Suspicious Activity Detection In Crowd Videos Using Deep Learning Model

R.Thirumalaisamy[1],* S.Kother Mohideen[2]

[1]Reg.No:18221252161005, Research Scholar, Department of Information Technology,
Sri Ram Nallamani Yadava College of Arts & Science, Tenkasi - 627804, Tamilnadu, India,
Affiliation of Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-627012,
Tamilnadu, India.

[2]Associate Professor & Head, PG Research Department Information Technology,
Sri Ram Nallamani Yadava College of Arts &Science, Tenkasi-627804, Tamilnadu, India.

Abstract

Video surveillance systems have become a focal point of interest for both researchers and industries. With the emergence of advanced surveillance cameras equipped with powerful processing capabilities, the development of intelligent visual surveillance systems has become feasible. These systems are required to track crowded areas and detect abnormal activities, including unusual behavior and instances of violence. Due to the continuous nature of surveillance video, it is impractical for humans to manually track and analyze it. Therefore, the prime target of this work is to build a fully automated system that can analyze and identify suspicious activities occurring in crowded places. To achieve this, a deep learning model named Modified DenseNet201 is introduced. Input video sequences are preprocessed using filter to eliminate noise and are then converted into frames. Key frames are extracted, and the frames are augmented to expand the size of the database size. The segmentation task is addressed using U-Net. Following this, the modified Densenet201 is trained on the training data, enabling it to establish the relationship between input and output variables. The performance of the modified DenseNet201 is tested with test data. To validate the effectiveness of the developed system, two different datasets are utilized: an examination dataset and a violent/non-violent dataset. Experimental results demonstrate that the proposed system outperforms the other methods, producing excellent results on both datasets.

Keywords: Abnormal behavior detection, deep learning, examination hall, random forest, and surveillance system

1. INTRODUCTION

Behavior analysis in computer vision is an extremely challenging task, particularly when it comes to analyzing human abnormal activity in crowed environment. While there has been substantial focus on analyzing actions performed by individuals, the attention given to crowded scenes has been comparatively minimal [1]. However, crowd behavior analysis has the potential to greatly impact various domains, including public safety. detecting chaotic behavior within crowds is of immense value to relevant authorities.

Crowd scene analysis poses even greater challenges compared to analyzing individual human activity due to several reasons. The high density of people in such scenes often poses difficulties for detection algorithm that struggle to accurately identify individual entities [2]. The available data for analysis is frequently of low quality and real-world data to be analyzed are typically accessible only to authorized personnel due to legal and privacy restrictions.

Examinations play a crucial role in every educational program. Dealing with academic dishonesty is typically handled at the classroom or institutional level [3]- [5]. However, in the context of covid-19, there is a significant need for student cheating detection systems in universities to ensure the safety of both invigilators and students. These systems are in high demand, and they require an intelligent framework that can issue alarms when suspicious activity is detected. By accurately detecting student activities, invigilators can respond to instances of misconduct during exams. The primary contributions are as follows: -

- A fully automated system is built to detect suspicious activity in crowded places.
- A new deep learning network, a modified DenseNet201is proposed.
- Effectiveness of the system is verified using crowd scene and examination hall databases
- The results exhibit the satisfactory achievement of the intended task.

The paper is arranged as follows: Section 2 provides a brief review of related works. Section 3 explains the proposed method. Section 4 analyses the effectiveness of the developed system. Section 5 summarizes the paper.

2. LITERATURE REVIEW

In this section, various approaches utilized in literature to detect human activity are explored. Deep learning networks are used across a diverse range of tasks and have showed impressive outcomes in a wide range of applications. Jalel et al. [1] used Convolutional Neural Network (CNN) to detect human activity. In this approach, the combination of multiple feature extraction techniques and CNN was employed to improve the classification rate. Nevertheless, the utilization of multiple feature extraction methods leads to a considerable increase in computational overhead. Mohammadi et al. [2] introduced a violence detection system utilizing the concept of Substantial Derivative (SD). Their approach focused on identifying violent activities in videos.

The network architecture was specifically designed for feature extraction, and two classifiers, namely Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), were employed to classify these extracted features. A deep learning model for unusual activity detection in examination hall was presented by Devi et al. [5].

They employed a motion based approach to extract significant frames from the video. These frames were fed into a CNN for classification purposes. The effectiveness of the system was evaluated using database related to exam hall scenarios. Majd et al. [6] combined CNN- Long

Short Term Memory (CNN-LSTM) to improve classification accuracy. However, this model takes up a lot of memory.

Beddiar et al. [7] implemented the used on Improved Fish Vectors (IFV) for the detection of suspicious activities. Their study aimed to improve the accuracy of identifying unusual behaviors. Xu [8] presented a video surveillance system using Faster R-CNN. Input video sequences were subjected to enhancement to improve quality of the videos. Faster R-CNN was trained to categorize the images into normal and suspicious. Violence detection system for football stadium using bidirectional LSTM (BiLSTM) detects unusual activity in crowd [9]. The system accepted videos as input. The videos were transformed to non-overlapping frames. Features were computed and then BiLSTM was trained to detect violent behavior.

A hybrid model by combining CNN and auto encoders for anomaly detection [10]. Auto encoder has encoder and decoder part. The encoder part includes convolutional and pooling layers whereas decoder part has deconvolutional and pooling layers. Sreenu and durai [11] presented a detailed survey on video surveillance systems using deep learning models. Authors deeply rooted review, which starts from object segmentation, detection, recognition and violent detection in crowded environment. Mehmood [12] detected unusual activity in crowd using pretrained 2D CNN. This method employed a pretrained 2d CNN for motion information and high classification rate.

3. PROPOSED SYSTEM

The block diagram of the introduced systems is given in Figure. 1. The key elements of the system are data collection, preprocessing, key frame extraction, augmentation, segmentation, and classification.

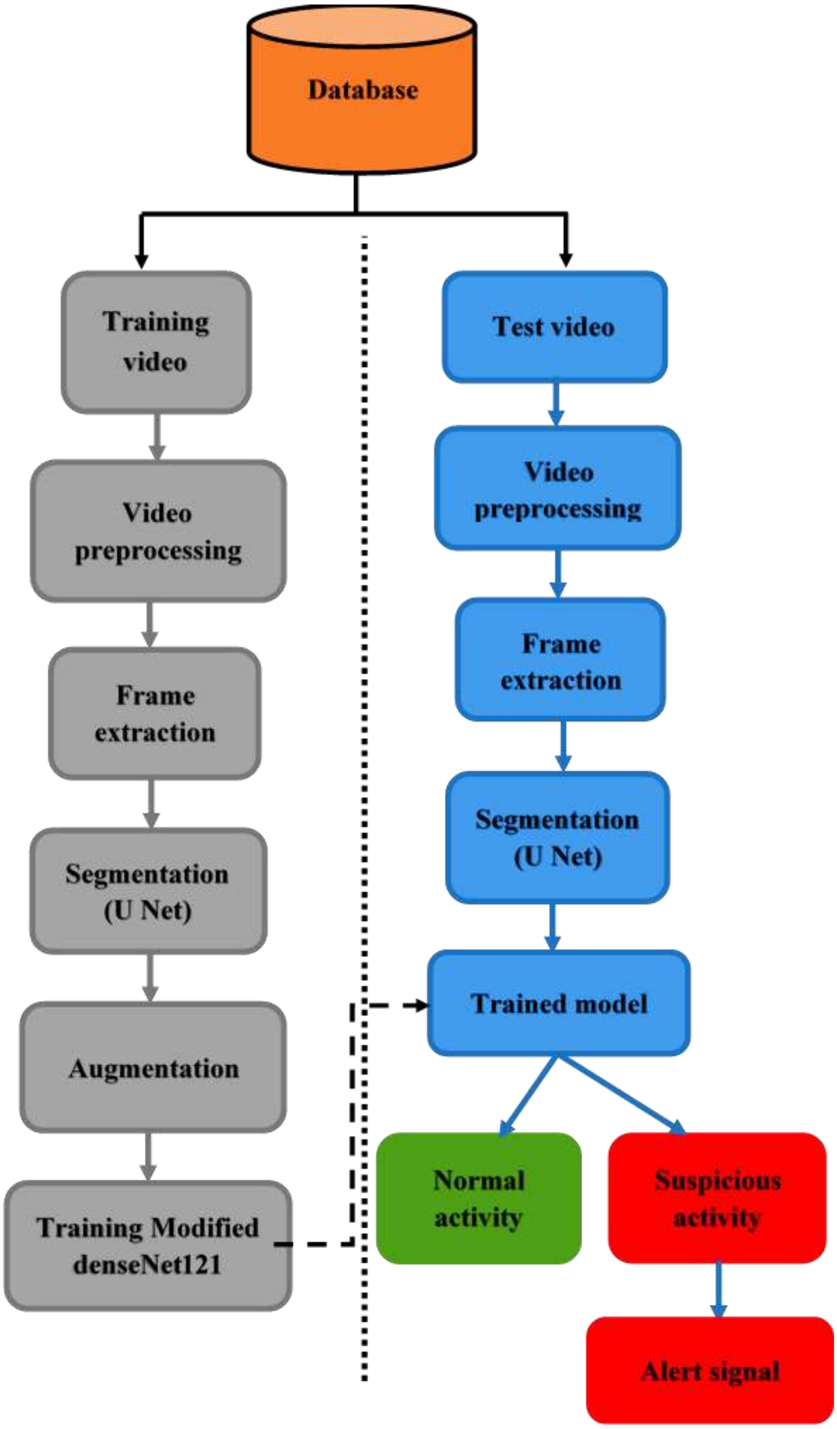

Figure.1 Workflow of the developed system

3.1 Data gathering

The objective of this work is to achieve a high level of accuracy in detecting abnormal activity within an examination hall. Additionally, the work aims to detect unusual behavior in crowded scenes to reduce the crime rate. To accomplish this, two datasets have been utilized for analysis:

- The Examination Unusual Behaviour (EUB) dataset, comprising 550 videos. Among these videos, 145 depict normal activities while 405 depict abnormal activities. All videos in this dataset are in the .mp4 format. Figure.2 presents sample frames from both normal and abnormal within the EUB dataset.

- The crowd Violence/non-violence database [2][5][6], containing 1000 normal and 1000 violence videos. Sample frames from this database are Figure.3, illustrating instances of crowd violence.

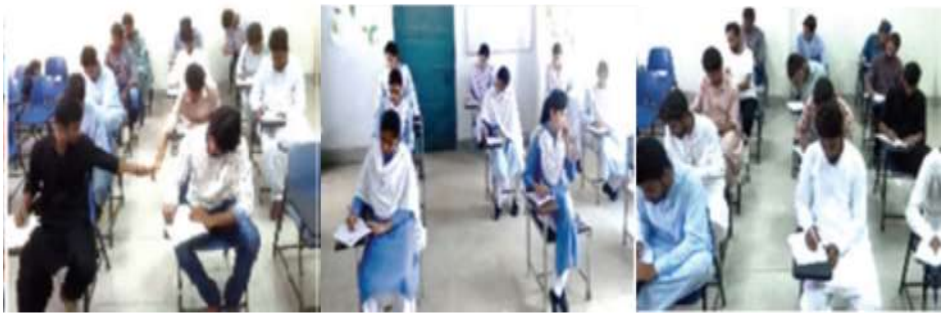

Figure.2 Sample frames from the EUB database

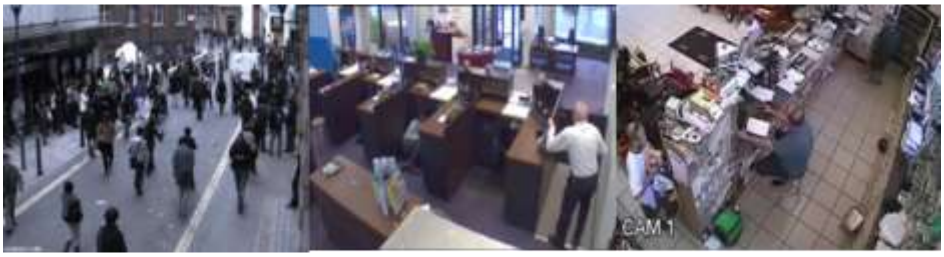

Figure.3 Sample frames from the crowd violence/non-violence database

3.2 Preprocessing

The mean filter is employed to reduce noise, while adaptive histogram equalization is implemented on frames to enhance their contrast. All processed frames are resized to 128x128 to ensure their suitability for subsequent analysis.

3.3 Key frame extraction

The absolute difference of histogram is a two-step process where the first step calculates a threshold by using mean and standard deviation of the histogram obtained from the absolute difference between consecutive image frames. In the second step, key frames are extracted by comparing the threshold with the absolute difference of consecutive image frames. This algorithm begins by sequentially extracting video frames. Initially, each frame histogram difference between two consecutive frames are computed.

At first, the histogram difference between two consecutive frames is computed. Then, the average and standard deviation of the absolute difference of the histograms are determined in order to establish a threshold.

Let the frames, f, video, v= {f1, f2, …..fi….fN} i= 1, 2, 3,…..N. the absolute difference between two consecutive frames, fi and fi+1 is expressed as,

$$d_i = |f_i - f_{i+1}| \qquad (1)$$

The mean, μ and standard deviation, σ of absolute difference is computed. Threshold can be defined as,

$$\text{Threshold}, T = \mu + \sigma$$

The Key Frame (KF) can be extracted as follows,

$$KF = \begin{cases} \text{Keyframe} & d < T \\ \text{Not a key frame} & d > T \end{cases} \qquad (2)$$

3.4 Segmentation

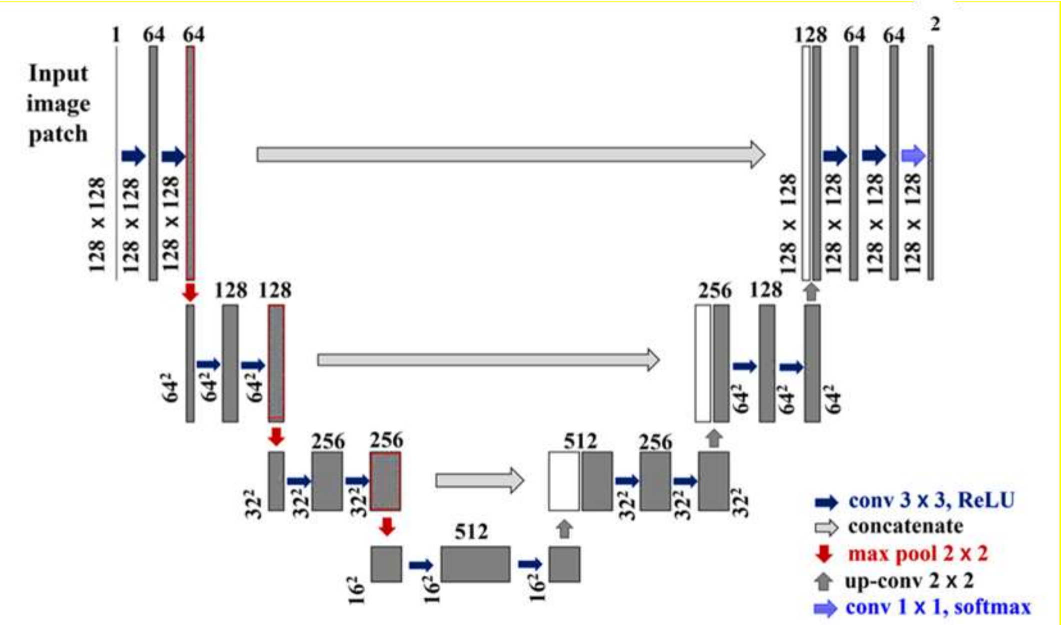

Figure.4 Architecture of U-Net

Figure.4 illustrates the structure of the U-net for segmentation. The input frame undergoes a sequence of operations. First, it passes through consecutive 3X 3 convolutional layers, both followed by Rectified Linear Unit (ReLU) activations. Subsequently, a 2X2 max-pooling layer operation with stride of 2 is employed for down sampling purposes. Following each down sampling step, the quantity of feature maps is doubled. This iterative process continues until the feature map reaches a dimension of 16x16 pixels, forming the contracting pathway of the network. Transitioning to the expansive pathway, the feature maps are up-sampled then subjected to a 2x2 convolution that reduced the feature channel count by half. This is followed by combining the corresponding feature map from the contracting path and applying two 3x3 convolutions, each followed by a ReLU activation function. The final layer involves mapping each 64 feature vector to the requisite number of classes through employment of a 1x1 convolution operation.

3.5 Classification

In this work, modified DenseNet201 is proposed to classify the images. The DenseNet has 201 layers. It starts with a convolutional layer and pooling layer. This is followed by a dense block, then a transition layer, another dense block pursued by a transition layer, one denser layer coupled with a transition layer, and ultimately, a closing dense block followed by fully connected layer and classification layer. In this investigation, conventional softmax function in the classification layer is replaced with WNN. The dimension of various layers of modified DenseNet201 is listed in Table.1. Structure of the developed Modified DenseNet 201 is given in Figure.4.

Table.1 Hyperparameters of the Modified DenseNet201

Layer	Kernel/stride
Input layer	
Conv1	7X7/2
Poo1	3X3/2
Dense block 1	$\begin{bmatrix} 1X1 \\ 3X3 \end{bmatrix}$ X 6
Transition layer 1	1x1, convolution 2x2,pooling /2
Dense block 2	$\begin{bmatrix} 1X1 \\ 3X3 \end{bmatrix}$ X 1
Transition layer 2	1x1, convolution 2x2,pooling /2
Dense block 3	$\begin{bmatrix} 1X1 \\ 3X3 \end{bmatrix}$ X 48
Transition layer 3	1x1, convolution 2x2,pooling /2
Dense layer 4	$\begin{bmatrix} 1X1 \\ 3X3 \end{bmatrix}$ X 32
Pooling layer	7X7/1 global average
Fully connected layer	1000
Classification layer	WNN

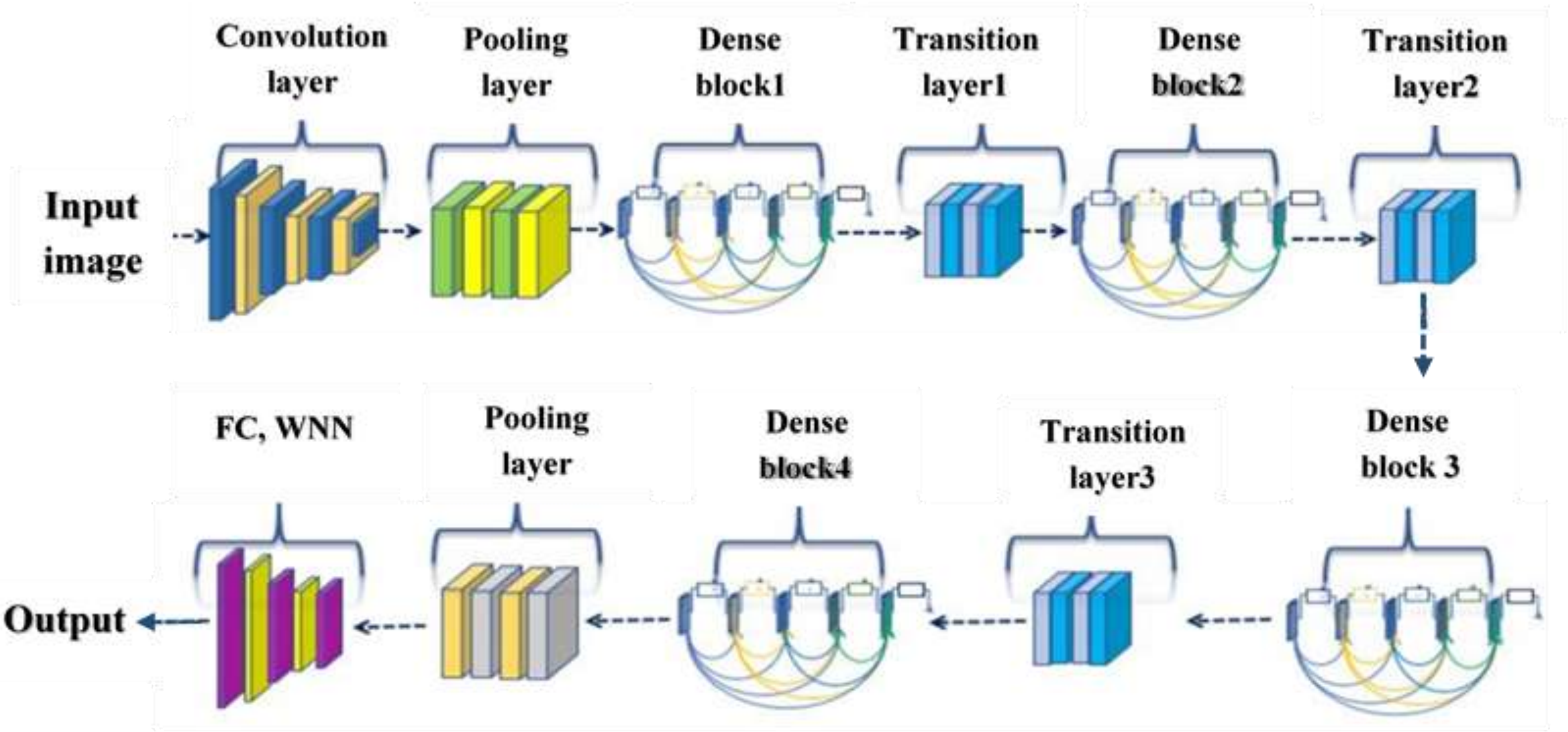

Figure.4 Structure of the modified DenseNet 201

4. SIMULATION RESULTS

4.1 Evaluation metrics

A research investigation was conducted to examine the efficacy of an introduced system by conducting experiments with two datasets. The experiments evaluated the effectiveness of different methods based on these datasets, examining four parameters including classification accuracy (ACC), Recall (R), Precision (P), and F1-score. Table. 2 lists the metrics used for assessment.

Table.2 Performance metrics

Metrics	**Equation**
Classification accuracy (ACC)	$\mathrm{ACC} = \frac{\mathrm{a+b}}{\mathrm{a+b+c+d}}$
Recall (R)	$\mathrm{R} = \frac{\mathrm{a}}{\mathrm{a+d}}$
Precision (P)	$\mathrm{P} = \frac{\mathrm{a}}{\mathrm{a+c}}$
F1-score	$\mathrm{F1 - score} = 2\ \mathrm{X}\ \frac{\mathrm{P\ X\,R}}{\mathrm{P+R}}$
Where, a-abnormal activity, b-normal activity, c-False abnormal, and d-False normal	

4.2 Results on EVB dataset

As per the proposed system, initially, all the datasets were preprocessed using data augmentation techniques to increase the database size as well as boost classification rate. After augmentation, the was trained with training samples. 10-cross fold validation tactic was used to assess the performance. The proposed system was tested using MATLAB 2019a. The proposed system was trained and tested using both augmented and non-augmented data. The tuning of hyper parameters is essential for training a deep learning model. drop out, learning rate, batch size and optimized were set to as 0.4, 0.001, 64 and Adamoptimizer.

Table.3 exhibits the outcomes of the introduced system assessed on the EVB dataset. Table.3 demonstrates that the evaluation results of metrics indicate an enhancement in the proposed system performance when augmented data was used, as compared to the system without augmented data. In Table.23, with respect to ACC, it is noted that the proposed system possessed value of 89.27% for non-augmented data, and the system with augmented data resulted in achieving a value of 99.44%, showing its effectiveness. On observing other metrics such as P, R and F1-score, the proposed system evolved results better than the system without augmentation data. Regarding both the original and augmented data, the proposed system attained P values of 91.59% and 99.33% R values of 94.07% and 99.56% and F1-score values of 92.81% and 99.45%, respectively. Figure.5 shows the Receiver Operating Characteristic (ROC) curve. The graph shows the superiority of the proposed with augmented data by reaching Area Under Curve (AUC) of 0.994.

Table.3 Activity detection efficacy on the EUB dataset

Data	ACC (%)	P(%)	R (%)	F1-score (%)	AUC
Data without augmentation	89.27	91.59	94.07	92.81	0.850
Data with augmentation	99.44	99.33	99.56	99.45	0.994

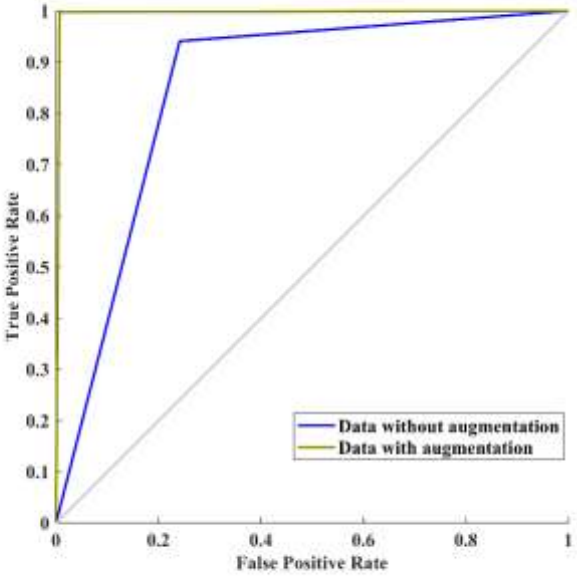

Figure. 5 ROC curve for the developed system on the EUB data

4.3 Results on violence/non-violence data

In this experiment, the proposed system was employed to detect unusual activity in crowd scenes. The developed system was trained and tested using both non-augmented and augmented data. Efficacy of the system is reported in Table.4. The outcomes reveal that for non-augmented data, the designed system achieved a classification rate of 92.67%, while for augmented data, the accuracy improved significantly to 99.40%, indicating a performance improvement of 6.73%. Furthermore, for original and augmented, the designed system achieved P, R and F1-score values of 96.35% and 99.20%, 92.50% and 99.60%, and 94.39% and 99.40%, respectively. Empirical findings showed superior outcomes for violence/ non-violence data among all metrics. The ROC curve for violence/ non-violence data is shown in Figure.6, which demonstrates an AUC of 0.928 for non-augmented data and 0.994 for augmented data.

Table.4 Activity detection efficacy on the violence/non-violence dataset

Data	ACC (%)	P(%)	R (%)	F1-score (%)	AUC
Data without augmentation	92.67	96.35	92.50	94.39	0.928

Data with augmentation	99.40	99.20	99.60	99.40	0.994

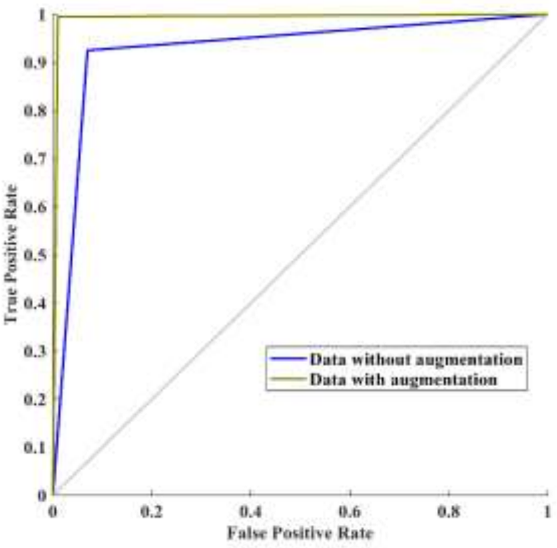

Figure.6 ROC curve for the developed system on the violence/non-violence data

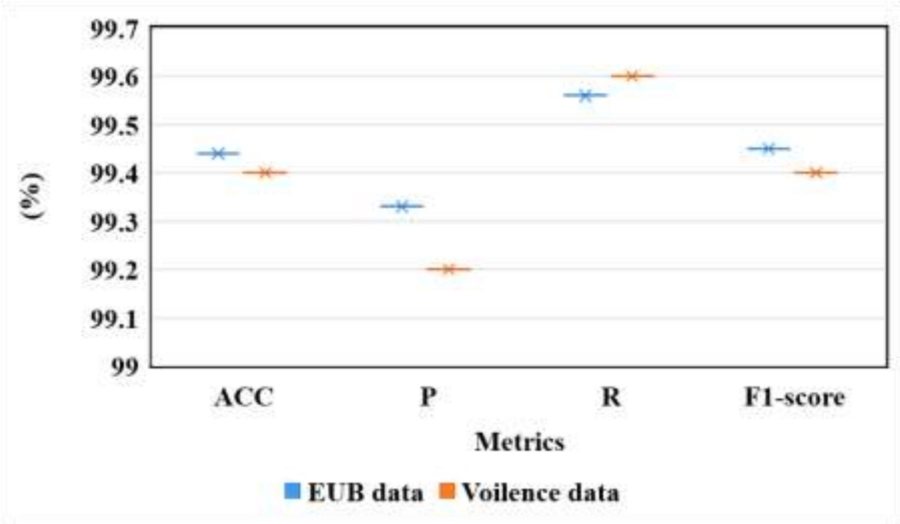

Figure.7 Effectiveness of the developed system

Figure.7 compares and illustrates the effectiveness of the developed system in detecting suspicious activity in crowded environments by analyzing its generalization capability. The system's performance was evaluated using the EUB and violence/non-violence data sets. The outcomes showed that the introduced system performed well for both datasets, achieving high values for most metrics. Figure. 7 provides a clear visual representation of these findings, indicating that the developed system can effectively detect suspicious activity in crowded environments.

4.4 Comparison with other approaches

Comparison on the EUB data

To further evaluate the developed system for detection of abnormal behaviors in crowded place and examination hall, a comparison was done with earlier approaches. Figure. 8 illustrates the effectives of the introduced system in comparison to the earlier approaches such as AUAR [6], Semi-supervised method, ConvLSTM, and CNN-LSTM.As depicted in Figure.8, the proposed system exhibited outstanding performance across all metrics when compared to the earlier systems taken for comparison. On observing recognition accuracy, AUAR, semi-supervised, ConvLSTM, and CNN-LSTM models attained the values f 77%, 73.5%,78%, and 83%, respectively. The proposed model poses the accuracy value of 99.44%, proving its effectiveness.

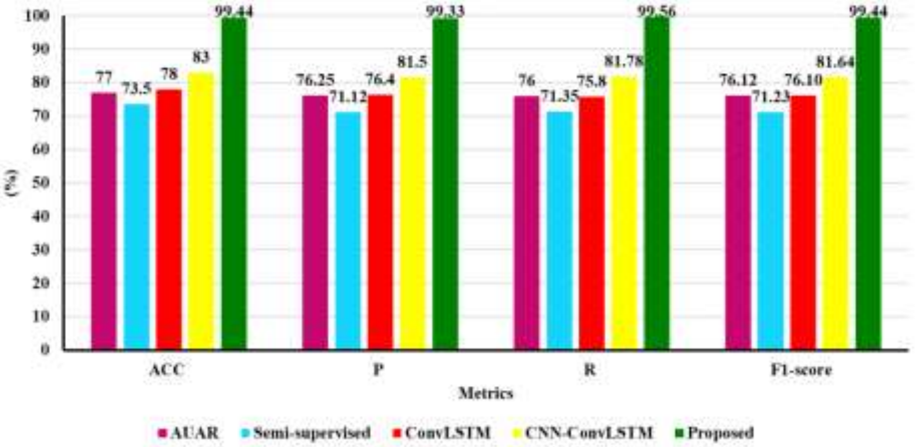

Figure.8 Comparison between the proposed and the earlier methods

Comparison on the violence/non-violence data

The graphical representation in Figure.9 compares the results of the developed system with four earlier methods, namely AUAR [5], SD [2], CNN-LSTM [6], and IFV [7], Semi-supervised method, ConvLSTM, and CNN-LSTM. The developed system surpasses the performance of the earlier methods, achieving the highest accuracy of 99.40%. in this way, P, R and F1-score values of developed system are better than the earlier methods. Including Mohammadi et al. [2]. (91.16%), Majd et al. [6] (96.74%), Beddiar et al. [7] (97.29%), Oh et al. [13] (93.25%). Similarly, the designed system achieved higher R and F1-score values as compared to the earlier methods, IFV (97.53% and 97.41%), CNN-LSTM (95.31% and 96.02%), SD (89.14% and 90.14%), AUAR (98.52% and 98.03%), Semi-supervised (91.2% and 92.21%), ConvLSTM (91.3% and 90.87%), and CNN-LSTM (94.99% and 95.33%).

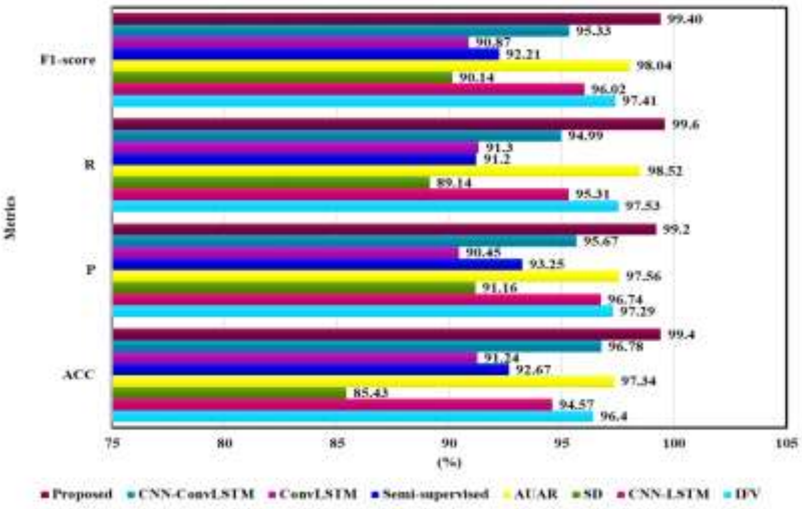

Figure.9 Performance comparison of different methods

5. CONCLUSION

In recent years, there has been significant research on suspicious activity recognition, particularly in crowed environments and examination halls. The timely and automated identification of suspicious activities plays a crucial role in enabling authorized person to respond accurately and fairly. The input video sequences undergo preprocessing, followed by the extraction of key frames. These computed frames were than employed as input for the developed system to differentiate between normal and suspicious activity. In case of any abnormal behavior or presence of a weapon is detected, the designed system triggers an alert signal to notify the authorized person or security guard. The performance of the system was validated on both the EVB data and violence data. The result indicated that the system outperforms other models when applied to these two datasets, delivering excellent outcomes.

REFERENCES

1. Jalal, A., Mahmood, M. and Hasan, A.S. (2019) Multi-features descriptors for human activity tracking and recognition in Indoor-outdoor environments. In: 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST),371-376.
2. Mohammadi, S., Kiani, H., Perina,A. and Murino,V.(2015)Violence detection in crowded scenes using substantial derivative. 12th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, Karlsruhe, Germany,1–6.
3. Jordan, A.E. (2001) College student cheating: The role of motivation, perceived norms, attitudes, and knowledge of institutional policy. Ethics Behav. 11, 233-247.
4. Miller, A.D., Murdock, T.B. and Grotewiel, M.M. (2017) Addressing academic dishonesty among the highest achievers. Theory into Pract. 56, 121-128.
5. Devi, S., Suvarna,G. and Chandini,S. (2017) Automated video surveillance system for detection of suspicious activities during academic offline examination. Int. J. of Compu. and Inform. Engg.,11(12),1265–1271.
6. Majd, M. and Safabakhsh, R. (2020) Correlational convolutional LSTM for human action recognition. Neurocomputing,396(1),224-229.

7. Beddiar,D. R., Nini,B., Sabokrou.M. and Hadid, A. (2020) Vision-based human activity recognition: A survey. Multimedia Tools and Applications, 79, 30509-30555.
8. Xu,J (2021) A deep learning approach to building an intelligent video surveillance system, Multimedia Tools and Applns., 80,5495-5515.
9. Samuel, R.D.J., Fenil, E., Manogaran, G., Vivekananda, G.N., Thanjaivadivel, T, Jeeva, S. and Ahilan, A (2019) Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM. Comput Netw., 151,191-200
10. Ribeiro, M, Lazzaretti, A.E., and Lopes, H.S (2018) A study of deep convolutional auto-encoders for anomaly detection in videos. Pattern Recogn. Lett.,105:13-22
11. Sreenu, G. A and Durai, M.A.S (2019) Intelligent video surveillance: a review through deep learning techniques for crowd analysis, J. Big Data, 6(48), 1-27
12. Mehmood, A (2021) Efficient anomaly detection in crowd videos using pre-trained 2D convolutional neural networks, IEEE Access, 9, 138283-138295.
13. Oh S, Ashiquzzaman A, Lee D, Kim Y, and Kim J. (2021) Study on Human Activity Recognition Using Semi-Supervised Active Transfer Learning. Sensors (Basel). 21(8):2760