Electronic Letters on Computer Vision and Image Analysis 24(1):31-50; 2025

Supervised Deep Learning Approaches for Anomaly Detection and Recognition in Crowd Scenes

Kinjal V Joshi* and Narendra M Patel+

*The Charutar Vidya Mandal University, Vallabh Vidyanagar, Gujarat, India
+Birla Vishvakarma Mahavidyalaya, Vallabh Vidyanagar, Gujarat, India
Received 9th of December 2022; accepted 9th of January 2025

Abstract

These days consciousness about public safety increases and Closed-Circuit Television (CCTV) cameras are installed at almost all public places. In general, automated smart surveillance systems are not commonly available, and most surveillance videos are monitored manually. This study emphasizes the automatic detection and classification of abnormal events in surveillance video especially in crowd environments. Abnormal event detection is a challenging task because the definition of abnormality is subjective. In the surveillance video with a dense crowd, automatic anomaly detection becomes very difficult because of clutter and severe occlusion. This research represents Convolutional Neural Network (CNN) and Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) based approaches for detection and classification of abnormal events. The CNN architecture is developed from scratch and used for spatial domains. LSTM architecture is developed for the temporal domain. Feature sequences are generated using CNN model and given as input to LSTM model. Experiments are carried out using five different publicly available benchmark datasets. The performance is measured by accuracy and Area Under the ROC (Receiver Operating Characteristic) Curve (AUC). The CNN-LSTM approach performs better than the CNN.

Key Words: Abnormal Event Detection, Abnormal Event Classification, CNN, LSTM

1 Introduction

With the increasing need to secure people and personal property, video surveillance has become a major concern in daily life. The increasing demand has led to the widespread installation of CCTV cameras to generate video footage. Most existing video surveillance systems are entirely supervised by humans. Video monitoring is a challenging and labor-intensive task, and it is difficult for humans to identify abnormal events in large video files. However, even a small mistake can lead to unacceptable consequences. Thus, it is essential to develop a system dealing with many video frames and alert people for a punctual and functional response when an abnormal event occurs. So, considerable research on automatic video surveillance is going on. The major applications of automatic abnormal event detection and recognition in surveillance scenes are building security, traffic analysis, video monitoring etc. Because of usefulness and complexity, currently, it is an open research area and many real-world benchmark datasets are publicly available for research. Automatic

Correspondence to: <kinjaljoshi@gcet.ac.in>
Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.1631
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

abnormal event detection is a challenging task because of the subjective definition of abnormality. Whenever a crowd environment is considered for a surveillance scene, it becomes more difficult because of clutter and occlusion. Research [1-3] represents that abnormal event detection in a crowd scene is one of the four modules of Crowd Analysis and 30% work is done in this area from the absolute work done in the Crowd Analysis domain.

Most of the researchers [42-46] have used unsupervised learning-based approaches. For an unsupervised learning-based approach, only Normal event videos are used in training. If any abnormal event exists in the test videos, it can be detected based on a logic that if it differs from the training video, it contains an abnormal event. For this task different researchers have used various autoencoders like convolutional, variational, two-stream, spatiotemporal etc., as autoencoder regenerates the input. Major drawback of this kind of approach is that any new normal event in test data, which is not available in the training set, can also be considered abnormal.

To use supervised learning-based approach labeled data is required. It is impossible to generate labels of all types of abnormal behaviors, so general-purpose abnormality detection may not be possible using supervised learning. However, as per requirements at a particular place, it is possible to generate labels of abnormal events and supervised learning can be used. For example, only pedestrians are allowed in some areas, so vehicle entry is abnormal. Access without making a payment is considered deviant behavior at some paid entry points. Robbery can be viewed as an abnormal event at some places like banks or shops. For surveillance cameras positioned on the road, the events like suddenly running, fighting, accident, crowd formation can be considered abnormal. In an examination hall, talking or copying the other student's answer sheet is abnormal behavior. So as per the situation, it is possible to generate labels of abnormal events and supervised learning can be used. Here supervised learning-based approaches utilizing CNN and CNN-LSTM are proposed, as both architectures have shown the most promising results in various applications.

The major contributions of the paper are as follows:

- Detailed study of various techniques for abnormal event detection and recognition in a crowd scene.
- Designed two supervised deep learning architectures for abnormal event detection and recognition in spatial domain and spatiotemporal domain.
- Conducted exhaustive experiments on various types of datasets to test and validate the proposed approaches.

The rest of the paper is organized as follows. In section 2 related findings are investigated to propose efficient methods for abnormal event detection and recognition. Section 3 represents dataset description, proposed approaches and experimental results. Section 4 represents comparison between the proposed approaches and comparison of both with state-of-the-art methods using specific datasets. Finally, the conclusion and future work are indicated in section 5.

2 Related Work

As the definition of an abnormal event is subjective, machine learning and deep learning algorithms can be applied for a particular event on a specific dataset. Detecting all types of abnormalities in video surveillance is highly challenging, which is why extensive research is being conducted in this field. To identify abnormalities, it is essential to extract features from the images. Feature descriptors can be classified as handcrafted features and deep learning-based features. Handcrafted features relate to properties derived using various algorithms using the information present in the image itself like edge, corner, Histogram of Oriented Gradients (HOG), Optical flow etc., [16][21]. These features are extracted and given as input to any supervised or unsupervised machine learning algorithm. Deep learning-based features are the feature descriptors extracted by the deep learning

model. The images are given as input, and the deep learning model automatically extracts features as per tuned parameters and hyperparameters. Recent research work is reported with deep learning-based approaches, so existing methods based on supervised and unsupervised deep learning are represented in this section. The proposed models give better results compared to all the methods mentioned in this section.

Mostafa, T. et al. [8] have analyzed local spatial-temporal motion patterns in video frames. They have used a motion heat map to find the region of interest. After identifying the motion structure, different classifiers are used like Support Vector Machine (SVM), Naive Bayes, Neural Network and CNN. The CNN classifier gives a good result to detect anomalies in a crowd scene. The achieved accuracy is 96.74% and 94.28% on UMN and UCSD datasets, respectively. Ravanbakhsh, M. et al. [10] employed a Fully Convolutional Network as a pre-trained model and plugged an effective binary quantization layer as the final layer to the net. The temporal CNN patterns were captured to denote motion in a crowd. The achieved AUC values are 0.95, 0.88 and 0.98 for UCSDPed1, UCSDPed2 and UMN datasets respectively.

Sabokrou M. et al. [11] have used transfer learning. The authors have used pre-trained CNN, i.e. Alexnet. They have received AUC values 0.904 and 0.902 for Subway Entrance and Subway Exit datasets, respectively. Feng Y. et al. [12] have extracted appearance and motion features using the PCANet. The deep Gaussian Mixture Model (GMM) is constructed with observed regular events. The authors have got a 0.925 AUC value for the UCSDPed1 dataset. Zhou, S. et al. [13] have used a Spatial- temporal CNN to detect features like appearance and motion from spatial and temporal dimensions. The achieved AUC values are 0.9963 and 0.927 using UMN and Subway Entrance datasets, respectively.

Smeureanu, S. et al. [14] have used the VGG pre-trained CNN model to extract features, and then one class Support Vector Machine (SVM) classifier is used to learn normal event patterns. They have got 0.85 AUC value for the UMN dataset. Sun, J. et al. [15], the authors have proposed the Deep One Class (DOC) model in which CNN is used to extract features, and One-class SVM is used to learn decision function for abnormal event detection from the given normal event images. They have done experiments using SVM with linear kernel and RBF kernel. With linear kernel, the achieved AUC value is 0.808, and for RBF kernel, it is 0.914 using the UCSDPed1 dataset. Hinami, R. et al. [17] have used a fast R-CNN model to detect an abnormal event. R-CNN is a Region-based Convolutional Neural Network in which region proposals are used during training and testing, so algorithm's performance slows down significantly. The achieved AUC values are 0.89 and 0.92 for Avenue and UCSDPed2 datasets respectively.

Yan, S. et al [18], authors have proposed a two-stream R-convolutional Variational Autoencoder (RconvVAE). They have done experiments with unsupervised learning-based approaches like a convolutional autoencoder, convolutional variational autoencoder, recurrent convolutional autoencoder, recurrent convolutional variational autoencoder. They have used appearance and motion features to describe the probabilistic distribution. The achieved AUC value using two-stream R- ConvVAE is 0.75 for UCSDPed1 dataset. Chong, Y. et al. [19] proposed a spatiotemporal autoencoder that includes two main components, one for spatial feature representation and the other for learning the temporal evolution of the spatial features. The achieved AUC values are 0.899 and 0.847 for UCSDPed1 and Subway Entrance datasets, respectively. Vu, H. et al. [20] have used denoising Autoencoder and Conditional Generative Adversarial Networks to detect an anomaly from video. They got AUC values 0.82, 0.99 and 0.71 for UCSDPed1, UCSDPed2 and Avenue datasets respectively.

Sultani, W. et al. [23] have developed a UCFCrime challenging multiclass dataset and experimented with two existing approaches for events classification. In the first one, they have used a 3D convolutional network for feature extraction and the nearest neighbor classifier for classification. The

achieved accuracy for classification is only 23%. In the second one, they have used Tube Convolutional Neural Network (TCNN), which includes a tube of interest pooling layer. It combines all clips' features and generates one feature vector for one video. The reported classification accuracy is 28.4%. Landi, F. et al. [9] have used a 3D convolutional network with a regression network to find out anomaly scores. They have done two experiments. In the first one, they have used the whole frame as input. In the second experiment, the authors extracted a spatiotemporal tube that locates abnormal events and is given as input. Here UCFCrime2Local dataset is used. For the whole frame, the obtained AUC value is 0.5612, and for the spatiotemporal tube, the obtained AUC value is 0.7413.

For unsupervised learning, labels of abnormal events are not used. Autoencoders are trained using only normal events. Test video frames are given as input to the autoencoder, and reconstructed images are compared with original images. If reconstruction error is greater than the threshold, then it is considered that the image contains an abnormal event. Especially, whenever scenes are of surveillance video, they can have lots of activities. To train autoencoder with all kinds of normal events is not possible. This matter affects the performance of the system. Hence, supervised learning-based approaches are proposed in this study.

3 Material and Methods

In this research two supervised deep learning-based methods are proposed to detect and classify abnormal events in surveillance scenes. One is using CNN for spatial domain and the other is using CNN-LSTM for spatiotemporal domain.

3.1 Dataset Description

The main objective of this research is to propose the efficient supervised deep learning architecture for abnormal event detection and classification. Here, Experiments are done using five benchmark publicly available datasets namely UMN, UCSDPed1, Violent Flows, Subway Entrance and UCFCrime2Local. The abnormal events like running, vehicle entry in restricted areas, avoidance of payment at entry gate, violent behavior of crowd, arrest, assault, burglary, robbery, stealing, and vandalism are detected. In all these datasets, ground truth and enough abnormal images are available. The UMN [4] dataset is developed by University of Minnesota. It contains two outdoor and one indoor video samples with 320×240 pixels image resolution. Each video starts with a segment of walking, signifying the normal state, and ends with sequences of running, representing the abnormal state.

The UCSDPed1 [5] is developed by University of California and San Diego. It is a real-world dataset that contains 34 training video samples and 36 testing video samples. Both train and test samples have 200 frames of dimension 238 x158. In this dataset, pedestrians walking on the walkway is considered normal behavior. Commonly occurring anomalies include bikers, skaters, small carts, and people walking across a walkway or in the grass surrounding it.

Violent Flows dataset [6] is real-world video footage of crowd violence along with standard benchmark protocols. It mainly presents the crowd violence behavior, and most of the scenes are dynamic, which significantly increases the detection difficulty. This dataset contains 246 video clips with 123 violent samples and 123 nonviolent ones. The resolution for video frames is 320 × 240.

The real-world dataset Subway is provided by A. Adam et al. [7]. This dataset contains two videos. One video monitors the entrance gate, which is 1 hour 36 minutes long and the second monitors the exit gate, which is 43 minutes long. In this research work, the video sample of the entrance gate is used, which contains 384 × 512 pixels image resolution. People moving in the wrong direction and avoiding payment at the entry gate are abnormal events. The number of anomalies is less in this dataset.

The UCFCrime2Local dataset [9] is subset of UCFCrime dataset. It contains 300 videos of seven different categories: Arrest, Assault, Burglary, Normal, Robbery, Stealing, and Vandalism. From these 300 videos, 200 videos are of regular activities and 100 videos are of six different categories' abnormal events. This dataset contains a train-test split. The training set includes 141 normal event videos and 69 abnormal event videos. The test set contains 59 normal event videos and 31 abnormal event videos. This dataset contains weakly labeled videos. Video level labeling is available, i.e., video is normal or has an anomaly somewhere.

Normal: Walking | Abnormal: Running | Normal: Pedestrian | Abnormal: Vehicle

UMN **UCSDPed1**

Normal: Nonviolent Crowd | Abnormal: Violent crowd | Normal: Making payment at entry | Abnormal: Avoiding payment at entry

Violent Flows **Subway Entrance**

Figure 1 Normal and abnormal sample images of various datasets

Arrest | Assault | Burglary | Normal

Robbery | Stealing | Vandalism

Figure 2 Sample images of UCFCrime2Local dataset

Except for UCFCrime2Local dataset, all datasets contain only two categories Normal and Abnormal. The UMN, UCSDPed1, Violent Flows and Subway Entrance datasets are only used for anomaly detection because they do not contain various categories of abnormalities. While UCFCrime2Local dataset contains six different kinds of abnormalities like Arrest, Assault, Burglary, Robbery, Stealing, and Vandalism and one category i.e. Normal surveillance videos. The dataset is large enough to work with deep learning approach, so it is used for detection as well as for recognition of abnormal events.

Because of various categories of anomalies, it fulfills the goal of the study of recognition of abnormal events. Figure 1 represents the normal event image and the abnormal event image of various datasets. Figure 2 represents the images of seven different categories of the UCFCrime2Local dataset. The implementation is done with MATLAB-2019a, 16 GB RAM, i9 Processor CPU machine.

3.2 CNN Based Approach

3.2.1 Convolutional Neural Network

A Convolutional Neural Network [36][38] is a deep learning algorithm that takes the image as an input and extracts features from it using learnable weights and biases. It can be used for object detection and classification task. It contains the sequence of various layers with tuned parameters and hyperparameters.

Transfer learning is typically used in Computer Vision and Natural Language Processing tasks. Pretrained networks are trained using large datasets like the ImageNet, which consists of more than a million images classified into many classes. The various pre-trained CNN architectures like LeNet, AlexNet, VGGNet, GoogLeNet, ResNet, ZFNet etc. are used by different researchers for transfer learning. The transfer learning approach using a pre-trained network is a significantly faster and easy way of training. Although it has achieved good results in many cases, it is not appropriate for all applications because it is hard to know how much the previous training process helps. In addition, only retraining the last few layers cannot guarantee the best results because categories for source training data are generally different and the semantic representations in the higher layer are relative to the training categories. Therefore, using the highly specific semantic representation in the new recognition task is not worthy. In this research, the pre-trained image classification network GoogLeNet is retrained and evaluated using UMN dataset. The achieved accuracy is 86.21% and AUC value is 0.9487. As the achieved performance was not sufficient, so a new architecture is developed from scratch.

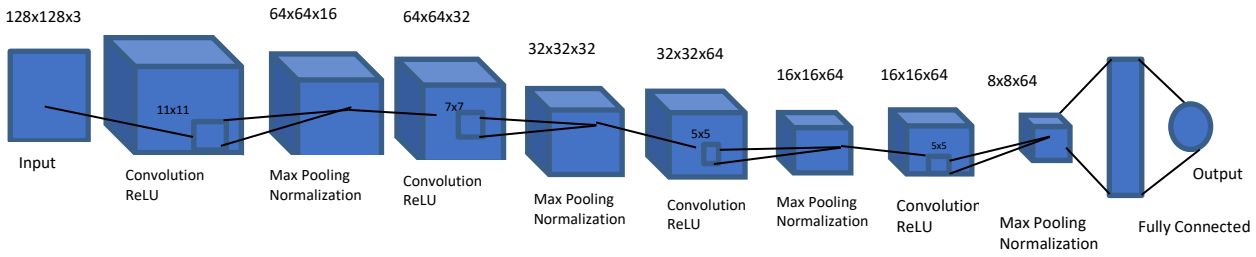


Figure 3 The proposed CNN architecture

To develop a new model, the experiments were started with small architecture, and by deep analysis of the performance, the final model is proposed. Figure 3 represents the proposed CNN Architecture. In this model, twenty layers are used. The size of the input image is taken 128x128. For color images, the input layer size is 128x128x3, and for grayscale images, the size is 128x128x1. As preprocessing all images are resized to 128x128 as per requirements.

In the proposed model, four convolution layers are used. For each convolution layer, zero padding is set such that the output size of the image remains the same as the input size. Biases and weights are initialized with the Glorot initializer [35] for each convolution layer. For initial convolution operation, fewer large size filters are used to extract coarse details. Then sequentially, the filter size is decreased, and the number of filters is increased to extract fine details. Initially, 16 filters of size 11x11 are taken for the first convolution layer. ReLU is used as an activation function. After applying the activation function, the max-pooling layer is used, which summarizes the maximum presence of a feature. The max pooling layer is used with pool size 2 and stride 2, So now the image size is 64x64.

During the study of CNN, it was found that few researchers [28] use batch normalization before the non-linearity, i.e., ReLU layer, and few researchers reported that adding batch normalization after the non-linearity improves accuracy [26, 27]. Here experiments are done using both architectures, and almost identical results are obtained using all datasets. So, the batch normalization layer is used after the max pooling layer in the proposed architecture. The same process is repeated three times with 32 filters of size 7x7, 64 filters of size 5x5 and again 64 filters of size 5x5, respectively.

Different researchers have various opinions about the use of the dropout layer. In [27], authors have reported that accuracy is decreased significantly by using the dropout layer. It should be used carefully to design CNN, as it cannot give better results in each application. At the same time, [29] shows dropout helps accuracy. As dropout is the one way to remove overfitting of the model, in this work, various experiments with the proposed architecture are done like (1) No dropout, (2) single dropout before fully connected layer and (3) one dropout layer after each batch normalization layer. Significant variation is not found in all three cases while doing the experiments using the described datasets. As batch normalization fulfils the goal of using dropout, it is not included in the proposed architecture.

3.2.2 Results and Discussion

CNN architecture extracts all important features from the images. Figure 4, Figure 5, Figure 6, and Figure 7 show the largest activation at four convolution layers, RELU layers, Max pooling layers and Batch normalization layers, respectively for the sample image of the arrest event of UCFCrime2Local dataset.

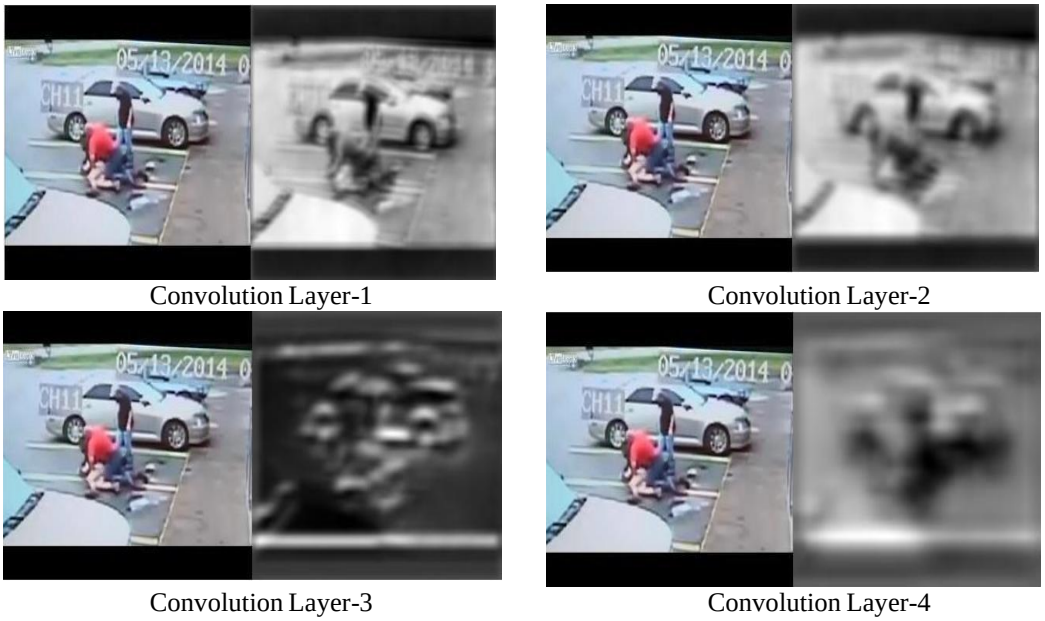


Figure 4 The largest activation at each convolution layer for the arrest event

For abnormal event detection using the UCFCrime2Local dataset, all images of various categories except for ‘Normal’ are considered abnormal event images. Here experiments are done on UMN, UCSDPed1, Violent Flows and Subway Entrance datasets using 80% training data and 20% test data. While for the UCFCrime2Local dataset, the experiment is performed as per the given train-test split.

ReLU Layer-1

ReLU Layer-2

ReLU Layer-3

ReLU Layer-4

Figure 5 The largest activation at each ReLU layer for the arrest event

Max Pooling Layer-1

Max Pooling Layer-2

Max Pooling Layer-3

Max Pooling Layer-4

Figure 6 The largest activation at each max pooling layer for the arrest event

The proposed Architecture is used in two ways. In the first method, the CNN model is used for feature extraction as well as for classification. So, after the fully connected layer, the classification layer is used to classify an image as normal or abnormal. In the second method, the CNN model is used to extract features, and SVM is used for classification. The extracted features by the CNN model are used to train the linear SVM classifier. CNN and SVM both classifiers almost give the identical results. Table 1 represents the achieved accuracy and AUC values for all mentioned datasets. Table-2 and Table-3 represent confusion matrices for all datasets using CNN and SVM classifiers, respectively.

Batch Normalization Layer-1

Batch Normalization Layer-2

Batch Normalization Layer-3

Batch Normalization Layer-4

Figure 7 The largest activation at each batch normalization layer for the arrest event

Table 1 Experimental results using CNN architecture

Dataset	Number of Images			Accuracy (%)		Area Under Curve	
	For Whole Dataset	For Training set	For Test set	Classification by CNN	Feature extraction by CNN & classification by SVM	Classification by CNN	Feature extraction by CNN & classification by SVM
UMN	7003	5603	1400	99.50	99.43	0.9999	0.9991
UCSDPed1	14000	11200	2800	99.89	99.89	1	0.9995
Violent Flows	21800	17440	4360	99.86	99.95	1	1
Subway Entrance	143996	115197	28799	99.99	99.99	1	1
UCFCrime2 Local	251151	174673	76478	74.38	71.16	0.7163	0.7303

Table 2 Confusion matrices for various datasets using CNN classifier

True class		Predicted class									
		UMN		UCSDPed1		Violent Flows		Subway Entrance		UCFCrime2Local	
		Ab	N	Ab	N	Ab	N	Ab	N	Ab	N
	Ab	224	3	808	1	2463	5	751	1	4180	9836
	N	4	1169	2	1989	1	1891	3	28044	9754	52708

Table 3 Confusion matrices for various datasets using SVM classifier

True class		Predicted class									
		UMN		UCSDPed1		Violent Flows		Subway Entrance		UCFCrime2Local	
		Ab	N	Ab	N	Ab	N	Ab	N	Ab	N
	Ab	224	3	808	1	2466	2	751	1	9171	4845
	N	5	1168	2	1989	-	1892	1	28046	17208	45254

The reason for getting such accurate results is the tuning of parameters and hyperparameters of the model. If the number of layers is taken less or more than the number described in the proposed architecture, performance decreases. The number of filters, filter size and learning rate value also affect much to the model performance. Many experiments are done using the various parameter and hyperparameter values and the best values are selected to propose the perfect architecture. Therefore, the same model gives the best performance on all the datasets without altering parameter or hyperparameter values. The parameters are set as shown in the following Table-4.

Table 4 Tuned parameters for the CNN architecture

Sr No	Parameter Name	Value
1	Optimizer	Stochastic Gradient Descent with Momentum (SGDM)
2	Initial Learning Rate	0.01
3	Learning Rate Drop Factor	0.2
4	Learning Rate Drop Period	5
5	Number of Epochs	5
6	Batch size	64

The same model with the same tuned parameters and hyperparameters is used for classification of various abnormal events. As discussed in section 3.1, UCFCrime2Local dataset contains seven different events: Arrest, Assault, Burglary, Normal, Robbery, Stealing and Vandalism. So, classification of abnormality is done using this dataset. Experiments for classification are done in three ways. In the first one, video frames are taken as per the given train-test split. For the other two ways, all images of a particular event's videos are combined according to class labels. Now using all images, two types of experiments are done (1) Non-stratified holdout validation with 80% training and 20% test data and (2) 5-fold cross-validation.

Table 5 Classification results using UCFCrime2Local dataset as per the given train-test split

Event	Number of Images			Area under Curve (AUC)	
	Total For each class	For Training Set	For Test Set	Classification by CNN	Classification by SVM
Arrest	8739	6861	1878	0.7061	0.6755
Assault	5560	3505	2055	0.7276	0.6928
Burglary	7841	5431	2410	0.5160	0.3161
Normal	203272	140810	62462	0.7115	0.5955
Robbery	8650	5618	3032	0.6263	0.6717
Stealing	8668	5440	3228	0.6323	0.4976
Vandalism	8421	6794	1627	0.5845	0.5427

Table 5, Table 6 and Table 7 represent the achieved AUC values for classification using the given train- test split, 80% training data and 5-fold cross-validation, respectively. Table 8 shows overall classification accuracy for each experiment. As this dataset contains each video for a specific event

with different background, as per given train-test split the videos available in training set are completely different compared to test set considering background information. For 80% training data and 5-fold cross-validation performance is better compared to the given train-test split because few images of each specific event's video with same background are present in the training set. So, it can be concluded that if the model is used for a particular place and is trained by standard background images, it gives outstanding classification results.

Table 6 Classification results using UCFCrime2Local dataset with 20% test data

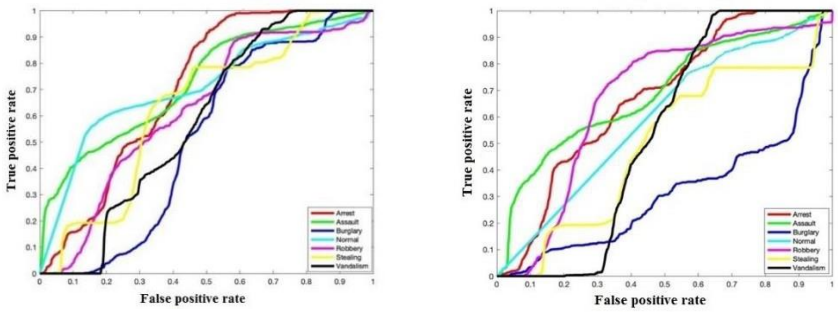
Event	Number of Images			Area under Curve (AUC)	
	Total For each class	For Training Set	For Test Set	Classification by CNN	Classification by SVM
Arrest	8739	6991	1748	0.9876	0.9949
Assault	5560	4448	1112	0.9991	0.9991
Burglary	7841	6273	1568	0.8335	0.8130
Normal	203272	162618	40654	0.8852	0.8423
Robbery	8650	6920	1730	0.8943	0.8712
Stealing	8668	6935	1733	0.9687	0.9158
Vandalism	8421	6737	1684	0.7346	0.7602

Table 7 Classification results using UCFCrime2Local dataset with 5-fold cross-validation

Event	Number of Images			Area under Curve (AUC)	
	Total For each class	For Training Set	For Test Set	Classification by CNN	Classification by SVM
Arrest	8739	6991	1748	1	1
Assault	5560	4448	1112	1	1
Burglary	7841	6273	1568	1	1
Normal	203272	162618	40654	1	1
Robbery	8650	6920	1730	1	1
Stealing	8668	6935	1733	1	1
Vandalism	8421	6737	1684	1	1

Table 8 Overall classification accuracy

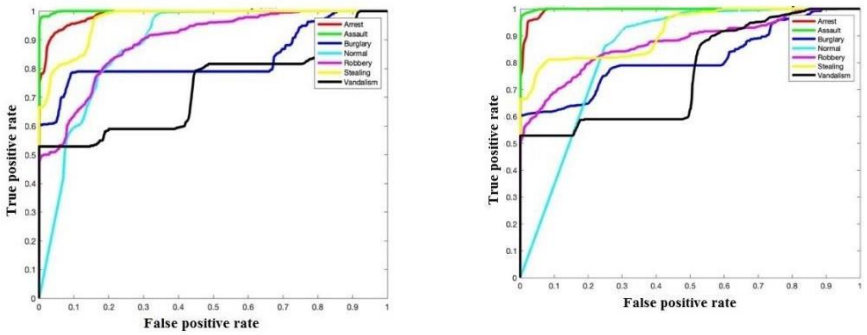
Dataset Partition	Accuracy (%)	
	Classification by CNN	Classification by SVM
As per the train-test split	71.79	70.20
80% training data	87.29	88.13
5 fold cross-validation	99.9924	99.9944



(a) Using CNN classifier **(b) Using SVM classifier**

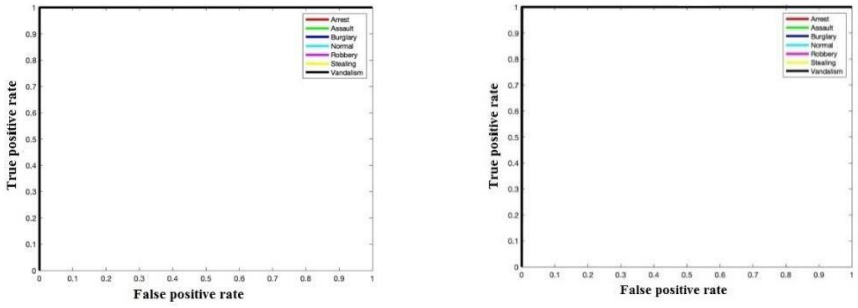
Figure 8 ROC curves for events classification using the UCFCrime2Local dataset as per given train-test split

Figure 8, Figure 9 and Figure 10 represent the ROC curves for classification using given train-test split, 80% training data and 5-fold cross-validation respectively. In figure 10, only one curve is visible because the achieved AUC value is 1 for each category.



(a) Using CNN classifier **(b) Using SVM classifier**

Figure 9 ROC curves for events classification using the UCFCrime2Local dataset with 20% test data



(a) Using CNN classifier **(b) Using SVM classifier**

Figure 10 ROC curves for events classification using the UCFCrime2Local dataset with 5-fold cross-validation

3.3 CNN-LSTM Based Approach

3.3.1 LSTM Network

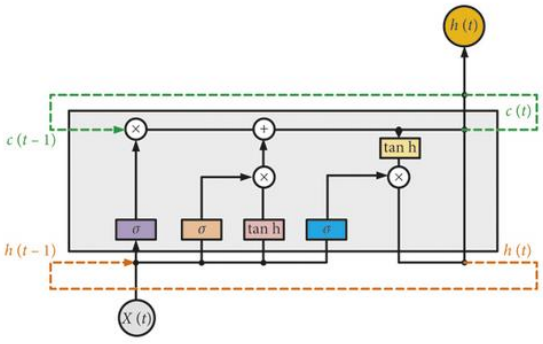


Figure 11 LSTM architecture [31]

LSTM is a Recurrent Neural Network (RNN) architecture introduced by Sepp Hochreiter et al. [30]. It remembers values over arbitrary intervals. RNN is used to classify, process, and predict time series

data. Traditional RNN can remember short duration past sequence. If output depends on the past long duration sequence, it fails as it forgets the starting point. In contrast, the LSTM is insensitive to gap length. It can remember past long sequence.

The Figure 11 shows the basic architecture of the LSTM cell. The cell is a repeated module that contains three different gate structures, forget gate, input gate and output gate. In the shown Figure, c, h and x represent cell state, hidden state and input, respectively. The first sigmoid activation function is the forget gate. It defines which information should be omitted from the previous cell state (C_{t-1}). Sigmoid function outputs a number between 0 and 1. If output is 1 then the whole information is passed, and if the output is 0, then nothing is passed from the previous cell state to the current cell state. The first tanh and second sigmoid activation functions represent the input gate. The sigmoid function decides which values will be updated, and the tanh function creates a new candidate value that could be added to the state. The last sigmoid is the output gate and highlights which information should be going to the next hidden state. As the tanh function outputs values between -1 and 1, the cell state is represented by tanh activation function and multiplied with the output of the last sigmoid function, so the result of the output gate contains decided information.

3.3.2 BiLSTM Network

Bi-directional long short-term memory network is known as BiLSTM network. It flows the sequence information in both directions forward and backward. Figure 12 shows BiLSTM Network. It preserves the future and the past information; therefore, it is usually employed where the sequence-to-sequence tasks are needed. This kind of network can be used in text classification, speech recognition and forecasting models.

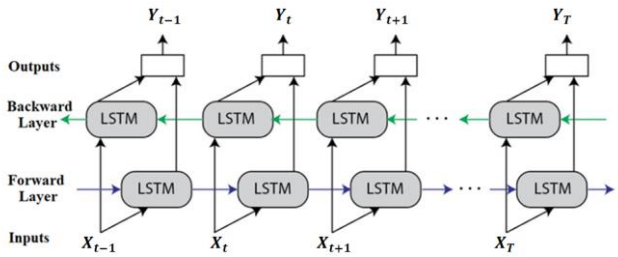


Figure 12 BiLSTM network [37]

3.3.3 CNN-LSTM Architecture

The CNN based approach works only in the spatial domain, i.e., images. To consider the temporal domain with spatial domain, the CNN-LSTM architecture is used, as shown in Figure 13. The sequences of feature vectors are generated from the input video sequences using the proposed Convolutional Neural Network architecture. Here BiLSTM model is used to consider input for both forward and backward directions. The detailed description of the proposed LSTM architecture is given in Table 9. The LSTM network is trained using the feature sequences with tuned parameters as shown in Table 10.

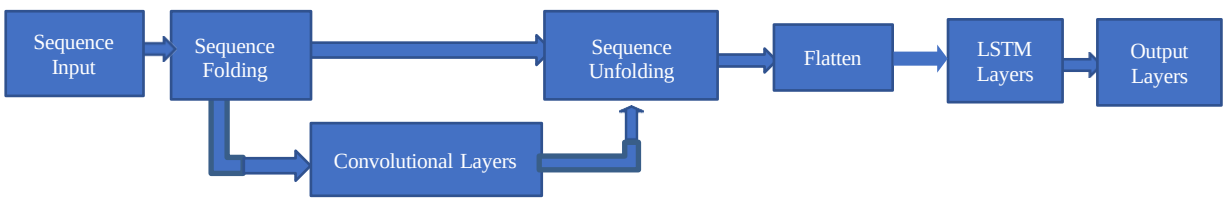


Figure 13 CNN-LSTM architecture

Table 9 The proposed LSTM architecture

Sr	Layer Name	Operation	Description
1.	'sequence'	Sequence Input	Sequence input with 1024 dimensions
2.	'bilstm'	BiLSTM	BiLSTM with 500 hidden units
3.	'fc'	Fully Connected	2 fully connected layer
4.	'softmax'	Softmax	softmax
5.	'classification'	Classification Output	crossentropyex with classes 'Abnormal' and 'Normal'

Table 10 Tuned parameters for the LSTM architecture

Sr No	Parameter Name	Value
1	Number of Nodes	500
2	Optimizer	SGDM
3	Initial Learning Rate	0.01
4	Batch size	8
5	Number of Epochs	5

3.3.4 Experimental Results

The experiments here are performed with 80% of the data allocated for training and 20% for testing using the UMN, UCSDPed1, Violent Flows and Subway Entrance datasets. For UCFCrime2Local dataset experiments are done as per the given train-test split. The Table 11 represents the achieved accuracy and AUC values using various datasets for abnormal event detection. Table 12 represents confusion matrices for the same. Table 13 represents classification results using the UCFCime2Local dataset as per given train-test split, in which the AUC value for each event is represented. Figure 14 shows the ROC curves for the same. The achieved overall classification accuracy is 73.86%. The UCFCrime2Local is a recently published complex dataset, and fewer experiments are done using it by researchers. Available state-of-the-art methods, using this dataset, are for abnormality detection, not for classification. Although it is a complex dataset, the proposed model performs well on it.

Table 11 Experimental results for abnormal event detection using the CNN-LSTM architecture

	Number of videos			Accuracy (%)	Area under Curve
	Total for each class	For Training Set	For Test Set		
UMN	22	18	4	100	1
UCSDPed1	70	56	14	100	1
Violent-Flows	246	196	50	98	1
Subway Entrance	152	121	31	96.77	0.9846
UCFCrime2Local	300	210	90	73.33	0.79

Table 12 Confusion matrices for various datasets using the CNN-LSTM architecture

True Class		Predicted Class									
		UMN		UCSDPed1		Violent Flows		Subway Entrance		UCFCrim2Local	
		Ab	N	Ab	N	Ab	N	Ab	N	Ab	N
	Ab	2	-	5	-	27	1	2	1	18	13
	N	-	2	-	9	-	22	2	26	11	48

Table 13 Experimental results for events classification using the CNN-LSTM architecture

Event	Number of Videos			Area under Curve (AUC)
	Total for each class	For Training Set	For Test Set	
Arrest	13	10	3	0.7471
Assault	19	14	5	0.7153
Burglary	17	10	7	0.6024
Normal	200	141	59	0.8387
Robbery	18	11	7	0.7349
Stealing	15	11	4	0.7674
Vandalism	18	13	5	0.7929

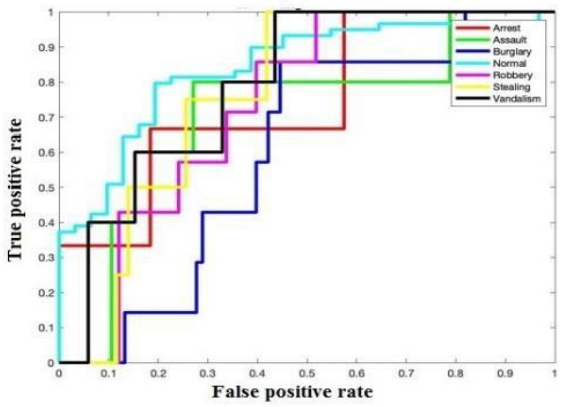


Figure 14 ROC curves for classification results using the CNN-LSTM architecture

4 Comparisons of The Proposed Approaches

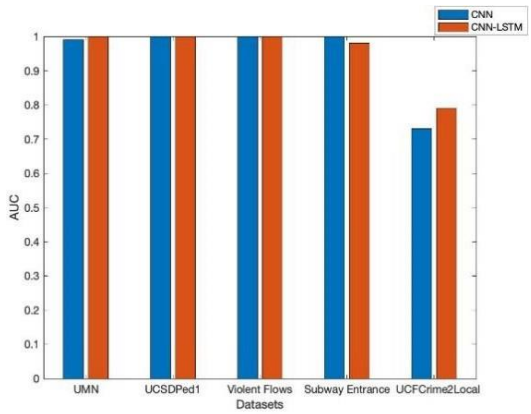


Figure 15 Performance of the proposed approaches for abnormal event detection

This section represents the comparison between the proposed approaches for abnormal event detection and classification and comparison of them with state-of-the-art methods. Figure 15 shows the performance of the proposed supervised deep learning-based approaches for abnormal event detection. Horizontal axis represents datasets and vertical axis represents the AUC values. Figure represents that both CNN and CNN-LSTM based approaches provide excellent results using various

datasets. The CNN-LSTM architecture gives higher performance on challenging weakly labeled dataset UCFCRime2Local because of extraction of temporal information with spatial information.

Figure 16 represents the performance of the proposed approaches for events classification. Horizontal axis represents events and vertical axis represents the AUC values. The Figure represents that the CNN- LSTM based approach provides higher AUC values for various events except for assault event.

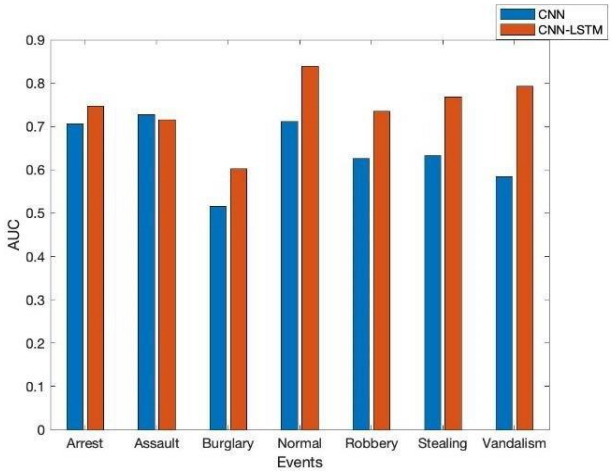


Figure 16 Performance of the proposed approaches for events classification

Table 14 The Performance of various methods on the UMN dataset

Approach	Accuracy (%)	AUC
Motion Structure + CNN Classifier [8]	96.74	-
Plug and Play CNN[10]	-	0.98
Generative Adversarial Nets (GAN)[22]	-	0.99
Abnormal Event Detection Network [25]	-	0.997
The Proposed CNN Model	99.5	0.9999
The Proposed CNN LSTM Model	100	1

Table 15 The performance of various methods on the UCSDPed1 dataset

Approach	Accuracy (%)	AUC
Motion Structure + CNN Classifier [8]	94.28	-
Plug and Play CNN [10]	-	0.957
Appearance and Motion DeepNet (AMDN) [32]	-	0.921
Generative Adversarial Nets (GAN) [22]	-	0.974
Spatiotemporal Autoencoder [19]	-	0.899
Convolutional LSTM- Autoencoder [18]	-	0.43
Variational Autoencoder [18]	-	0.63
Convolutional Autoencoder [18]	-	0.726
Convolutional Variational Autoencoder [18]	-	0.727
Recurrent Convolutional Autoencoder [18]	-	0.694
Recurrent Convolutional Variational Autoencoder [18]	-	0.727
Two-Stream R-ConvVAE [18]	-	0.75
Convolutional LSTM [18]	-	0.67
The Proposed CNN Model	99.89	1
The Proposed CNN LSTM Model	100	1

Comparisons between the proposed approaches and state-of-the-art methods are demonstrated here using various datasets. Table 14, Table 15, Table 16, Table 17 and Table 18 show the performance of the proposed and state-of-the-art deep learning methods using UMN, UCSDPed1, Violent-Flows, Subway Entrance and UCFCrime2Local datasets, respectively. Each table displays the AUC values and/or accuracy (%) reported by different researchers for a specific dataset, using both supervised and unsupervised deep learning methods. As AUC is scale invariant and classification threshold invariant, it is widely used as performance measurement metric. In the proposed methods, both accuracy and AUC values are computed, while most existing methods report only AUC values.

Table 16 The performance of various methods on the Violent-Flows dataset

Approach	Accuracy (%)	AUC
3D CNN [33]	98	0.98
Convolutional LSTM [34]	94.57	-
The Proposed CNN Model	99.86	1
The Proposed CNN LSTM Model	98	1

Table 17 The performance of various methods on the Subway Entrance dataset

Approach	Accuracy (%)	AUC
Fully Convolutional Neural Network [11]	-	0.904
Convolutional Autoencoder (ConvAE) [18]	-	0.842
Convolutional Variational Autoencoder (ConvVAE) [18]	-	0.844
Recurrent Convolutional Autoencoder (R-ConvAE) [18]	-	0.821
Recurrent Convolutional Variational Autoencoder (R-ConvVAE) [18]	-	0.846
Two-Stream R-ConvVAE [18]	-	0.851
Multilevel Anomaly Detector [20]	-	0.8234
Spatiotemporal Autoencoder [19]	-	0.847
Two Stream Autoencoder [24]	-	0.873
The Proposed CNN Model	99.99	1
The Proposed CNN LSTM Model	96.77	0.9846

Table 18 The performance of various methods on the UCFCrim2Local dataset

Approach	Accuracy (%)	AUC
Inflated 3D ConvNet (I3D) + Regression for Whole frame [9]	-	0.5612
I3D + Regression for Spatiotemporal Tube [9]	-	0.7473
The Proposed CNN Model	74.38	0.7163
The Proposed CNN-LSTM Model	73.33	0.79

5 Conclusion and Future Scope

5.1 Conclusion

In this work, two methods based on CNN and CNN-LSTM are proposed for automatic abnormal event detection in surveillance scenes. Both approaches give outstanding performance on various benchmark datasets. For the CNN based approach, the proposed architecture performs better than the pre-trained GoogleNet architecture. The proposed CNN architecture provides AUC values 0.99, 1, 1, 1 and 0.73 using UMN, UCSDPed1, Violent-Flows, Subway Entrance and UCFCrime2Local datasets, respectively while the CNN-LSTM architecture provides AUC values 1, 1, 1, 0.9846 and 0.79 respectively. Because of consideration of both spatial and temporal domain, the CNN-LSTM

architecture performs better than CNN architecture using complex weakly labeled UCFCrime2Local dataset.

Both proposed approaches perform efficiently for multiclass events classification. The CNN-LSTM model gives better AUC values than the CNN model for various events using UCFCrime2Local dataset except for Assault event. The overall classification accuracies for the CNN and CNN-LSTM architectures are 71.79% and 73.86%, respectively.

The proposed approaches show strong performance in automatic abnormal event detection in crowd scenes across various challenging real-world datasets. The CNN uses only spatial information, whereas the CNN-LSTM incorporates both spatial and temporal information, leading to better performance than the CNN. Given its promising results in simulations, it has the potential to be extended to real-time applications.

5.2 Future Work

Public and private sectors demand smart surveillance system which requires accurate solutions for automatic detection and recognition of anomalies in surveillance scenes, and there is still a large room for improvement. In this research specific types of abnormal events are examined. Further, more events can be considered with large datasets with the use of GPUs to reduce the overall training time. For further research, emotional aspects can also be considered to classify abnormal events because changes in people's emotions are usually a precursor of abnormal events.

References

[1] N. Sjarif, S. Shamsuddin, S. Hashim and S. Yuhaniz, "Crowd Analysis and Its Applications" In Proc. International Conference on Software Engineering and Computer Systems,2011. DOI:10.1007/978-3-642-22170-5_59

[2] M. Zitouni, H. Bhaskar, J. Dias and M. Al-Mualla, "Advances and Trends in Visual Crowd Analysis: A Systematic Survey and Evaluation of Crowd Modelling Techniques", Neurocomputing, Vol. 186, pp.139- 159, 2016. https://doi.org/10.1016/j.neucom.2015.12.070

[3] G. Tripathi, K. Singh and D. Vishwakarma, "Convolutional Neural Networks for Crowd Behaviour Analysis: A Survey" The Visual Computer International Journal of Computer Graphics, Vol. 35, pp.753– 776, 2019. https://doi.org/10.1007/s00371-018-1499-5

[4] Unusual Crowd Activity Dataset by University of Minnesota, http://mha.cs.umn.edu./movies/crowdactivity-all.avi/

[5] UCSD Anomaly Detection Dataset by University of California and San Diego, http://www.svcl.ucsd.edu/projects/anomaly/dataset.html

[6] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent Flows: Real-Time Detection of Violent Crowd Behavior", In Proc. 3rd IEEE Conference on Computer Vision and Pattern Recognition, June 2012.

[7] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. "Robust Real-Time Unusual Event Detection Using Multiple Fixed-Location Monitors", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, No.3, pp. 555–560, 2008 DOI: 10.1109/TPAMI.2007.70825

[8] T. Mostafa, J. Uddin and Md. Haider Ali, "Abnormal Event Detection in Crowded Scenario", In Proc. 3rd International Conference on Electrical Information and Communication Technology, 2017 DOI: 10.1109/EICT.2017.8275217

[9] F. Landi, C. Snoek and R. Cucchiara, "Anomaly Locality in Video Surveillance", arXiv:1901.10364, 2019 https://doi.org/10.48550/arXiv.1901.10364

[10] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto and N. Sebe, "Plug-and-Play CNN for Crowd Motion Analysis: An Application in Abnormal Event Detection", In Proc. IEEE Winter Conference on Applications of Computer Vision, pp.

1689- 1698, 2018. DOI:10.48550/arXiv.1610.00307

[11] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayedd, and R. Klette, "Deep-Anomaly : Fully Convolutional Neural Network for Fast Anomaly Detection in Crowded Scenes", Computer Vision and Image Understanding, 2018. https://doi.org/10.48550/arXiv.1609.00866

[12] Y. Feng, Y. Yuan, and X. Lu, "Learning Deep Event Models for Crowd Anomaly Detection", Neurocomputing, Vol. 219, pp.548–556 , 2017. https://doi.org/10.1016/j.neucom.2016.09.063

[13] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, and Z. Zhang, "Spatial-temporal Convolutional Neural Networks for Anomaly Detection and Localization in Crowded Scenes", Signal Processing Image Communication, Vol. 47, pp 358–36 , 2016. DOI:10.1016/j.image.2016.06.007

[14] S. Smeureanu, R. Ionescu, M. Popescu and B. Alexe, "Deep Appearance Features for Abnormal Behavior Detection in Video", Image Analysis and Processing—ICIAP 2017. DOI:10.1007/978-3-319-68548-9_70

[15] J. Sun, J. Shao and C. He, "Abnormal Event Detection for Video Surveillance Using Deep One-Class Learning", Multimedia Tools and Application, 2017. https://doi.org/10.1007/s11042-017-5244-2

[16] W. Li, V. Mahadevan and N. Vasconcelos , "Anomaly Detection and Localization In Crowded Scenes", IEEE Transactions on Pattern Analysis and Machine Intelligence, June - 2013. DOI: 10.1109/TPAMI.2013.111

[17] R. Hinami, T. Mei, and S. Satoh, "Joint Detection and Recounting of Abnormal Events by Learning Deep Generic Knowledge", In Proc. International Conference on Computer Vision, 2017.

[18] S. Yan, J. Smith, W. Lu and B. Zhang, "Abnormal Event Detection from Videos Using a Two-stream Recurrent Variational Autoencoder", IEEE Transactions on Cognitive and Developmental Systems, 2018 DOI:10.1109/TCDS.2018.2883368

[19] Y. Chong, and Y. Tay "Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder", In Proc. 14th International Symposium on Advances in Neural Networks, 2017. DOI:10.1007/978-3-319-59081-3_23

[20] H. Vu, T. Nguyen, T. Le, W. Luo, and D. Phung, "Robust Anomaly Detection in Videos Using Multilevel Representations", In Proc. AAAI Conference on Artificial Intelligence, Vol. 33, No.1, pp. 5216- 5223, 2019. DOI: https://doi.org/10.1609/aaai.v33i01.33015216

[21] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded Scene Analysis: A Survey", IEEE Transactions on Circuits and Systems for Video Technology, 2015. DOI: 10.1109/TCSVT.2014.2358029

[22] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, N. Sebe, "Abnormal Event Detection in Videos Using Generative Adversarial Nets", In Proc. IEEE International Conference on Image Processing, 2017 DOI: 10.1109/ICIP.2017.8296547

[23] W. Sultani, C. Chen, M. Shah, "Real-world Anomaly Detection in Surveillance Videos", In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.6479-6488, 2018. DOI**:** 10.1109/CVPR.2018.00678

[24] J. Feng, Y. Liang and L. Li, "Anomaly Detection in Videos Using Two-Stream Autoencoder with Posthoc Interpretability", Computational Intelligence and Neuroscience, Vol. 2021, 2021 https://doi.org/10.1155/2021/7367870

[25] T. Wang, Z. Miao, Y. Chen, Y. Zhou, G. Shan, H. Snoussi, "AED-Net: An Abnormal Event Detection Network", Engineering, Vol. 5, No.5, pp 930-939, 2019 https://doi.org/10.1016/j.eng.2019.02.008

[26] D. Mishkin, N. Sergievskiy, and J Matas, "Systematic evaluation of convolution neural network advances on the imagenet" Computer Vision and Image Understanding,2017 https://doi.org/10.1016/j.cviu.2017.05.007

[27] C. Garbin, X. Zhu, and O. Marques, "Dropout vs. batch normalization: An empirical study of their impact to deep learning", Multimedia Tools and Applications, 79, 12777–12815,2020 https://doi.org/10.1007/s11042-019-08453-9

[28] Ioffe and Szegedy, "Batch Normalization: Accelerating deep network training by reducing internal covariate shift"Academic Press, 2015 https://doi.org/10.48550/arXiv.1502.03167

[29] X. Li, S. Chen, X. Hu and J. Yang, "Understanding the Disharmony Between Dropout and Batch Normalization by Variance Shift", In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019 https://doi.org/10.48550/arXiv.1801.05134

[30] S. Hochreiter and J. Schmidhuber, "Long Short Term Memory", Neural Computation, Vol. 9, No.8, pp.1735-1780,1997. https://doi.org/10.1162/neco.1997.9.8.1735

[31] D. Xiao, Y. Huang, C. Qin, H. Shi, and Y. Li, "Fault diagnosis of induction motors using recurrence quantification analysis and LSTM with weighted BN," *Shock and Vibration.*, vol. 2019, pp. 1-14, 2019 https://doi.org/10.1155/2019/8325218

[32] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting Anomalous Events in Videos by Learning Deep Representations of Appearance and Motion," Computer Vision and Image Understanding, Vol 219, issue C, pp- 548-556, 2017 https://doi.org/10.1016/j.cviu.2016.10.010

[33] F. Ullah, A. Ullah, K. Muhammad, I. Haq and S. Baik, "Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network", Sensors, MDPI, Vol. 19, No. 10, 2019 https://doi.org/10.3390/s19112472

[34] S. Sudhakaran and O. Lanz, "Learning to Detect Violent Videos using Convolutional Long Short-Term Memory" 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp.1- 6,2017 DOI**:** 10.1109/AVSS.2017.8078468

[35] X. Glorot and Y. Bengio. "Understanding the Difficulty of Training Deep Feedforward Neural Networks." In Proc. 13th International Conference on Artificial Intelligence and Statistics, pp. 249-256, 2010

[36] S. Saha, "A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way", Towards data Science, 2018 https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-theeli5-way-3bd2b1164a53

[37] E. Lodhi, F-Y. Wang, G. Xiong, L. Zhu, T. S. Tamir, W. U. Rehman, M. A. Khan, "A Novel Deep Stack-Based Ensemble Learning Approach for Fault Detection and Classification in Photovoltaic Arrays" Remote Sensing, Vol 15, No.1277, 2023 https://doi.org/10.3390/rs15051277

[38] A R, Bushara. "A deep learning-based lung cancer classification of CT images using augmented convolutional neural networks." *ELCVIA Electronic Letters on Computer Vision and Image Analysis* 21, no. 1, 2022 https://doi.org/10.5565/rev/elcvia.1490

[39] C. Cortes and V. Vapnik, "Support-Vector Networks", Machine Learning, Vol. 20, pp. 273–297, 1995 https://doi.org/10.1007/BF00994018

[40] V. Salunkhe, "Support Vector Machine", 23 July, 2021 https://medium.com/@viveksalunkhe80/support-vector-machine-svm-88f360ff5f38

[41] A. Vijayan, B. Meenaskshi, A. Pandey, A. Patel, and A. Jain, "Video anomaly detection in surveillance cameras", In IEEE International Conference for Advancement in Technology (ICONAT), pp. 1-4,2022 DOI: 10.1109/ICONAT53423.2022.9726078

[42] X. Wang, Z. Che, B. Jiang, N. Xiao, K Yang, J. Tang, J. Ye, J. Wang, and Qi, Q.Qi, "Robust unsupervised video anomaly detection by multipath frame prediction", IEEE transactions on neural networks and learning systems, 33(6), pp.2301-2312, 2021 DOI**:** 10.1109/TNNLS.2021.3083152

[43] M. Z Zaheer, A. Mahmood, M.H Khan, M.Segu, F. Yu, and S.Lee, S.I., "Generative cooperative learning for unsupervised video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition" pp. 14744-14754, 2022. https://doi.org/10.48550/arXiv.2203.03962

[44] D. Gong, L. Liu, V. Le, B. Saha, M.R. Mansour, S. Venkatesh and A.V.D. Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection" In Proceedings of the IEEE/CVF International Conference on Computer Vision pp. 1705-1714, 2019. https://doi.org/10.48550/arXiv.1904.02639

[45] V.T. Le, and Y.G. Kim, "Attention-based residual autoencoder for video anomaly detection", Applied Intelligence, 53(3), pp.3240-3254, 2023 https://doi.org/10.1007/s10489-022-03613-1