

Detecting anomalies in security cameras with 3D-convolutional neural network and convolutional long short-term memory

Esraa A. Mahareek¹, Eman K. Elsayed^{1,2}, Nahed M. ElDesouky¹, Kamal A. Eldahshan³

¹Mathematics Department, Faculty of Science, Al-Azhar University (Girls branch), Cairo, Egypt

²School of Computer Science, Canadian International College, Cairo, Egypt

³Mathematics Department, Faculty of Science, Al-Azhar University, Cairo, Egypt

Article Info

Article history:

Received Mar 26, 2023

Revised Jul 4, 2023

Accepted Jul 17, 2023

Keywords:

3D-convolutional-neural-network

Anomaly detection

Bidirectional convolutional long

short-term memory

Fight detection

Surveillance videos

Violence detection

ABSTRACT

This paper presents a novel deep learning-based approach for anomaly detection in surveillance films. A deep network that has been trained to recognize objects and human activity in movies forms the foundation of the suggested approach. In order to detect anomalies in surveillance films, the proposed method combines the strengths of 3D-convolutional neural network (3DCNN) and convolutional long short-term memory (ConvLSTM). From the video frames, the 3DCNN is utilized to extract spatiotemporal features, while ConvLSTM is employed to record temporal relationships between frames. The technique was evaluated on five large-scale datasets from the actual world (UCFCrime, XD-Violence, UBIFights, CCTV Fights, UCF101) that had both indoor and outdoor video clips as well as synthetic datasets with a range of object shapes, sizes, and behaviors. The results further demonstrate that combining 3DCNN with ConvLSTM can increase precision and reduce false positives, achieving a high accuracy and area under the receiver operating characteristic-area under the curve (ROC-AUC) in both indoor and outdoor scenarios when compared to cutting-edge techniques mentioned in the comparison.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Esraa A. Mahareek

Mathematics Department, Faculty of Science, Al-Azhar University

Cairo, Egypt

Email: esraa.mahareek@azhar.edu.eg

1. INTRODUCTION

Even for individuals, a significant concern is the monitoring capacity to keep the people safe and its quick response to serve this purpose as protection is the main driver for the deployment of surveillance systems. Although the usage of monitoring devices has risen, human potential has not [1]. As a result, even if there is a large loss of labor and time considering relative to normal events, how improbable abnormal events are to occur, a lot of oversight is necessary to identify odd events that could damage people or a business [2].

For organizations like law enforcement, security, and others, surveillance footage is a valuable source of information. It is an automated system that is used to keep an eye on both interior and outdoor spaces including parking lots, malls, and airports. With the use of 2D or 3D cameras, the captured video streams are transformed into images. Computer vision algorithms analyze these photos to find objects, people, and actions in the scene. In order to detect and respond to unforeseen incidents like robberies, assaults, vandalism, or traffic accidents, video surveillance systems must be able to recognize anomalous actions in these settings.

However, compared to typical events, anomalous occurrences are uncommon. The development of computer vision systems that automatically detect anomalous action in surveillance movies is essential since monitoring surveillance videos is very vital and time-consuming. It might be challenging to detect changes in the scene in many surveillance recordings due to their low quality and discontinuous character. Hand-crafted feature extractors are used in conventional methods to solve this problem to find abnormal events. These methods take a lot of work, and they are challenging to keep up as the video format changes over time. Machine learning innovations in recent years have made it possible to train algorithms to detect anomalies without explicitly defining features.

The problem definition for detecting anomalies in surveillance videos involves developing algorithms that can identify events or behaviors that deviate from the expected norm in a given environment. This task can be particularly challenging due to the complexity and variability of real-world environments, as well as the need for algorithms to operate in real-time, with minimal delay or latency. Furthermore, the algorithms must be able to distinguish between normal and abnormal events with a high level of accuracy to avoid false positives or negatives. To achieve this, various approaches have been proposed such as machine learning techniques that can automatically learn from training data and identify patterns of normal behavior. Another approach is to use deep learning techniques, such as convolutional neural networks (CNNs), which have shown promising results in detecting anomalies in surveillance videos. However, the effectiveness of these approaches depends heavily on the quality and quantity of training data available, as well as the specific features that are used to represent normal and abnormal behaviors. Additionally, the development of more advanced sensors and cameras that can capture high-quality video data with greater detail and resolution has also contributed to improving accuracy.

In this research, we suggest a novel strategy for instantly identifying and employing CNN, which draw attributes from video frames, to classify anomalies in video recordings in accordance with different anomaly classes such as assault, robbery, and fighting. In this method, we first choose convolutional long short-term memory (ConvLSTM) to learn the long-term spatial and temporal characteristics of anomalies, then a 3D-convolutional neural network (3DCNN) to learn the short-term spatial and temporal characteristics. In order to improve training stability and performance, we then merge these networks into a single architecture to perform classification of surveillance videos. In order to learn certain picture properties that are discriminative for various anomaly classes and for each class, multiple layers of convolutional networks are trained on thousands of photos, they are taught to identify the normal from the abnormal frames in a video clip. In order to do this, a characteristic taken from a typical frame is compared to a feature retrieved from an anomaly frame of the same class to determine how similar they are, and the frame is then classified as normal or abnormal by generating a similarity score between the two feature vectors. The biggest disadvantage of this approach is the large amount of training images and datasets required to train the network to recognize important image features. It is a sizable dataset on which we trained our proposal because more than 128 hours of recorded video are available in UCFCrime, which is classified into 8 anomaly courses and 1 normal class. Using the held-out test data, we evaluate the performance of our proposal and find that it outperforms other existing approaches and has a suitable classification accuracy for various anomaly events kinds.

In order to identify various forms of anomalies, we first detail the data set that was used in this study as well as how it was pre-processed, trained, and evaluated using a 3DCNN technique. Following that, we give a summary of the test dataset's findings and display each dataset's classification accuracy and area under the receiver operating characteristic-area under the curve (ROC-AUC). This essay is structured as follows: The literature review of numerous publications connected to this research study is described in section 2. 3D-CNN is discussed in section 3. The method is described in section 4. The dataset is described in section 5, and the pre-processing of the training data is briefly described in section 6, which is followed by a discussion and conclusions.

2. RELATED WORK

Using computer vision to recognize certain actions in security cameras has gained prominence in the action detection industry. The field of computer vision is relevant to this work. In order to automate the task of video anomaly detection, many academics have been working to build effective machine-learning approaches. Figure 1 displays the distribution of papers on anomaly detection from works that were published in the public domain between 2015 and 2021. A model for detecting anomalies in surveillance footage is presented in [3], [4]. There are two phases to the system. Numerous handcrafted elements have been shown on this platform.

Cool-3D (C3D) characteristics have also been extracted using deep learning approaches and anomaly detection using support vector machine (SVM) from video data. Sultani *et al.* [5] used these methods. The next stage is behavior modeling. During this step, SVM is trained using a bag-of-visual-words (BOVW) to understand how to represent typical behavior.

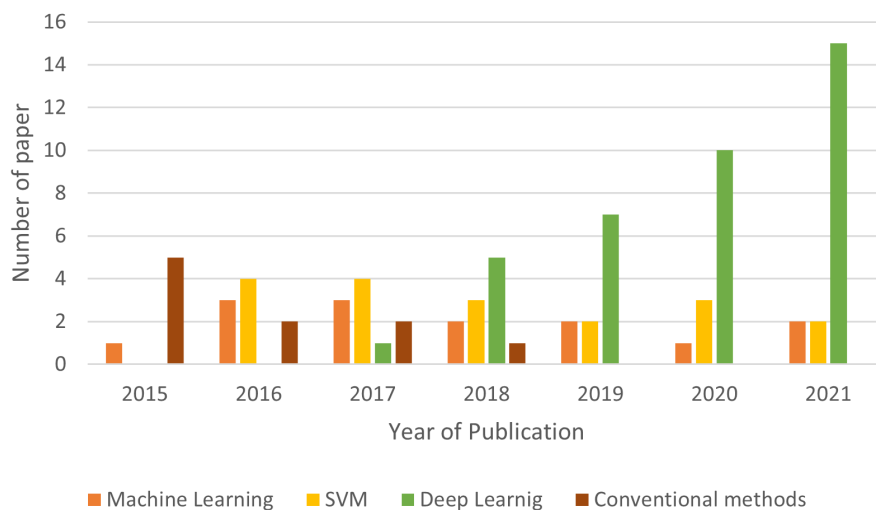


Figure 1. Papers on violence detection are distributed annually

The most dangerous form of bullying in schools is campus violence and is a problem for society worldwide. There are a number of potential techniques, including video-based ones, to prevent college violence as remote monitoring and artificial intelligence (AI) capabilities develop. In order to identify campus violence, Ye *et al.* [6] utilize aural and visual data. Role-playing is employed to collect information about campus violence, and each of the video's 16 frames is extracted to create 4096-dimension feature vectors. When applying the 3DCNN for classification and feature extraction, a total precision of 92% is attained.

Lv *et al.* [7] publish the weakly supervised anomaly localization (WSAL) technique, which focuses on temporally localizing anomalous regions inside anomalous films. influenced by the striking contrast of odd videos. The evolution of adjacent temporal segments is evaluated in order to identify abnormal portions. To achieve this, a high-order context encoding model is proposed that measures dynamic fluctuations in addition to extracting semantic representations to make effective use of the temporal environment. Video classification is more difficult than it is for static images because it is challenging to accurately capture both the spatial and temporal information of succeeding video frames. Ji *et al.* [8] proposed the 3D convolution operator for computing features from both geographical and temporal data. Wu *et al.* [9] provide a self-supervised-sparse-representation (S3R) framework in 2022 that represents the idea of anomalous at the feature level by looking at the synergy between dictionary-based representation and self-supervised learning. In order to improve the discriminativeness of feature magnitudes for recognizing anomalies. Chen *et al.* [10] proposed the magnitude-contrastive-loss and the feature amplification mechanism. Test results using benchmark datasets from XD and UCF for crime and violence.

3. 3D-CONVOLUTIONAL NEURAL NETWORK

A particular kind of neural network termed a 3DCNN is composed of several 2D convolutional layers, followed by many layers of fully linked nonlinear units, all of which are organized in multiple parallel planes like 3D. Similar to how convolutional layers can extract spatial patterns from picture data. To extract temporal patterns from the data, a convolution can be used along the time dimension. However, if our data includes both temporal patterns and spatial, as is the case with video data, we should investigate these two types of patterns jointly since they can combine to produce more complex spatio-temporal patterns. The fundamental principle behind a 3DCNN is to successively process an image or a video clip in two dimensions (spatial and temporal) in order to produce the desired outcome.

By increasing the convolution kernel, 3DCNN accomplishes this by extending CNN. Utilizing 3DCNN [11] is efficient for extracting video features. 3DCNN extracts the spatial-temporal information from the entire video for a more thorough analysis. Given the data format of the video, the 3D convolution kernel is utilized to extract regional spatio-temporal neighborhood information. In (1) represents the formula 3DCNN:

$$v_{ij}^{xyz} = Relu(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-1} w_{ijm}^{pqr} k_{(i-1)m}^{(x+p)(y+q)(z+r)}) \quad (1)$$

where the activation function of the buried layer is denoted by $Relu$. The i^{th} and j^{th} feature graph sets' current value at point (x, y, z) is represented as v_{ij}^{xyz} . The term b_{ij} denotes the bias of the i^{th} layer and the j^{th} feature map. The kernel value $(p, q, r)^{th}$ associated with the m^{th} feature map in the layer before is represented by w_{ijm}^{pqr} . The convolution kernel's height and breadth are denoted by P_i, Q_i , while its size along the temporal axis is denoted by R_i .

4. CONVOLUTIONAL LONG SHORT-TERM MEMORY

ConvLSTM was developed specifically to help with problems predicting spatial-temporal sequences. Compared to regular LSTM, ConvLSTM may be more effective at extracting spatial and temporal characteristics from feature graph sets [11]. This allows ConvLSTM, which analyzes and predicts the occurrences in time series, to incorporate the spatial data from a single feature map. ConvLSTM can therefore be applied to dynamic anomaly recognition to more successfully address timing difficulties. In order to create the ConvLSTM [12] equations, flowing equations are used.

$$i_t = \sigma(W_{pi} * X_t + W_{hi} * K_{(t-1)} + W_{ci} \circ C_{(t-1)} + y_i) \quad (2)$$

$$f_t = \sigma(W_{pf} * X_t + W_{hf} * K_{(t-1)} + W_{cf} \circ C_{(t-1)} + y_i) \quad (3)$$

$$C_t = f_t \circ C_{(t-1)} + i_t \circ \tanh(W_{hc} * K_{(t-1)} + W_{xc} * P_t + y_c) \quad (4)$$

$$O_t = \sigma(W_{po} * P_t + W_{ho} * K_{(t-1)} + W_{co} \circ C_t + y_o) \quad (5)$$

$$f_t = O_t \circ \tanh(C_t) \quad (6)$$

The inputs are P_1, P_2, \dots, P_t , the cell outputs are C_1, C_2, \dots, C_t , and the hidden states are K_1, K_2, \dots, K_t . The gates $i_t, f_t, \text{ and } O_t$ represent the 3D tensors of ConvLSTM, respectively. The last 2D, which are spatial, are rows and columns. The operators “*” and “ \circ ”, respectively, stand for the convolution operator and “Hadamard product”. In this instance, the ConvLSTM is supplemented with the batch normalization layer and dropout layer.

5. PROPOSED METHOD

To classify videos, 3DCNN and ConvLSTM are combined. We will go over the 3DCNN ConvLSTM model's design in this part. We proposed a 3DCNN followed by a ConvLSTM network as a feature extraction model for the dynamic anomaly detection process. Figure 2 displays the architecture of our model. The input layer is composed of a stack of 1632323 downsampled continuous anomalous video frames. The architecture consists of four 3D convolutional layers, each with a distinct filter (32, 32, 64, and 64). the same 333 kernel size, though. After that, a ConvLSTM layer with 64-unit sizes was introduced. ReLU and batch normalization layers are placed after each 3DCNN layer. The 3D max between each pair of 3DCNN layers were pooling and dropout layers. With values of 0.3 and 0.5, dropout layers were used. The softmax activation function follows the output probability in a fully connected layer with 512 and has a significant number of output units equal to the number of anomalous video classes.

To categorize the test video, it split into 16 consecutive frames, and fed into the trained model. The features that the model has discovered are used to determine the likelihood score for each frame. The majority voting schema predicts the label of the video sequence using the probability score of each frame after receiving a prediction of 16 frames as input (7) contains the voting formula for a majority.

$$Y = \text{mode}C(X1, C2, C3, \dots, C(X16)) \quad (7)$$

$X1, X2, \dots, \text{and} X16$ denote the frames collected from the tested video, and Y denotes the class name for the sign gesture video. For each frame, $C(X1), C(X2), C(X3), \dots, C(X16)$ reflect the expected class designation.

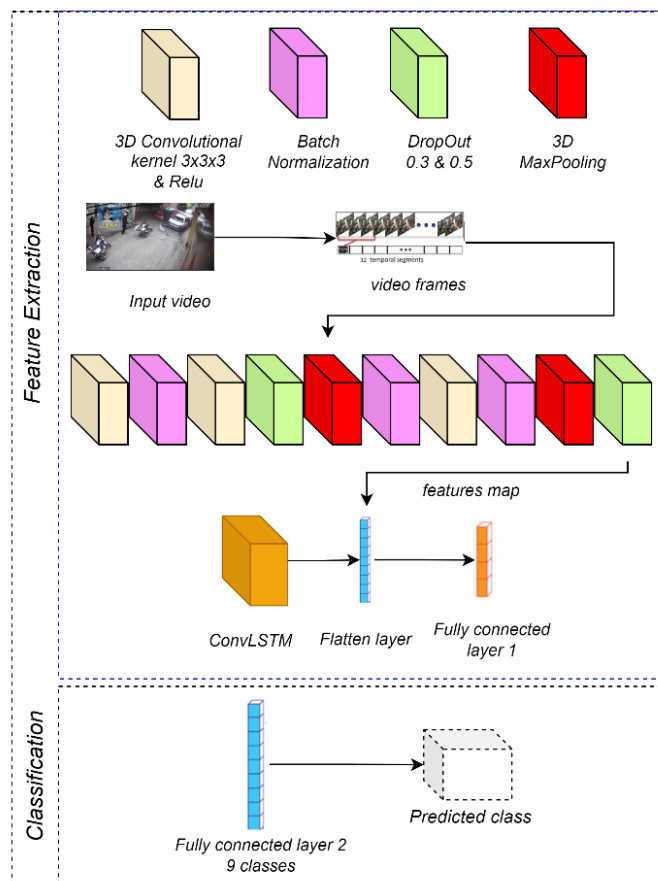


Figure 2. 3DCNN-ConvLSTM model's construction

6. DATASETS

The practice of identifying and analyzing anomalies in video data is gaining popularity. We employ our method to identify and analyze anomalies in numerous important video datasets in order to satisfy this need. For instance, the citations for UCFCrime [5], XDViolence [13], UBIfights [14], NTU CCTVFights [15], and UCF101 [16].

The first dataset consists of 128 hours of video in various sizes and types. Eight categories of crimes, totaling 1,900 words each, are listed in these films. These offenses consist of assault, arson, fighting, breaking and entering, explosion, arrest, abuse, and traffic accidents. Additionally, "Normal" videos—those without any footage of crimes are part of the collection. This dataset can be used to complete two tasks. The first step is to do a general analysis of anomalies, taking into account all anomalies in one group and all usual activities in another. Figure 3 depicts the distribution of the number of videos by class for each UCFCrime course.

With a runtime of 217 hours and a total of 4,754 untrimmed films with audio signals and shaky labels, the second dataset, XDViolence [13], is a sizable, multi-scene dataset. The third dataset, UBIfights, is focused on a specific anomaly detection while still providing a wide range of fighting scenarios. It consists of 80 hours of film that has been fully annotated at the frame level, consisting of 1,000 films, of which 784 show typical daily events and 216 show war scenarios. All unnecessary video clips, including video introductions and news, were removed to avoid interfering with the learning process. The titles of the videos include descriptions of the several types of videos, such as those shot with fixed, rotated, or moveable cameras, or those shot in indoor and outdoor settings, in red, green, and blue (RGB) or grayscale, or both.

The fourth dataset, UCF101, consists of 101 different real-world activity categories of YouTube videos. There are 13,320 videos in 101 different activity categories. The movies are separated into 25 groups, each including four to seven videos of a different activity drawn from the 101 action categories. Similar backdrops, points of view, and other traits may be presented in videos from the same group.

distribution of the percentage of videos in each UCF-Crime per class

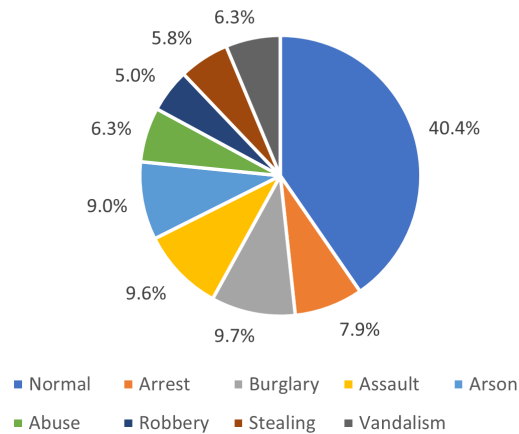


Figure 3. The percent of UCFCrime classes

CCTVfights, the final dataset, contains 1,000 videos of actual fights that were recorded by CCTVs or handheld cameras. There are 280 CCTV films in total, with bouts lasting an average of 2 minutes and anywhere from 5 seconds and 12 minutes. In addition, it contains 720p video of actual battles obtained from other sources (referred to as non-CCTV in this text), primarily from mobile cameras but also on occasion from dashcams, drones, and helicopters. These movies average 45 seconds in length and range in length from 3 seconds to 7 minutes, however some of them have several fights that help the model draw broader conclusions. Table 1 gives a detailed description of the datasets that were used in this experiment.

Table 1. Specifics about each dataset used for comparison

Dataset	No. Samples	No. Hours	No. Classes	Size
UCFCrime	1,900	128	9	60 GB
XDViolence	4,754	217	6	123 GB
CCTV	1,000	17.68	1	7.2 GB
UBIfight	1,000	80	2	7.9 GB
UCF101	13,320	27	101	7 GB

7. IMPLEMENTATION

For evaluation, we split each dataset into 75:25 training and testing divisions. The remaining films are for testing. Each split is further divided into five folds, each of which contains about one-third of the total number of movies for training or validation. Using a Windows 10 Pro machine, an Intel Core i7 CPU,

and 16 GB of RAM, the deep learning model was tested. The Anaconda environment, the Spyder editor, and Python were all used in the system's implementation. In the deep learning libraries, Keras and TensorFlow both appeared. Data handling and pre-processing were done with the Python OpenCV library.

Deep learning models have a variety of characteristics that affect how well they develop and perform. We'll discuss how our network's functioning is impacted by the number of iterations. The number of iterations is one of the most important hyperparameters in modern deep learning systems. As a result of graphic processing units (GPU) parallelism, the model may be trained with fewer iterations in practice, which drastically speeds up computation. As opposed to that, training took longer while using larger iteration numbers than when using smaller ones, but testing accuracy was higher. The number of iterations and batch size can both be significantly impacted by the size of the model training dataset.

8. EXPERIMENTAL RESULTS

A crucial responsibility is performance review. Therefore, the performance of the multi-class classification issue is evaluated or represented using the AUC. It is one of the most basic evaluation criteria for determining whether a categorization model is effective. The level or measure of separability is known as AUC. It demonstrates how well the model can distinguish between classes.

Accuracy and AUC are the two metrics employed by classification methods. A highly accurate model produces extremely few incorrect predictions. However, the cost of those wrong projections to the company is not taken into account. When applied to these business problems, accuracy measurements abstract away the TP and FP characteristics and provide model forecasts with an excessive degree of confidence, which is detrimental to business objectives. AUC is the preferable statistic in such cases because it calibrates the trade-off between sensitivity and specificity at the best-selected threshold. Additionally, while AUC compares two models and evaluates a single model's performance at several thresholds, accuracy evaluates the performance of a single model.

The performance of the trained models was assessed using the AUC test and recognition accuracy. Table 2 summarizes the results of our trials for recognition accuracy and AUC utilizing the suggested 3DCNN-ConvLSTM model with batch sizes of 32 and iterations of 10, 30, and 50. Figures 4 and 5 depict the performance on the UCFCrime datasets' training and validation runs across 10 and 30 iterations, respectively. The training dataset clearly showed that the model performed effectively, performance on the UCF101 training and validation datasets is shown in Figure 6 at 100 iterations. The training accuracy for the model was almost 100%. The best recognition accuracy rate for the UCF101 dataset was 100 percent after 50 iterations when 25% of the dataset was used to test the trained model. While the model's accuracy rates for the UCFCrime, XDViolence, CCTVFight, and UBIfight datasets are 98.5%, 95.1%, 99.1%, and 97.1%, respectively. When trained for 50 iterations, which takes 25 hours for the UCFCrime dataset, the model gets the highest recognition accuracy among the five datasets. While the model's AUC for the UCFCrime, XDViolence, CCTVFight, and UBIfight datasets, respectively, are 92.2%, 87.7%, 94.3%, 93.3%, and 92.3%. Given that the results are competitive with those of the recent studies compared in Tables 3, 4, 5, and 6.

Table 2. Performance evaluation of our model using the UCFCrime, CCTVFights, UBIfight, XDViolence, and UCF101 datasets

Measure	Dataset				
	UCFCrime	XDViolence	CCTVFights	UBIfight	UCF101
Accuracy_10	89%	81.9%	91.7%	89.7%	90.7%
AUC_10	80%	79%	83%	82.6%	87%
Accuracy_30	93.4%	92.3%	94.1%	93.1%	95.1%
AUC_30	85.6%	83.2%	89%	89.8%	89.3%
Accuracy_50	98.5%	95.1%	99%	97.1%	100%
AUC_50	92.2%	87.7%	94.3%	93.3%	92.3%

In order to properly assess the model, Table 3 contrasts the outcomes for additional models provided by other studies for the UCFCrime dataset. It shows that our proposal produces the top AUC outcomes, 92.2 across 50 iterations. and achieves 87.7 in AUC and 95.1 in accuracy for the CCTVFights dataset, respectively. In order to fully assess the model, Table 4 contrasts the outcomes for additional models provided by previous studies

for the XDViolence dataset. It shows that our proposal offers the top AUC outcomes 87.7% at 50 iterations. In order to fully assess the model, Table 5 contrasts the outcomes for additional models provided by other methods for the UBIfights dataset. It shows that our proposal offers the top AUC outcomes 93.3 percent in 50 iterations. In order to fully assess the model, Table 6 contrasts the outcomes for additional models provided by previous research for the UCF-101 dataset. It shows that our proposal offers highest levels of accuracy, 100% in 50 iterations. Figures 7 and 8 show characteristics typical of real-time abuse and explosion videos, for instance.

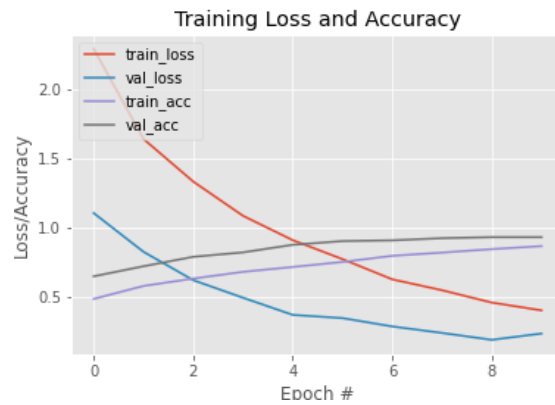


Figure 4. Model accuracy during validation/training for 10 iterations on the UCFCrime dataset

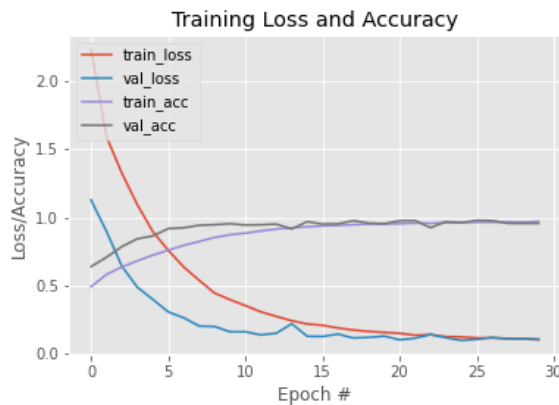


Figure 5. Model accuracy during validation/training for 30 iterations on the UCFCrime dataset

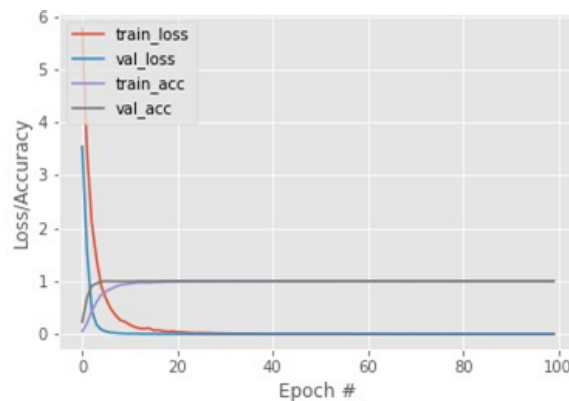


Figure 6. Model accuracy during validation/training for 100 iterations on the UCF101 dataset

Table 3. Comparison of our model's output with that of additional models for UCFCrime dataset

Reference	AUC	Technique	Year
[10]	86.98%	Magnitude-contrastive glance-and-focus network (MGFN)	2022
[9]	85.99%	Self-supervised sparse representation (S3R)	2022
[7]	85.38%	Weakly supervised anomaly localization (WSAL)	2020
[17]	84.89%	Learning causal temporal relation (LCTR) and Feature discrimination for anomaly detection (FDAD)	2021
[18]	84.48%	Multi-stream-network with late-fuzzy-fusion	2022
[19]	84.03%	Real time floor monitoring (RTFM)	2021
ours	92.2%	3DCNN-ConvLSTM	2023

Table 4. Comparison of our model's output with that of additional models for the XDViolence dataset

Reference	AUC	Technique	Year
[9]	80.26%	Self-supervised sparse representation (S3R)	2022
[10]	82.11%	Magnitude-contrastive glance-and-focus network (MGFN)	2022
[19]	77.81%	Real time floor monitoring (RTFM)	2021
[20]	83.54%	Cross-modal-awareness-local-arousal (CMA-LA)	2022
[21]	83.4%	Modality-aware-contrastive-instance-learning-with-self-distillation (MACIL-SD)	2022
ours	87.7%	3DCNN-ConvLSTM	2023

Table 5. Comparison of our model's output with that of additional models for the UBIfights dataset

Reference	AUC	Technique	Year
[1]	90.6%	Gaussian mixture model-based (GMM)	2020
[5]	89.2%	Sultani <i>et al.</i>	2018
[22]	61%	Variational-autoEncoder (S2-VAE)	2018
ours	93.3%	3DCNN-ConvLSTM	2023

Table 6. Comparison of our model's output with that of additional models for the UCF101 dataset

Reference	AUC	Technique	year
[23]	98.64%	Frame selection SMART	2020
[24]	98.6%	OmniSource	2020
[25]	98.2%	Text4Vis	2022
[26]	98.2%	Local and global diffusion (LGD-3D)	2019
ours	100%	3DCNN+ConvLSTM	2023

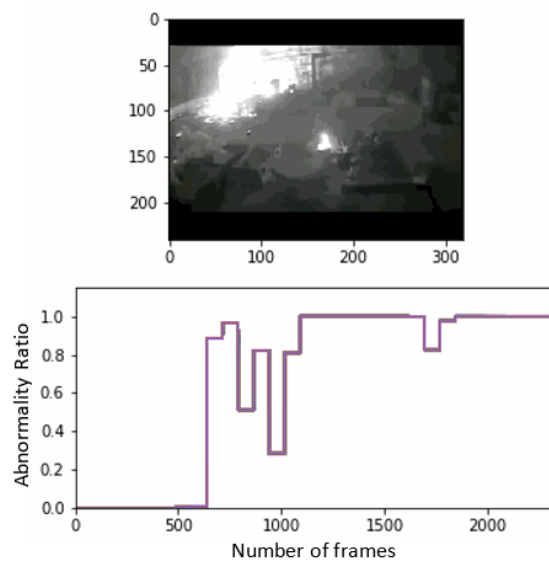


Figure 7. The real-time detection of anomalies in explosion videos

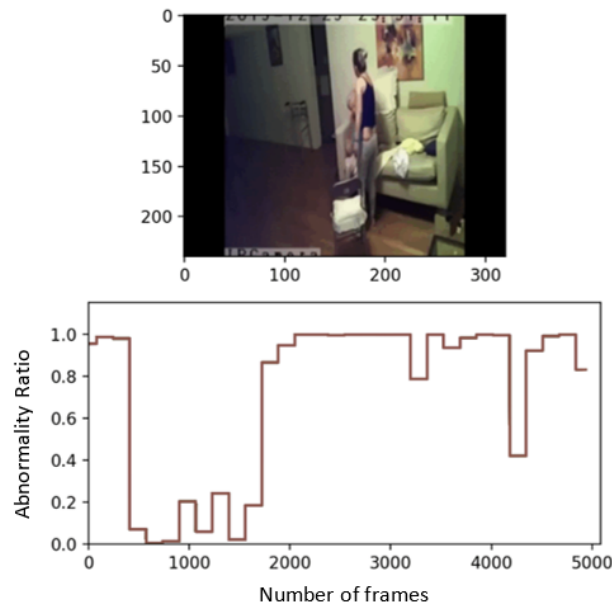


Figure 8. The real-time detection of anomalies in abuse videos

9. CONCLUSION

We proposed an anomaly detection model using deep learning in this work since it is an effective artificial intelligence method for categorizing videos. In order to solve the problem of anomaly detection, the 3DCNN and ConvLSTM models collaborate. We evaluated the proposed method by applying it to five large-scale datasets. The five datasets displayed excellent performance, and model training accuracy was 100%. The recognition's reliability was 98.5%, 99.2%, and 94.5%, respectively. In comparison to 3DCNN, 3DCNN+ConvLSTM performed admirably on the datasets. Our study's findings demonstrate that the model is superior to the competing models in terms of accuracy. As a continuation of our current work, we want to develop a model for anticipating anomalies in surveillance footage.





REFERENCES

- [1] B. M. Degardin, "Weakly and partially supervised learning frameworks for anomaly detection," *Engenharia Informática*, 2020, doi: 10.13140/RG.2.2.30613.65769.
- [2] P. Patel and A. Thakkar, "The upsurge of deep learning for computer vision applications," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 538–548, Feb. 2020, doi: 10.11591/ijece.v10i1.pp538-548.
- [3] B. Omarov, S. Narynov, Z. Zhumanov, A. Gumar, and M. Khassanova, "State-of-the-art violence detection techniques in video surveillance security systems: a systematic review," *PeerJ Computer Science*, vol. 8, Apr. 2022, doi: 10.7717/peerj-cs.920.
- [4] A. M. Kamoona, A. K. Gostar, A. Bab-Hadiashar, and R. Hoseinnezhad, "Sparsity-based naive bayes approach for anomaly detection in real surveillance videos," in *2019 International Conference on Control, Automation and Information Sciences (ICCAIS)*, Oct. 2019, pp. 1–6, doi: 10.1109/ICCAIS46528.2019.9074564.
- [5] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," *arXiv:1801.04264*, Jan. 2018.
- [6] L. Ye, T. Liu, T. Han, H. Ferdinando, T. Seppänen, and E. Alasaarela, "Campus violence detection based on artificial intelligent interpretation of surveillance video sequences," *Remote Sensing*, vol. 13, no. 4, Feb. 2021, doi: 10.3390/rs13040628.
- [7] H. Lv, C. Zhou, Z. Cui, C. Xu, Y. Li, and J. Yang, "Localizing anomalies from weakly-labeled videos," *IEEE Transactions on Image Processing*, vol. 30, pp. 4505–4515, 2021, doi: 10.1109/TIP.2021.3072863.
- [8] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.





- [9] J. J.-C. Wu, H.-Y. Hsieh, D.-J. Chen, C.-S. Fuh, and T.-L. Liu, "Self-supervised sparse representation for video anomaly detection," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13673, 2022, pp. 729–745.
- [10] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y.-C. Wu, "MGFN: magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection," *arXiv:2211.15098*, Nov. 2022.
- [11] E. Elsayed and D. Fathy, "Semantic deep learning to translate dynamic sign language," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 1, pp. 316–325, Feb. 2021, doi: 10.22266/ijies2021.0228.30.
- [12] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. Wong, and W. Woo, "Convolutional LSTM network: a machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems*, pp. 802–810, Jun. 2015.
- [13] P. Wu *et al.*, "Not only look, but also listen: learning multimodal violence detection under weak supervision," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12375, 2020, pp. 322–339.
- [14] B. Degardin and H. Proenca, "Human activity analysis: Iterative weak/self-supervised learning frameworks for detecting abnormal events," *IEEE/IAPR International Joint Conference on Biometrics*, 2020, doi: 10.1109/IJCB48548.2020.9304905.
- [15] M. Perez, A. C. Kot, and A. Rocha, "Detection of real-world fights in surveillance videos," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 2662–2666, doi: 10.1109/ICASSP.2019.8683676.
- [16] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: a dataset of 101 human actions classes from videos in the wild," Dec. 2012, *arXiv:1212.0402*.
- [17] P. Wu and J. Liu, "Learning causal temporal relation and feature discrimination for anomaly detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3513–3527, 2021, doi: 10.1109/TIP.2021.3062192.
- [18] K. V. Thakare, N. Sharma, D. P. Dogra, H. Choi, and I.-J. Kim, "A multi-stream deep neural network with late fuzzy fusion for real-world anomaly detection," *Expert Systems with Applications*, vol. 201, Sep. 2022, doi: 10.1016/j.eswa.2022.117030.
- [19] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 4955–4966, doi: 10.1109/ICCV48922.2021.00493.
- [20] Y. Pu and X. Wu, "Audio-guided attention network for weakly supervised violence detection," in *2022 2/textsuperscriptnd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, Jan. 2022, pp. 219–223, doi: 10.1109/ICCECE54139.2022.9712793.
- [21] J. Yu, J. Liu, Y. Cheng, R. Feng, and Y. Zhang, "Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection," in *Proceedings of the 30/textsuperscriptth ACM International Conference on Multimedia*, Oct. 2022, pp. 6278–6287, doi: 10.1145/3503161.3547868.
- [22] T. Wang *et al.*, "Generative neural networks for anomaly detection in crowded scenes," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1390–1399, May 2019, doi: 10.1109/TIFS.2018.2878538.
- [23] S. N. Gowda, M. Rohrbach, and L. Sevilla-Lara, "SMART frame selection for action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, no. 2, pp. 1451–1459, doi: 10.1609/aaai.v35i2.16235.
- [24] H. Duan, Y. Zhao, Y. Xiong, W. Liu, and D. Lin, "Omni-sourced webly-supervised learning for video recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12360, 2020, pp. 670–688.
- [25] W. Wu, Z. Sun, and W. Ouyang, "Revisiting classifier: transferring vision-language models for video recognition," *arXiv:2207.01297*, Jul. 2022.
- [26] Z. Qiu, T. Yao, C.-W. Ngo, X. Tian, and T. Mei, "Learning spatio-temporal representation with local and global diffusion," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 12048–12057, doi: 10.1109/CVPR.2019.01233.

BIOGRAPHIES OF AUTHORS







Esraa A. Mahareek     received the B.Sc. degree in computer science, in 2012, the M.Sc. degree in computer science, in 2021. She is currently a teaching assistant in computer science at the Mathematics Department, Faculty of Science, Al-Azhar University, Cairo, Egypt. She has published research paper in the field of AI, machine learning, metaheuristic optimization. She can be contacted at email: esraa.mahareek@azhar.edu.eg.







Eman K. Elsayed     Prof. Eman K. Elsayed Dean of Canadian International College School of Computer Science, Bachelor of Science from Computer Science Department, Cairo University 1994, Master of Computer Science from Cairo university 1999, Computer Science Ph.D. 2005 from Al-Azhar University, Professor of Computer Science from 2019. She published 65 papers until Mars 2023 in different branches of AI. She can be contacted at email: eman_k_elsayed@cic-cairo.com.



Nahed M. EIDesouky     is an associate professor of computer science and information systems at Al-Azhar University (girls) in Cairo, Egypt. Bachelor of Science from the Faculty of Engineering, Cairo University in 1982, Master of Communication and Electronics from the Faculty of Engineering Cairo University 1990, communication and electronics Ph.D. 1999 from the Faculty of Engineering at Cairo University. Associate professor of computer science from 2010. She published 23 papers in different branches of computer science; security, cloud computing and optimization. She can be contacted at email: nahedeldesouky5922@azhar.edu.eg.



Kamal A. Eldahshan     is a professor of Computer Science and Information Systems at Al-Azhar University in Cairo, Egypt. At Al-Azhar, he founded the Centre of Excellence in Information Technology, in collaboration with the Indian government, and was also the founder and former president of the coordination bureau of the Egyptian Knowledge Bank, the country's largest initiative for academic access. An Egyptian national and graduate of Cairo University, he obtained his doctoral degree from the Université de Technologie de Compiègne in France, where he also taught for several years. During his extended stay in France, he also worked at the prestigious Institut National de Télécommunications in Paris. Professor El-Dahshan's extensive international research, teaching, and consulting experience has spanned four continents and include academic institutions as well as government and private organizations. He taught at Virginia Tech as a visiting professor; he was a Consultant to the Egyptian Cabinet Information and Decision Support Centre (IDSC); and he was a senior advisor to the Ministry of Education and Deputy Director of the National Technology Development Centre. Professor ElDahshan is a professional Fellow on Open Educational Resources (OER) as recognized by the United States Department of State and an Expert with the Arab League Educational, Cultural and Scientific Organization. A tireless advocate for equitable access to knowledge too all, he co-founded, in 2018, the Arab Foundation for the Deaf and Hearing Impaired, which aims at supporting and empowering its beneficiaries to contribute to the public, scholarly, and cultural lives of their communities, as equals. Among other accolades, he is a Fellow of the British Computing Society, and a founding member of the Egyptian Mathematical Society. She can be contacted at email: dahshan@gmail.com.