
Abuse	0.87	0.86	0.86	50
Arson	0.88	0.84	0.86	50
Assault	0.90	0.88	0.89	50
Burglary	0.89	0.90	0.89	50
Explosion	0.86	0.84	0.85	50
Fighting	0.91	0.89	0.90	50
Normal	0.92	0.94	0.93	50
Road Accidents	0.90	0.91	0.90	50
Robbery	0.88	0.89	0.88	50
Shooting	0.87	0.86	0.86	50

Metric	Precision	Recall	F1-Score	Support
Accuracy	**0.89**	**0.89**	**0.89**	**500**
Macro Average	0.89	0.88	0.88	500
Weighted Average	0.89	0.89	0.89	500

The results demonstrate that the proposed model achieves consistent performance across different crime categories. The Normal class achieved the highest F1-Score of 0.93, indicating strong ability to distinguish normal activities from anomalous behavior. High performance was also observed in Fighting, Road Accident, and Burglary classes, suggesting that the model effectively captures spatio-temporal patterns in surveillance videos.

Overall, the obtained validation accuracy of 89% confirms that the proposed CNN–LSTM architecture is capable of effectively detecting criminal activities in surveillance footage and can support real-time intelligent monitoring systems for public safety applications.

5. Conclusion and Future Scope

In this research, we developed a complete end-to-end AI-powered criminal activity detection system capable of monitoring CCTV video streams in real time. The core of the system integrates a spatio-temporal deep learning model—combining CNN-based spatial feature extraction with LSTM-based temporal behavior analysis—to accurately identify complex activities such as fighting, assault, burglary, and explosions. This approach allows the system to detect not just objects or movements, but meaningful behavioral patterns across consecutive frames.

Beyond the model architecture, our project also implements a fully functional deployment pipeline, as illustrated in the system diagram. The CCTV video feed is processed through a streaming server, which converts the continuous footage into frames. These frames are then transmitted to the backend (FastAPI), which communicates with the deep learning inference server to obtain predictions. Detected events are stored in a PostgreSQL database, while alerts and status updates are delivered to a web-based frontend dashboard for real-time monitoring. This practical integration of video streaming, backend processing, model inference, and frontend visualization demonstrates the system's readiness for real-world use.

The overall system design significantly reduces manual surveillance effort by automatically identifying high-risk activities and storing only relevant video clips. Its lightweight and modular architecture ensures scalability and easy deployment across schools, offices, apartments, and public spaces. This work bridges the gap between academic action-recognition models and operational surveillance systems by delivering a behavior-aware, real-time, and deployable security solution.

Future enhancements may include integrating attention mechanisms for finer motion understanding, extending the model to additional crime categories, and deploying edge-optimized versions of the system for low-power devices.

References

1. M. M. Ali, "Real-time video anomaly detection for smart surveillance," IET Image Processing, Dec. 2022. doi: https://doi.org/10.1049/ipr2.12720.

2. V. Mandalapu, L. Elluri, P. Vyas, and N. Roy, "Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions," IEEE Access, vol. 11, pp. 60153–60170, 2023. doi: https://doi.org/10.1109/access.2023.3286344.
3. A. Elmetwally, Reem Eldeeb, and Samir Elmougy, "Deep learning based anomaly detection in real-time video," Multimedia Tools and Applications, May 2024, doi: https://doi.org/10.1007/s11042-024-19116-9.
4. P. Shanthi and V. Manjula, "A systematic review on CNN-YOLO techniques for face and weapon detection in crime prevention," Discover Computing, vol. 28, no. 1, Sep. 2025, doi: https://doi.org/10.1007/s10791-025-09715-x.
5. E. Cesario, Paolo Lindia, and A. Vinci, "A scalable multi-density clustering approach to detect city hotspots in a smart city," Future Generation Computer Systems, vol. 157, pp. 226–236, Mar. 2024, doi: https://doi.org/10.1016/j.future.2024.03.042.
6. R. Jain and G. Gupta, "Design and Analysis of Novel Hybrid CNN-LSTM Approach for Detecting Cybersecurity Threats in IoT Networks," International Journal of Science and Research (IJSR), pp. 1012–1023, Jul. 2025, doi: https://doi.org/10.21275/sr25630141216.
7. A. Kumar and R. Sinha, "Graph Neural Network Approach for Crime Hotspot Prediction," Applied Intelligence, Springer, 2023.
8. T. Wang and L. Zhao, "3D-CNN with Attention Mechanism for Real-Time Violent Activity Detection," Pattern Recognition Letters, 2022.
9. A. Elmetwally, R. Eldeeb, and S. Elmougy, "Deep Learning Based Anomaly Detection in Real-Time Video," Multimedia Tools and Applications, 2024. doi:10.1007/s11042-024-19116-9.
10. W. Ullah, A. Ullah, T. Hussain, Z. A. Khan, and S. W. Baik, "An Efficient Anomaly Recognition Framework Using an Attention Residual LSTM in Surveillance Videos," Sensors, vol. 21, no. 8, p. 2811, 2021.
11. D. Manju et al., "Early Anomalous Action Detection in Surveillance Video Using MRCNN-LSTM Classification," Engineering, Technology & Applied Science Research, vol. 15, no. 4, pp. 25668–25676, 2025.
12. W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018.
13. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv preprint arXiv:1409.1556, 2014.
14. Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, pp. 436–444, 2015.
15. J. Donahue et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015.