Indonesian Journal of Electrical Engineering and Computer Science
Vol. 41, No. 3, March 2026, pp. 1105~1116
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v41.i3.pp1105-1116

An innovative deep learning based approach for anomaly detection in intelligent video surveillance

Megha G. Pallewar[1], Vijaya R. Pawar[2]
[1]Department of Electronics and Telecommunication Engineering, COEP Technological University, Pune, India
[2]Department of Electronics and Telecommunication Engineering, Bharati Vidyapeeth's College of Engineering for Women, Savitribai Phule Pune University, Pune, India

Article Info

Article history:

Received Apr 26, 2024
Revised Oct 15, 2025
Accepted Dec 13, 2025

Keywords:

Anomaly detection
Convolutional neural networks
Deep learning
Human activity detection
Long short-term memory

ABSTRACT

Nowadays, anomaly detection has gained vital importance as security is a major concern everywhere. This work focuses on developing an intelligent video surveillance system capable of detecting anomalous activities in videos, utilizing the UCF Crime dataset as the primary source. The proposed model employed a multistage method uniting the convolutional neural networks (CNN) and long short-term memory (LSTM) networks. In the proposed approach, video frames serve as input to the CNN, which processes them to extract key features. These features are then passed to an LSTM network to capture temporal dependencies and identify anomalous events over time. This CNN-LSTM architecture successfully detects twelve distinct types of anomalous activities: abuse, arrest, arson, assault, burglary, explosion, fight, road accident, robbery, stealing, shoplifting, and vandalism. The dataset is divided into portions for training, testing, and validation, along with cross-validation to ensure model generalization. The system achieves an accuracy of 98.6%, reflecting a significant improvement of 4-5% over existing systems. This demonstrates the robustness of the proposed method in detecting anomalous behavior in video data.

Corresponding Author:

Megha G. Pallewar
Department of Electronics and Telecommunication Engineering, COEP Technological University
Pune 411005, Maharashtra, India
Email: msyannawar@gmail.com

1. INTRODUCTION

Anomaly detection is the technique commonly used to identify data patterns or incidences that drastically diverge from the predicted norms. These anomalies may result from changes in the underlying phenomenon or from the appearance of previously unseen scenarios in the monitored environment [1]-[3]. The process of identifying and localizing anomalous instances in security footage can be done manually, somewhat mechanically, or entirely automatically. The majority of traditional surveillance systems are operated manually and only by humans. This method demands constant attention and concentration in order to judge events or behavior to determine whether they are unusual or suspicious. But this procedure is quite labor-intensive since one must monitor several video feeds at the same time. The operators' mental load often results in lower focus and operational inadequacies [4].

In order to ensure public safety in urban areas, video monitoring is crucial. Video surveillance plays a crucial role in ensuring the safety of individuals. Video data produced by security systems contains a vast amount of valuable information. The outputs from these systems are rich in content, offering important insights embedded within the footage. The need for public safety emphasizes the analysis and processing of video data. The primary aim is to enhance security in public areas, with the objective of streamlining and

simplifying the tasks of security personnel. Analyzing and interpreting this data becomes increasingly important as the need for public safety increases. The developed system's goal is to reduce the workload for security services and improve public area security. In typical surveillance scenarios, security personnel observe a display connected to video cameras, looking for anomalous events such as traffic accidents, fires, explosions, robberies, and stampedes. However, relying on human evaluation alone is far from ideal, as it demands sustained focus over extended periods, a task that is both mentally exhausting and prone to error. Thus, an automated system is crucial for detecting and recognizing abnormal human activities in real-time, with the added capability of notifying relevant authorities immediately. Automatically monitoring CCTV cameras to identify suspicious activities and generating timely responses can play a vital role in preventing accidents or mitigating damage. By eliminating the need for constant human monitoring, automated systems offer a highly efficient solution for video anomaly detection, saving time and labor while freeing operators from tedious tasks.

As a result, many researchers have shifted focus to semi-automatic surveillance systems that minimize human intervention. These systems can make decisions independently, without requiring continuous human oversight. In fully automated surveillance, the key challenge goes beyond recognizing abnormal events in video streams—it also involves accurately detecting and localizing these kinds of situations and human actions. The extraction of robust and discriminative features is essential to solving this complex problem effectively. Detecting unusual and abnormal events presents several challenges, largely due to the varying patterns in different scenes. As scenarios and applications evolve, so do the types of anomalies. Video anomaly detection systems are widely employed across diverse sectors, including healthcare services, security at airports, shopping malls, or railway stations, and surveillance for traffic monitoring or theft detection [5]-[10]. Abnormal activity recognition has vast potential across many fields, especially as data continues to expand in complexity and encompass various forms of information. Timely detection of abnormalities can prevent machinery failures, improve performance, control disease outbreaks, or even save lives. An intelligent video surveillance system is always useful for multiple operations associated [11].

Dubey *et al.* [12] suggested 3D deep multiple instance learning with ResNet (MILR) technique used in the extraction of temporal and spatial characteristics from the videos. The anomaly score was then obtained employing these features. Huang *et al.* [13] created the temporal spatial convolutional neural network (TSCNN), a convolutional neural network (CNN) that accepts continuous frame sequences as input. The network was created to detect human activity in real time using 3D convolution. The lowest derived accuracy is 81.8%, while the average classification accuracy is 94.6%. Then, prominent work is done by Tang *et al.* [14] who established a CNN for human activity recognition (HAR) that makes use of Lego filters. The issues with standard CNN, such as the requirement for temporal dimensions in processing units (filters) and unit sharing across several sensors, are resolved by employing Lego CNN. Thus, one can create a more effective HAR model by substituting Lego filters for the standard filters. The established model is then applied to five publicly available datasets, and the outcomes are contrasted. Kumar and Sailaja [15] proposed a deep learning based approach by using a bidirectional long short-term memory (Bi-LSTM) network combined with skeleton activity forecasting (SAF) to recognize anomalous human activities. The pose estimation performed using the human subject's skeleton joint points. Further the developments are majorly focused on the variety of deep learning techniques combinations always progress with the scalability and accuracy of the system. Few commonly used methodologies are ASRNet, Conv2D, convolutional long short-term memory (ConvLSTM), Bi-LSTM with CNN [3], [16]-[21]. The outcomes coming with these methodologies are having limitations related with limited dataset, constrained scenario for detection, limited focused activities and good accuracy only for simple activity base. This review of the literature indicates that automatic video surveillance systems with defined and restricted criteria have been reported to have good accuracy. Therefore, machine learning-based classifiers with an unconstrained parameter approach are used in the current system.

The primary challenges can be identified as follows:

- The challenges faced with environmental conditions such as lighting conditions which changes with day-night, indoor-outdoor footage in video.
- Background complexity is also a major issue.
- It is challenging to identify activities involving many people than activities involving a single person or two people.
- Scalability issues arise when managing large-scale surveillance systems with numerous cameras and enormous volumes of data.
- Activities that look similar (e.g., fighting vs. abusing or robbery vs. burglary) can be difficult to distinguish.

A key challenge in abnormal incident classification is the lack of sufficient datasets containing relevant videos. Few existing datasets include crucial event types such as robbery, abuse, or harassment. High-

quality datasets are the foundation of successful neural network models, as they enable these systems to learn and accurately classify incidents. To build an architecture that delivers truthful results, a database must be used which covers almost all abnormal events. So the current methodology employs the UCF Crime Database, which includes 13 abnormal activities, such as: abuse, arrest, arson, assault, burglary, explosion, fight, road accident, robbery, shooting, stealing, shoplifting, and vandalism. This comprehensive dataset supports the development of a proficient surveillance system capable of autonomously identifying suspicious behavior. By focusing on challenges related to background complexity, camera angles, and environmental conditions, the system leverages a versatile dataset to improve processing accuracy.

Anomaly detection has broad applications, from identifying traffic accidents to flagging suspicious activities. Convolutional neural networks (CNNs) excel at detecting anomalous activities in video surveillance by efficiently learning hierarchical features from raw data, allowing for precise identification of complex behaviors. The proposed system uses a convolutional neural network–long short-term memory (CNN-LSTM) model, where CNNs handle feature extraction and the long short-term memory (LSTM) manage classification, aiming most recent challenges in these years. The rest of the paper is structured as follows to provide a comprehensive overview of the work. Section 2 describes the present methodology. Section 3 elaborates on the experimentation. Section 4 discusses performance analysis and results. Section 5 presents the conclusion and future scope.

2. METHOD

The process for video-based anomaly detection with CNN and LSTM is shown in Figure 1. This methodology comprises six stages:

- Dataset selection: the first stage involves selecting a dataset that encompasses a wide variety of activities. For this study, the UCF crime database is used. Maintained by the University of Central Florida (UCF) and the UCF Police Department, this dataset includes videos of various criminal activities such as theft, assault, and vandalism, captured on campus. The dataset is extensive, with approximately 75 videos per activity, resulting in a diverse collection.
- Video to frame conversion: videos from the dataset are converted into individual frames for further processing. Each activity video generates approximately 6,000 to 7,000 frames.
- Pre-processing: the retrieved frames undergo several preprocessing techniques to enhance quality. These techniques include edge detection, noise elimination, and background removal. The histogram of oriented gradients (HoG) is applied to the preprocessed frames to obtain filtered images.
- Feature extraction: the preprocessed and filtered images are then fed into a multi-stage CNN-LSTM model. This model uses a layered structure to extract and analyze detailed feature attributes.
- Action detection: the multi-layered CNN-LSTM architecture processes the feature attributes to detect and classify actions effectively.
- Anomaly detection: finally, the system evaluates the processed data to identify and flag anomalous activities based on the learned patterns.

The study focuses on anomaly detection in video streams, with an emphasis on identifying complex activities. While previous methods have achieved good results, they primarily focused on simpler, single-viewed activities. This study aims to improve accuracy in more challenging environments through comprehensive experimentation. The empirical analysis is carried out by using the following stages:

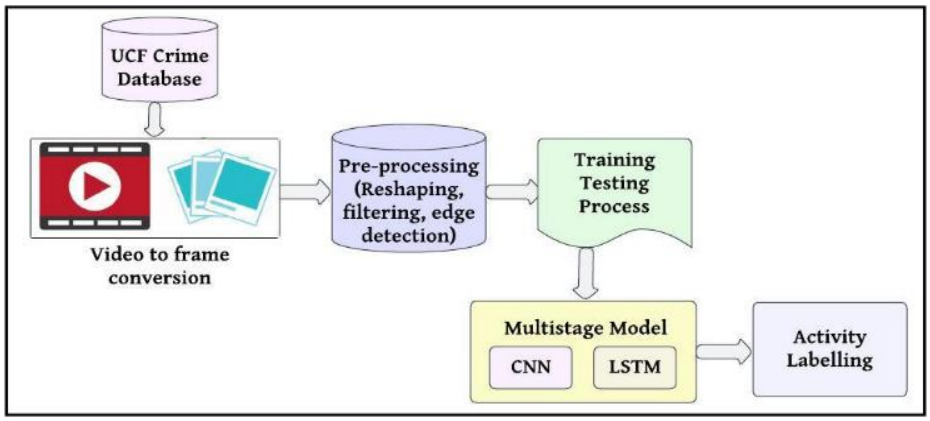

Figure 1. The proposed model for anomalous activity detection

2.1. UCF crime database

In terms of scale and complexity, the UCF Crime Database, created by Sultani *et al.* [22], represents a significant improvement over earlier databases. It contains raw CCTV footage showing thirteen typical real-life crime scenes, including robbery, explosion, fight, theft, shoplifting, and vandalism. The collection includes around 2,000 videos comprising 128 hours of footage. These videos have an average runtime of four minutes, which makes the dataset more valuable for classifying anomalies based on videos. The averages of 7,247 frames are generated for each video. The dataset is divided into two streams as normal recordings and raw real-world recordings. Initially, the research concentrated on classifying and predicting incidents of abuse and fighting. As the study progressed, it expanded to cover six distinct actions and ultimately extended to include all 12 actions represented in the dataset. The results were achieved using the same set of algorithms throughout these different research phases.

2.2. Multistage approach with layered CNN-LSTM

In order to get features from spatial and temporal data CNN and LSTM are combined as CNN handles spatial features well and LSTM works perfectly with temporal features [23]. The proposed work executed by CNN-LSTM architecture which includes three 2D convolution layers and then a LSTM layer. The typical CNN architecture is shown in Figure 2. The layers of convolutional network share equal number of filters and kernels. The input to the very initial convolution layer consists of images with size of 224×224×3 (height, width, and number of channels). The input is further handled by multiple convolutional layers stacked together, with rectified linear unit (ReLU) serving as the activation function for each layer, involving 4,096 nodes. Each layer contains 16 filters and a kernel of 3×3 size is used. Every convolution layer is followed by maxpooling layer. The CNN includes max-pooling operations, which help decrease the dimensions of the output feature vectors, making the representation more manageable and computationally efficient. Spatial features extracted by CNN are subsequently sent to LSTM layers. After the CNN layer, a LSTM layer used which has 3×3 kernel sizes and has 64 filters. This layer captures spatio-temporal information by processing sequential data, which is essential for video-based anomaly detection. The network architecture then consists of a dropout layer, a flatten layer, and a dense layer. An added pair of dropout and dense layers is used at the conclusion of the network. At the end of the network, additional pair of dropout and dense layers is employed and the softmax activation function is employed with every dense layer. This layer is responsible for making the final classification decisions and contains approximately 1,500 nodes.

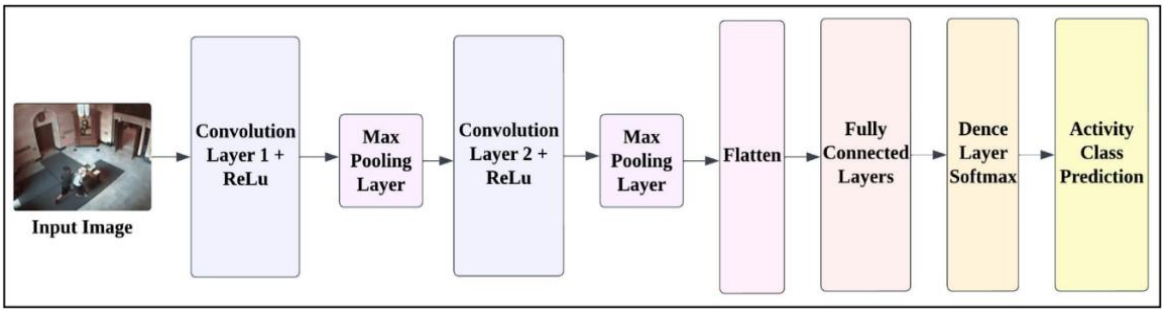

Figure 2. The CNN architecture

Tables 1 and 2 show the output shape of each layer, detailing the output shapes with total number of parameters. The LSTM architecture, shown in Figure 3, is a type of recurrent neural networks (RNNs). These are generally used to manage the vanishing gradient problem in traditional RNNs by enabling the capture of abiding dependencies in sequential data. Each LSTM cell contains three gates: the input gate, which decides what information to add; the forget gate, which determines what to discard; and the output gate, which selects what information to pass forward. These gates, controlled by sigmoid and tanh activations, allow LSTMs to retain or forget information over time. As a result, LSTMs are highly effective for temporal sequence modeling to detect anomalies which are used for video analysis. Temporal dependencies are captured by the LSTM layers through their sequential processing of the spatial information [17], [24], [25]. Understanding long-range dependencies in the data is made possible by the internal memory state that the LSTM layers maintain. Regression, classification, and other tasks can be performed with the final output of the LSTM layers.

Table 1. Layer-wise summary of CNN model

Layer (type)	Parameters
Rescaling 1	0
Conv2D	448
Maxpooling2D	0
Conv2D 1	4640
Maxpooling2D 1	0
Conv2D 2	18496
Maxpooling2D 2	0
Flatten	0
Dense	3965056
Dense 1	1548
Total parameters	3990188

Table 2. Layer-wise summary of LSTM model

Layer (type)	Number of parameters
Input layer	0
LSTM	1,287,168
LSTM	197,120
LSTM	49,408
Flatten	0
Dense	491,776
Dropout	0
Dense	32,896
Dropout	0
Dense	258

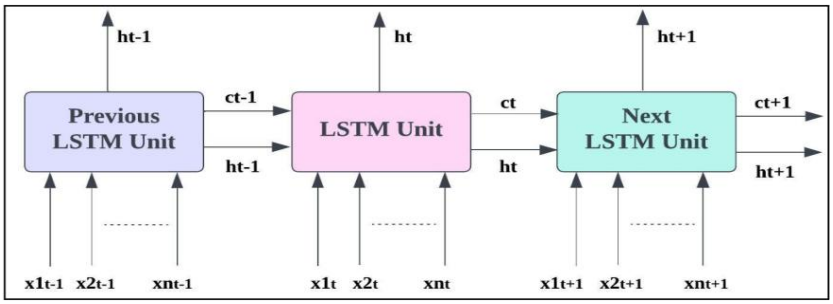

Figure 3. The LSTM model

LSTM architectures typically incorporate three fundamental types of logical gates that serve as the building blocks of the system. The essential elements within an LSTM are h_t and c_t, which play a crucial role in its functioning. These two units, h_t and c_t, are analyzed and computed using (1)-(5):

$$\text{Forget gate: } f_t=\sigma(W_f \cdot[h_{t-1},x_t]+b_f) \quad (1)$$

$$\text{Input gate: } i_t=\sigma(W_i \cdot [h_{t-1},x_t]+b_i)$$
$$\tilde{C}_t=\tanh(W_C \cdot [h_{t-1},x_t]+b_C) \quad (2)$$

$$\text{Cell state update: } C_t=f_t *C_{t-1}+i_t\tilde{C}_t \quad (3)$$

$$\text{Output gate: } o_t=\sigma(W_o \cdot [h_{t-1},x_t]+b_o) \quad (4)$$

$$\text{Hidden State: } h_t=o_t*\tanh(C_t) \quad (5)$$

Here, the sigmoidal function is denoted by σ, the hyperbolic tangent is nothing but tanh, and · gives the multiplication of matrix weights W_i and respective gate values. The LSTM model summary is provided in Table 2. To ensure continuous processing, the CNN-LSTM multistage architecture, as shown in Figure 4, starts by extracting frames from video clips, with a threshold set at 40 frames. Each image generates frames with specific height and width, which are fed to the neural network. During pre-processing, the input data were cropped and adjusted. The total dataset is divided with a split of 70% for training, 20% for validation, and 10%

for testing. The CNN model extracts features from the frames, producing 25,088 features per frame, resulting in 40 sets of feature vectors. These vectors are then sequentially processed by the LSTM network. The final output is obtained from the dense layer at the last stage of the CNN-LSTM model. Once the training process was completed using the training data, the model was evaluated on the testing data, and the accuracy and confusion matrix were computed.

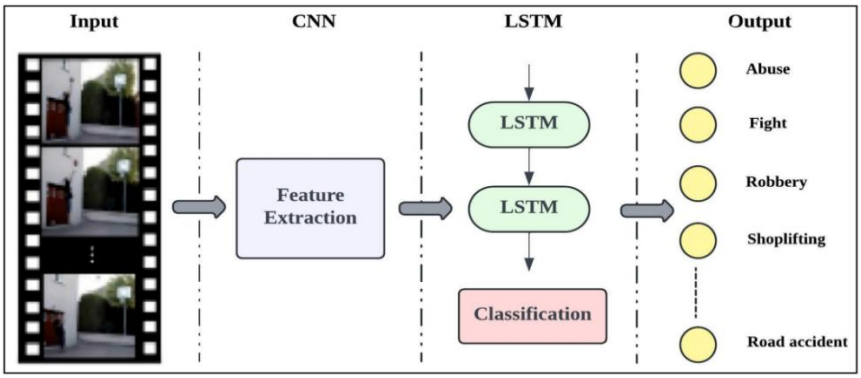

Figure 4. System architecture of CNN-LSTM model

3. EXPERIMENTATION

Extensive experimentation was conducted using a Python 3.7.0 environment with a 2.7 GHz Intel Core i7-4600U CPU, 8GB RAM, and a 64-bit operating system. The testing involved different frame counts of 20, 40, and 60 frames. Each action class was trained with 50 frames prior to testing. The experimentation varied the number of epochs (10, 20, and 30), along with adjustments to batch sizes, filter counts, and the number of dense layers. Ultimately, results were gathered using 30 epochs. Initially, the focus was on classifying two activities, which was later expanded to six activity classifications. The model performed exceptionally well with six classes. Building on this success, the model was extended to classify 12 activities, including: abuse, arrest, arson, assault, burglary, explosion, fight, road accident, robbery, shooting, stealing, shoplifting, and vandalism. Figure 5 illustrates the results for each class and the labeling of the 12 activity classes, respectively. Additionally, the model was tested with inputs of 60, 80, and 100 frames per class.

Figure 5. The 12 class activity labelling

The methodology is widely referenced for scene analysis and suspicious activity detection [26]-[30]. The model's performance was evaluated using the area under the curve (AUC) to measure its overall ability to distinguish between classes, and accuracy to determine the proportion of correctly classified instances. Table 3 compares the video classification results with various methodologies, including SVM, CNN-Bi LSTM, DenseNet, and ResNet. The present system achieved a performance accuracy of 98.6% and an AUC of 90%. The comparison highlights that the current multistage approach delivers superior accuracy for the UCF Crime dataset compared to other methodologies.

Table 3. Result of video classification

Model used	Year	Dataset	Accuracy	AUC
CNN-Bi LSTM [28]	2023	UCF11	98.9	--
		UCF50	96.04	
		UCF crime	61.04	
Cubic SVM KNN [29]	2021	CIFAR100	99.24	--
VAE [30]	2021	UCSDped1	--	92.3
		Avenue		82.1
DenseNet121 [31]	2023	UCF crime	--	86.63
ResNet (MILR) [12]	2019	UCF crime	--	
ST-GCN [32]	2023	UCSD Ped1	89.1	85.5
		UCSD Ped2	-	97.9
		ShanghaiTech	-	83.8
CNN-LSTM (our model)	2023	UCF crime	98.6	90

4. RESULTS AND PERFORMANCE ANALYSIS

The system can be analyzed using different parameters. These parameters are commonly used metrics that elaborate specifically about the system performance and behavior for applied conditions. The included performance parameters are accuracy, specificity, recall, and precision. This section helps to understand various metrics, firstly, and then the accuracy, loss curves, and confusion matrix are discussed. Further, the parametric plots are analyzed to understand the overall behavior of the established system.

4.1. Evaluation indicators

The following are the assessment metrics to analyze the effectiveness of the implemented model:
Accuracy: It is a parameter used to measure the proportion of correctly identified anomalies and normal instances among the total number of instances.

$$\text{Accuracy} = \frac{\text{True Positive+True Negative}}{\text{Total Samples}} \tag{6}$$

Precision: it is referred to as the proportion of true anomalies among the instances flagged as anomalous by the model.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+ True Negative}} \tag{7}$$

Recall: this indicator measures the ability of the system to correctly identify actual anomalies.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+ False Negative}} \tag{8}$$

F1-score: This one is the harmonic mean of precision and recall, which provides a stable metric to measure when classes are not balanced.

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision+ Recall}} \tag{9}$$

False positive rate (FPR): this is proportion of normal actions classified incorrectly as abnormal actions.

$$\text{FPR} = \frac{\text{False Positive}}{\text{True Negative+ False Positive}} \tag{10}$$

False negative rate (FNR): it is the proportion of actual abnormalities incorrectly identified as normal actions.

$$\text{FNR} = \frac{\text{True Negative}}{\text{True Negative+ False Negative}} \quad (11)$$

Area under the receiver operating characteristic curve (AUC-ROC): The changes occur with true positive rate and false positive rate due to various threshold values are graphically represented by the AUC-ROC curves. The value of AUC indicates the performance of the model. Higher AUC is considered better as it differentiates the normal and abnormal actions significantly.

Confusion Matrix: It is a table that presents the true positives, false positives, true negatives, and false negatives, offering a detailed insight into the model's performance.

4.2. Training and validation accuracy and loss curves

The CNN-LSTM model's performance can be assessed using various types of curves; here, the system was verified with two of them, training and validation accuracy as well as loss curves. Both the curves are referred here to assist in identifying any problems such as overfitting or underfitting by visually representing how the model learns over time. At every epochs (iteration) in the training process, the training accuracy curve illustrates the model's accuracy on the training data. The model's accuracy on the validation dataset—which it does not observe during training—is represented by the Validation accuracy curve. It is employed to observe the model's capacity for adaptation. Ideally, both the training and validation accuracy should increase over time. The training dataset's loss (error) is represented by the training loss curve at each epochs. System aims to reduce the loss. The loss on the validation dataset is displayed by the validation loss curve. Better model performance is indicated by a lower loss. The information provided by the training and validation loss values is crucial because it helps us identify any learning issues that may result in an underfit or an overfit model by allowing us to see how systems typical behavior verified with the iteration of succeeding epochs. At the predicting stage, they will also tell us which epochs to utilize the training model weights. Figure 6 shows the accuracy curves for training and validation after 2 and 10 epochs. It can be easily observed that the accuracy got improved as the number of epochs increases. Figure 7 shows the training and validation loss curves for 2 and 10 epochs. The loss function decreases with an increase in epochs.

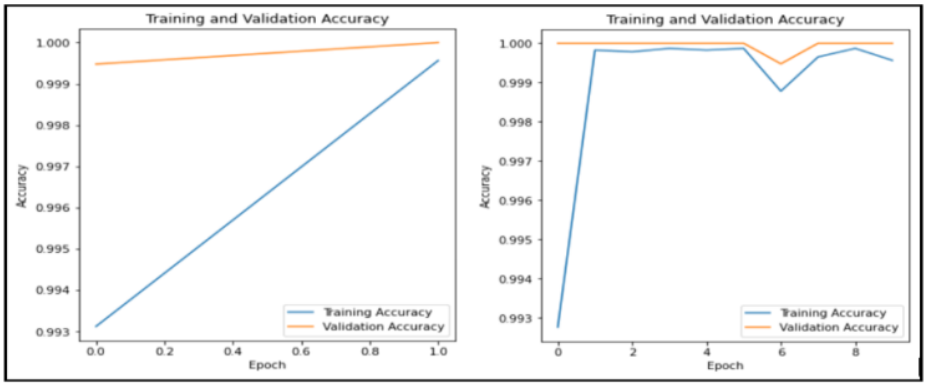

Figure 6. Training accuracy and validation accuracy curves for 2 and 10 epochs

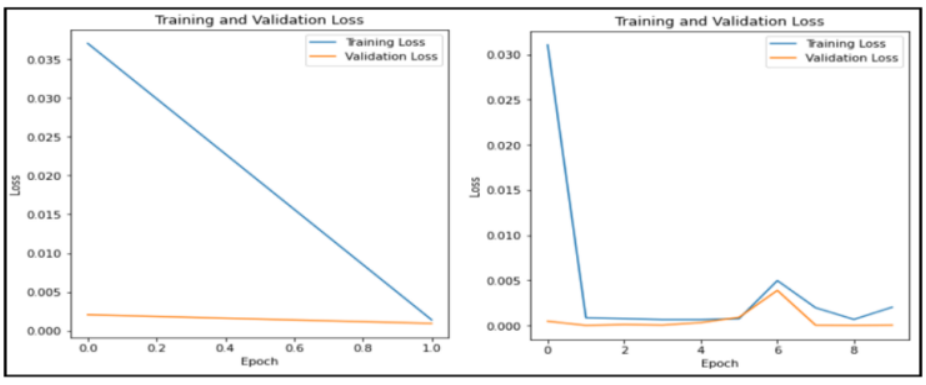

Figure 7. Training loss and validation loss curves for 2 and 10 epochs

4.3. Confusion matrix

A confusion matrix assists in the visualization of the outcome of a classification task by giving a table structure of the various predictions and outcomes. This confusion matrix shows you the proportion of times each activity was misidentified (false positives and false negatives) compared to the proportion of times it was correctly detected (true positives). This has the potential to bring emphasis on particular actions that the model struggles to recognize. Figure 8 shows the confusion matrix with individual class-wise accuracy. Through which the percentage accuracy of all 12 classes can be identified.

The matrix reveals that the highest misclassification occurs with the assault action, which is often mistaken for abuse or a fight. Similarly, abuse is frequently misclassified as assault or fight. This is likely due to the similarity in the actions involved. However, the percentage of accurately matching the actual and predicted actions remains relatively high, providing a clear indication of the implemented model's performance across the various activities.

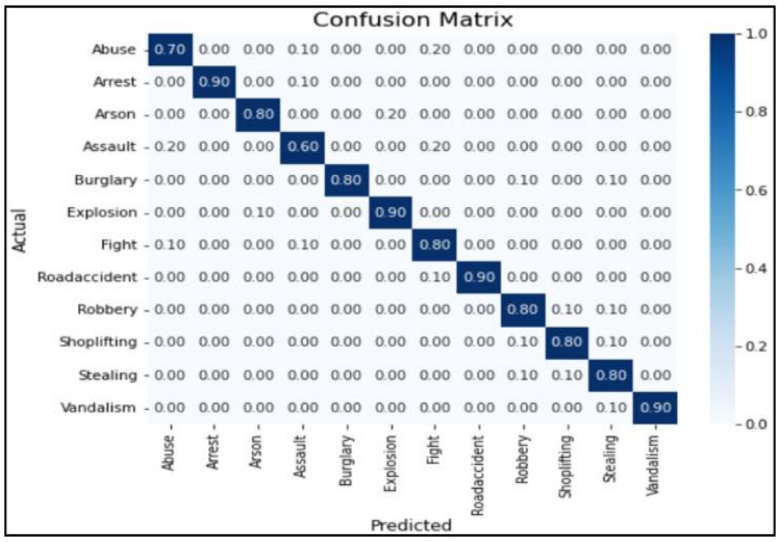

Figure 8. Confusion matrix for 12 activity recognition

4.4. Performance parameters and their plots

Precision, recall, F1 score, and accuracy are frequently represented graphically in order to demonstrate how well the classification models are performing. It is simpler to understand and evaluate the performance of classification models across several classes when precision, recall, F1 score, and accuracy are shown graphically. They provide an illustrative overview of a model's strengths and potential areas for improvement. The class-wise comparison is shown in Figures 9 and 10. A comparative analysis with existing systems is presented, as illustrated in Figure 11.

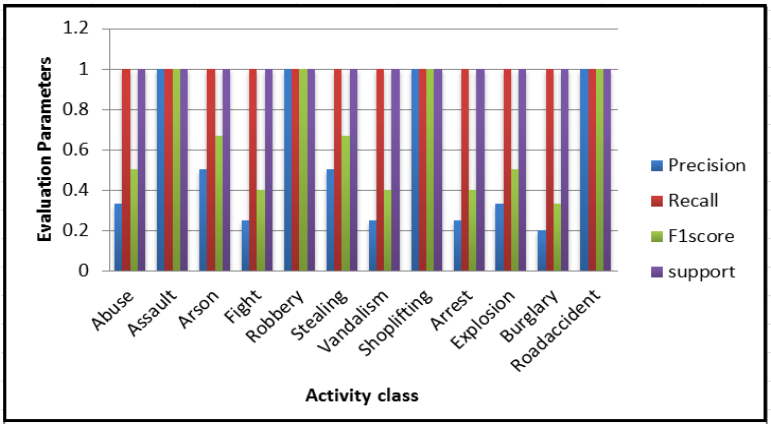

Figure 9. Class-wise performance parameter

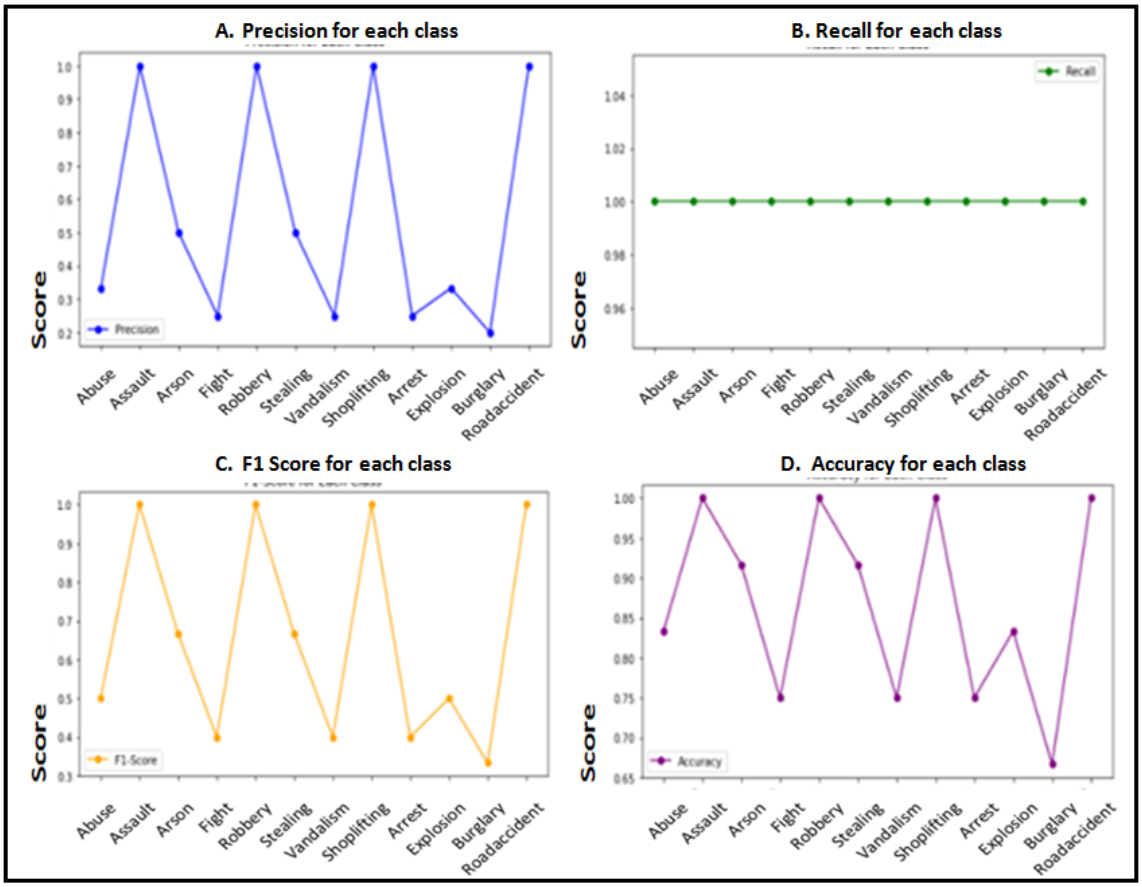

Figure 10. Class wise line graph for precision, recall, accuracy and F1 score

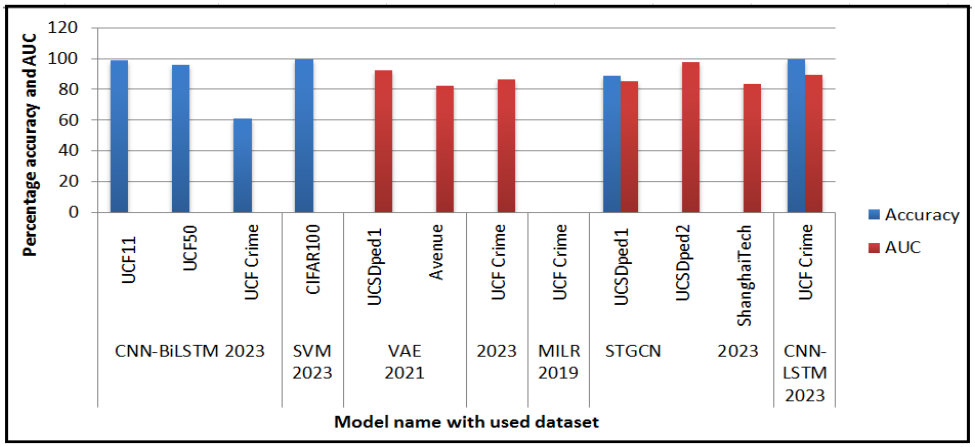

Figure 11. Comparison with other anomaly detection systems

The UCF Crime dataset is more complex since it gives a variety of actions and video kinds for each action, in contrast to other publicly available datasets that only include videos with fixed standpoints, uniform backdrops, and usual angles. This challenging UCF Crime dataset is used to assess the CNN-LSTM model's experimental outcomes. The UCF Crime dataset records crime activity in real time, captured by CCTV cameras in unstructured surroundings, while other datasets concentrate on frequent actions. Because of its diversity, it serves as a useful baseline for evaluating anomaly detection methods' performance in practical settings. The present model works excellently for maximum classes for the task of anomaly detection. The performance is improved with difficult activity database.

5. CONCLUSION AND FUTURE SCOPE

The experimental results demonstrate that the current CNN-LSTM model excels in identifying anomalous activities within the UCF Crime dataset. This approach effectively detects a variety of abnormal behaviors likely to occur in public spaces, enhancing human safety by enabling timely responses from monitoring systems or authorized personnel. In future work, the system can be expanded to include all 13 classes present in the UCF Crime dataset. It is anticipated that extending the model to handle these additional classes will yield similar high performance. However, a notable limitation of the current model is its lengthy training time. This issue can be addressed by focusing processing on regions of interest within the video frames, thereby ignoring irrelevant areas and reducing the system's response time. This improvement would enhance overall performance metrics. To further advance real-time video surveillance capabilities, future research could explore self-supervised learning techniques such as momentum contrast (MoCo) and Simple framework for contrastive learning of visual representations (SimCLR). Additionally, incorporating multimodal models like masked autoencoders (MAE) could provide further enhancements in detection accuracy and efficiency.

REFERENCES

[1] J. Ren, F. Xia, Y. Liu, and I. Lee, "Deep video anomaly detection: opportunities and challenges," in *2021 International Conference on Data Mining Workshops (ICDMW)*, IEEE, Dec. 2021, pp. 959–966. doi: 10.1109/ICDMW53433.2021.00125.

[2] K. A. Alaghbari, M. H. Md. Saad, A. Hussain, and M. R. Alam, "Activities recognition, anomaly detection and next activity prediction based on neural networks in smart homes," *IEEE Access*, vol. 10, pp. 28219–28232, 2022, doi: 10.1109/ACCESS.2022.3157726.

[3] I. Ullah and Q. H. Mahmoud, "Design and development of RNN anomaly detection model for IoT networks," *IEEE Access*, vol. 10, pp. 62722–62750, 2022, doi: 10.1109/ACCESS.2022.3176317.

[4] A. A. Sodemann, M. P. Ross, and B. J. Borghetti, "A review of anomaly detection in automated surveillance," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1257–1272, Nov. 2012, doi: 10.1109/TSMCC.2012.2215319.

[5] S. S. Priya and R. I. Minu, "Abnormal activity detection techniques in intelligent video surveillance: a survey," in *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, IEEE, Apr. 2023, pp. 1608–1613. doi: 10.1109/ICOEI56765.2023.10125671.

[6] H.-T. Duong, V.-T. Le, and V. T. Hoang, "Deep learning-based anomaly detection in video surveillance: a survey," *Sensors*, vol. 23, no. 11, p. 5024, May 2023, doi: 10.3390/s23115024.

[7] S. Hossain, A. Abtahee, I. Kashem, M. M. Hoque, and I. H. Sarker, "Crime prediction using spatio-temporal data," in *Computing Science, Communication and Security*, N. Chaubey, S. Parikh, and K. Amin, Eds., Singapore.: Springer, 2020, pp. 277–289. doi: 10.1007/978-981-15-6648-6_22.

[8] N. Shah, N. Bhagat, and M. Shah, "Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, p. 9, Apr. 2021, doi: 10.1186/s42492-021-00075-z.

[9] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 14360–14369. doi: 10.1109/CVPR42600.2020.01438.

[10] B. Degardin and H. Proença, "Iterative weak/self-supervised classification framework for abnormal events detection," *Pattern Recognition Letters*, vol. 145, pp. 50–57, May 2021, doi: 10.1016/j.patrec.2021.01.031.

[11] J. Huang, A. Huang, and L. Wang, "Intelligent video surveillance of tourist attractions based on virtual reality technology," *IEEE Access*, vol. 8, pp. 159220–159233, 2020, doi: 10.1109/ACCESS.2020.3020637.

[12] S. Dubey, A. Boragule, and M. Jeon, "3D ResNet with ranking loss function for abnormal activity detection in videos," in *2019 International Conference on Control, Automation and Information Sciences (ICCAIS)*, IEEE, Oct. 2019, pp. 1–6. doi: 10.1109/ICCAIS46528.2019.9074586.

[13] W. Huang, Y. Liu, S. Zhu, S. Wang, and Y. Zhang, "TSCNN: a 3D convolutional activity recognition network based on RFID RSSI," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jul. 2020, pp. 1–8. doi: 10.1109/IJCNN48605.2020.9207590.

[14] Y. Tang, Q. Teng, L. Zhang, F. Min, and J. He, "Layer-wise training convolutional neural networks with smaller filters for human activity recognition using wearable sensors," *IEEE Sensors Journal*, vol. 21, no. 1, pp. 581–592, Jan. 2021, doi: 10.1109/JSEN.2020.3015521.

[15] D. Kumar and S. R. Sailaja, "Abnormal activity recognition using deep learning in streaming video for indoor application," in *2021 ITU Kaleidoscope: Connecting Physical and Virtual Worlds (ITU K)*, IEEE, Dec. 2021, pp. 1–7. doi: 10.23919/ITUK53220.2021.9662095.

[16] R. Vrskova, R. Hudec, P. Kamencay, and P. Sykora, "A new approach for abnormal human activities recognition based on ConvLSTM architecture," *Sensors*, vol. 22, no. 8, p. 2946, Apr. 2022, doi: 10.3390/s22082946.

[17] A. Murad and J.-Y. Pyun, "Deep recurrent neural networks for human activity recognition," *Sensors*, vol. 17, no. 11, p. 2556, Nov. 2017, doi: 10.3390/s17112556.

[18] K. Deshpande, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Anomaly detection in surveillance videos using transformer based attention model," in *Communications in Computer and Information Science*, vol. 1794 CCIS, 2023, pp. 199–211. doi: 10.1007/978-981-99-1648-1_17.

[19] T. Yuan *et al.*, "Towards surveillance video-and-language understanding: new dataset, baselines, and challenges," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2024, pp. 22052–22061. doi: 10.1109/CVPR52733.2024.02082.

[20] Z. Wang and Y. Chen, "Anomaly detection with dual-stream memory network," *Journal of Visual Communication and Image Representation*, vol. 90, p. 103739, Feb. 2023, doi: 10.1016/j.jvcir.2022.103739.

[21] A. Patwal, M. Diwakar, V. Tripathi, and P. Singh, "An investigation of videos for abnormal behavior detection," *Procedia Computer Science*, vol. 218, pp. 2264–2272, 2023, doi: 10.1016/j.procs.2023.01.202.

[22] W. Sultani, Q. A. Arshad, and C. Chen, "Action recognition in real-world videos," in *Computer Vision*, Cham: Springer International Publishing, 2020, pp. 1–10. doi: 10.1007/978-3-030-03243-2_846-1.
[23] M. G. Pallewar, V. R. Pawar, and A. N. Gaikwad, "Human anomalous activity detection with CNN-LSTM approach," *Journal of Integrated Science and Technology*, vol. 12, no. 1, p. 704, 2024.
[24] Q.-A. Arshad *et al.*, "Anomalous situations recognition in surveillance images using deep learning," *Computers, Materials & Continua*, vol. 76, no. 1, pp. 1103–1125, 2023, doi: 10.32604/cmc.2023.039752.
[25] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: a deep learning approach for short-term traffic forecast," *IET Intelligent Transport Systems*, vol. 11, no. 2, pp. 68–75, Mar. 2017, doi: 10.1049/iet-its.2016.0208.
[26] M. Kumar, A. K. Patel, M. Biswas, and S. Shitharth, "Attention-based bidirectional-long short-term memory for abnormal human activity detection," *Scientific Reports*, vol. 13, p. 14442, Sep. 2023, doi: 10.1038/s41598-023-41231-0.
[27] 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, pp. 89181–89197, 2021, doi: 10.1109/ACCESS.2021.3091081.
[28] H. Gangloff, M.-T. Pham, L. Courtrai, and S. Lefevre, "Leveraging vector-quantized variational autoencoder inner metrics for anomaly detection," in *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, Aug. 2022, pp. 435–441. doi: 10.1109/ICPR56361.2022.9956102.
[29] S. Solanki, Y. Shah, D. Rohit, and D. Ramoliya, "Unveiling anomalies in surveillance videos through various transfer learning models," in *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, IEEE, Oct. 2023, pp. 300–306. doi: 10.1109/ICSSAS57918.2023.10331722.
[30] Y. Yang, Z. Fu, and S. M. Naqvi, "Abnormal event detection for video surveillance using an enhanced two-stream fusion method," *Neurocomputing*, vol. 553, p. 126561, Oct. 2023, doi: 10.1016/j.neucom.2023.126561.

BIOGRAPHIES OF AUTHORS

Megha G. Pallewar has completed M.E. in Digital Circuits in 2013 from Savitribai Phule Pune University, Pune. She has completed her Ph.D. from Savitribai Phule Pune University, Pune. She has teaching experience of 16 years and industrial experience of 2 years. She has filed 01 patent and published 15 papers in reputed journals and conferences. She is life member of ISTE. She can be contacted at email: msyannawar@gmail.com

Vijaya R. Pawar has completed M.E. in Electronics in 2002 from Shivaji University Kolhapur. Bharati Vidyapeeth Deemed University, Pune has awarded her a Ph.D. degree in Electronics Engineering in 2015. She has teaching experience of 29 years and research experience of 10 years. She has filed 02 patents and 2 copyrights in the technical field. She has published more than 47 papers, of which 28 papers are in international journals. Her 25 papers are indexed in Scopus and 12 papers are indexed in SCI. She was invited as session chair for 12 national and international conferences. She has also been invited by 06 colleges as a resource person for delivering a session on," Digital Signal Processing" and "Biomedical signal Processing". She has received one research grant and 08 development grants. She is the life Member of ISTE, IEI, and IETE. She can be contacted at email: Vijaya.kashid@bharatividyapeeth.edu.