# Violence Detection in Surveillance Videos Using Artificial Intelligence

**Atharva Wankhade, prof Snehil Jaiswal, Snehal Tingane**

*Abstract*— **In recent years, the role of monitoring systems in upholding public order and ensuring public safety has expanded. Video monitoring in locations such as train stations, schools, and hospitals necessitates the automatic identification of hostile and suspicious behaviours to avert any potential harm to society, the economy, and the environment. The optimal utilization of computerized violence detection by law enforcement is crucial. An all-encompassing methodology is necessary to identify instances of violence and the use of weapons in closed circuit television (CCTV) recordings. This work presents the Smart-City CCTV Violence Detection (SCVD) dataset, which is used for detecting occurrences of violence in surveillance films, including both cases including weapons and those without weapons. The proposed system utilizes a model based on deep learning approach to detect violence in videos. The model uses a comprehensive set of characteristics to describe violence occurrences in the dataset. The model was created using MATLAB software, utilizing image processing and machine and deep learning toolkit. It achieved an overall accuracy rate of 96.4%. The proposed endeavour has practical applications in the industry and is advantageous to society in terms of security**

*Index Terms*— **Violence Detection, Weaponized Violence Detection, Video Surveillance, Machine Learning, Deep Learning**

## I. INTRODUCTION

The presence of violence and gang-related activity in a city can pose a significant danger, especially when authorities are unable to intervene to mitigate further harm promptly. Occasionally, these occurrences can lead to fatalities and damage to belongings, particularly when weapons are present [1]. Unfortunately, instances of road rage, violence associated with gangs, and other impulsive acts of violent crime often happen suddenly and without any advance notice or the opportunity for authorities to take pre-emptive measures. These occurrences present a significant obstacle for law enforcement agencies and other pertinent entities. Regrettably, the documentation of such occurrences frequently takes place retrospectively, resulting in authorities having limited choices for prompt intervention and efficient prevention [2]. Despite the assistance provided by surveillance systems in identifying perpetrators and offenders through recordings, detecting, searching, and apprehending individuals after a crime has been committed often proves to be time-consuming. There is an increasing demand for automated detection and signalling systems to decrease the amount of time it takes to complete a task and improve overall effectiveness.

After the significant advancement of deep learning in the ImageNet 2012 competition, deep neural networks (DNNs) have emerged as the preferred artificial intelligence method for automating these jobs. Through the utilization of Deep Neural Networks (DNNs) and other Artificial Intelligence (AI) methodologies, smart cities across the globe can enhance their ability to identify and react to instances of violence, thereby ensuring the protection of both individuals and assets. The advantages of these technologies are evident. For example, integrated surveillance systems coupled with sophisticated AI algorithms may analyse live video feeds from CCTV cameras to detect and notify authorities of potential violent occurrences, facilitating prompt action. Moreover, AI-driven predictive analytics may examine diverse data sources, such as social media feeds and sensor data, to detect patterns and trends linked to violence. This allows authorities to strategically allocate resources and proactively prevent instances of violence in particular regions. As these technologies progress, they will have a growing significance in ensuring safety and security in our urban areas [4-6]. Existing violence detection techniques mostly depend on spatiotemporal models to recognize occurrences of violent actions in video footage. Nevertheless, it is crucial to acknowledge that violence spans a broad spectrum of behaviours, ranging from physical altercations to gunfights. Applying identical treatment to all instances of violence may not adequately prioritize the level of seriousness or possible damage involved [7] [9-10]. To tackle this difficulty, it is imperative to devise techniques for identifying firearms in surveillance films.

The ability to detect and analyse events in video feeds has seen significant advancements in recent years, mostly due to the increasing prevalence of acts of human violence in our daily lives. The surveillance film is often identified using manual detection. There are several cameras distributed worldwide, but the incidence of human violence may be relatively low, and the potential risks can occur anywhere. This text aims to assess the present state of human-violence systems and explore the various deep-learning approaches and methodologies employed in analysing them. This technology is designed to identify and track objects, motions, and activities. By analysing the data collected, we may integrate this technology to detect instances of human violence that occur daily. Hence, there is an urgent requirement for additional investigation and progress in open-world weapons identification, which can effectively recognize and categorize a wider array of objects that could potentially be used as weapons. This study specifically examines cases of violence detection, both including weapons and not involving weapons. The findings of this study have

Manuscript received May 15, 2024

**Atharva Wankhade,** G H Raisoni University Amravati
**prof Snehil Jaiswal,** G H Raisoni University Amravati
**Snehal Tingane,** G H Raisoni University Amravati

possible applications in various real-life situations, such as aiding security workers or facilitating emergency calls.

The paper structured into the five sections. Section 2 describes about the related work regarding the existing violence detection system. The proposed methodology with detail implementation is given in the section 3. Further, the experimental results are evaluated and analysed in the section 4. Finally, section 5 depicted the conclusion of the article.

## II. RELATED WORK

The existing violence detection system uses the intelligent learning approach based on various circumstances as discussed further. B Omarov et.al. [1] aims to address the problems as state-of-the-art methods in video violence detection, datasets to develop and train real-time video violence detection frameworks, discuss and identify open issues in the given problem. G. Sreenu et.al. [2] included a deep-rooted survey which starts from object recognition, action recognition, crowd analysis and finally violence detection in a crowd environment. Eknarin Ditsanthia et.al. [3] proposed a novel representation learning approach to improve the detection rate of violent behaviours in videos in which multiscale convolutional features extracted from image-based deep convolution neural network to describe spatial information in a video frame. Toluwani Aremu et.al. [4] introduced the Smart-City CCTV Violence Detection (SCVD) dataset, specifically designed to facilitate the learning of weapon distribution in surveillance videos with novel technique called, SSIVD-Net (Salient-Super-Image for Violence Detection). Ali Mansour Al-Madani et.al. [5] presented a 3D convolutional neural network-based technique for detecting violence in videos with performance evaluations shown that the suggested technique effectively identifies violence in video clips. S A Arun Akash et.al. [6] used deep learning as computer vision to predict and detect the action, properties from video with detection of violent acts from CCTV cameras using the Inception V3 and Yolo V5 models. Bermejo Nievas et.al. [7] introduced a new video database containing 1000 sequences divided in two groups: fights and non-fights. Experiments on this database and another one with fights from action movies show that fights can be detected with near 90% accuracy. Aqib Mumtaz et.al. [8] proposed deep representation-based model using concept of transfer learning for violent scenes detection to identify aggressive human behaviours. J.V. Vidhya et.al. [9] converted the given input video into frames and the preprocessing is done at the frame level based on feature extraction using the 2D convolutional neural network (Conv2D) which adapts the layers of VGG-19 net architecture with global average pooling and learns the spatial information in the given video. Muzamil Ahmed et.al. [10] proposed framework, the keyframe extraction technique eliminates duplicate consecutive frames which reduces the training data size and hence decreases the computational cost by avoiding duplicate frames.

Romas Vijeikis et.al. [11] presented a novel architecture for violence detection from video surveillance cameras with a spatial feature extracting a U-Net-like network that uses MobileNet V2 as an encoder followed by LSTM for temporal feature extraction and classification. A. Z. Kouzani et.al. [12] examined the latest technological innovations for tackling domestic violence which describes a range of technology platforms and tools, and discusses their capabilities and shortcomings. J. Su et.al. [13] show competitive violence detection results using a general action recognition CNN without modifying the original architecture. Elly Matul Imah et.al. [14] presented violence detection using the visual geometry group network-16 (VGGNet-16)-based deep transfer learning feature extraction, combined with ensemble decision fusion learning. Yassine Himeur et.al. [15] introduced, to the best of the authors' knowledge, the first overview of existing DTL- and DDA-based video surveillance to shed light on their benefits, discuss their challenges, and highlight their future perspectives. Guillermo Garcia-Cobo et.al. [16] proposed a novel deep learning architecture that accurately and efficiently detects violent crimes in surveillance videos. Seymanur Akt et.al. [17] explored LSTM based approaches to solve it based on new dataset, which consists of fight scenes from surveillance camera videos available at YouTube. PAJON Quentin et.al. [18] presented an investigation into machine learning techniques for violence detection in videos and their adaptation to a federated learning context. Ali Mansour Al-Madani et.al. [19] presents a 3D convolutional neural network-based technique for detecting violence in videos. Accuracy is improved by employing machine and deep learning techniques in a suggested manner. Ji Li et.al. [20] propose a deep learning model based on 3D convolutional neural networks, without using hand-crafted features or RNN architectures exclusively for encoding temporal information.

Balika J. Chelliah et.al. [21] tackles the difficult subject of detecting violence in videos. Unlike previous work focusing on merging multimodal features, visual subtypes connected to violence. Mai Magdy et.al. [22] presented a deep learning architecture in this study using four-dimensional video-level convolution neural networks includes residual blocks that are used with three-Dimensional Convolution Neural Networks 3D (CNNs) to learn long-term and short-term spatiotemporal representation from the video as well as record inter-clip interaction. Fath U Min Ullah et.al. [23] proposed a triple-staged end-to-end deep learning violence detection framework in which persons are detected in the surveillance video stream using a light-weight convolutional neural network (CNN) model to reduce and overcome the voluminous processing of useless frames. Romas Vijeikis et.al. [24] presented a novel architecture for violence detection from video surveillance cameras based on a spatial feature extracting a U-Net-like network that uses MobileNet V2 as an encoder followed by LSTM for temporal feature extraction and classification. David Choqueluque-Roman et.al. [25] proposing a weakly supervised method to detect spatially and temporarily violent actions in surveillance videos using only video-level labels which follows a Fast-RCNN style architecture, that has been temporally extended. Xiang Wang et.al. [26] proposed a two-stream deep learning architecture for video violent activity detection named SpikeConvFlowNet in which RGB frames and their optical flow data are used as inputs for each stream to extract the spatiotemporal features of videos and fed to the classifier for the final decision. Shakil Ahmed Sumon et.al. [27] explored different strategies to find out the saliency of the features from different pretrained models in detecting violence in videos from a dataset which has been created which consists of violent and non-violent videos of different settings.

## III. PROPOSED WORK

The entire working process of the presented method is shown in Figure 1. As shown in figure, the presented model consists a series of processes which are discussed in the figure.

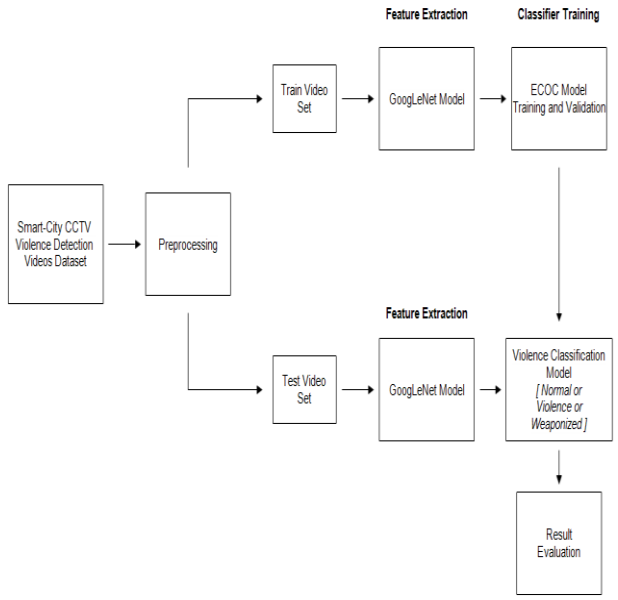

Feature Extraction
Classifier Training
Train Video Set
GoogLeNet Model
ECOC Model Training and Validation
Smart-City CCTV Violence Detection Videos Dataset
Preprocessing
Feature Extraction
Test Video Set
GoogLeNet Model
Violence Classification Model [ Normal or Violence or Weaponized ]
Result Evaluation

Figure1: Proposed methodology of Violence Detection System

***Violence Video Dataset***

In this scenario, standard benchmark of Smart-City CCTV Violence Detection video database [4], which is publicly available is used for this experimentation. The Smart-City CCTV Violence Detection dataset (SCVD) is a recently developed benchmark. The existing datasets, such as the NTU CCTV-Fights dataset and the Real-Life Violence Situations dataset (RLVS), along with other commonly utilized datasets for violence detection, consist of films captured by phone cameras. This can potentially distort the distribution and emphasis of violence detection based on CCTV footage. In addition, our dataset includes a category specifically designed for identifying weapons in films, making it the pioneering dataset for weapons detection in video format. This is apart from other datasets for weapons recognition, which mostly consist of photos featuring guns and knives. The SCVD dataset is designed to account for the possibility that any handheld object capable of causing harm to individuals or property may be considered a weapon. The ultimate dataset comprises three distinct categories: Violence (V), Normal (N), and weaponized violence (WV). The video dataset is divided into separate train and test folders, as specified in the dataset, to create training and testing sets of films for the proposed modelling.

***Pre-processing***

The presented input movies are utilized as input for the described model. The submitted input movies are used as input for the described model. During the initial phase, pre-processing occurs by resizing each frame in the video according to the dimensions of the learned model.

Subsequently, frame segmentation is conducted to identify and pick the frames that pertain to significant instances of violence and non-violence from a particular set of video samples. When analysing a video clip, it is conceivable that instances of violent activity, which is the focus of our investigation, occur only in certain parts of the clip. These instances are particularly noticeable in specific frames where the movement of a person is observed, indicating their involvement in the violent or other actions. It is well observed that human actions are always accompanied by motion. To identify frames that include significant instances of violent data, the system uses Peak Signal Noise Ratio (PSNR) estimation between two consecutive frames.

If the video sequence being analysed consists of 'n' frames, we shall transmit all the frames from frame number 1 to n to the frame fragment selection unit, using the current and next frames. When the current frame is 'i', the unit will initially calculate the Mean Square Error between frame 'i' and frame 'i+1'. When the system is provided with video in RGB format, there will be more than two channels available in each frame. In these instances, the Peak Signal to Noise Ratio (PSNR) will be computed for each channel of two consecutive frames being analyzed, and the average PSNR of all channels will be regarded as the PSNR value of the two frames. Afterward, a series of PSNR values for each frame will be generated. From this series, frames that include significant events can be chosen. PSNR is determined by calculating the mean square error (MSE), as seen below.

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

$$MSE = \frac{1}{m\,n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2.$$

Here, MAXi is the maximum possible pixel value of the image. When the pixels are represented using 8 bits per sample, this is 255. This novel frame selection unit drastically reduces the number of frames which are considered for feature extraction and as a result the computational cost is reduced.

***Feature Extraction***

During the subsequent phase, a set of significant characteristics is derived from the divided image, specifically from each video frame, utilizing convolutional neural network models that are built upon deep learning principles. The deep learning model is built around a convolutional neural network with multiple deep layers that capture a diverse range of features. The model was selected based on its exceptional categorization performance.

GoogLeNet is a convolutional neural network that utilizes the Inception architecture. The network employs Inception modules, enabling it to select between several convolutional filter sizes inside each block. An Inception network vertically arranges these modules, occasionally incorporating max-pooling layers with a stride of 2 to reduce the resolution of the grid by half.

GoogLeNet has undergone training using a dataset of more than one million photos. It possesses the ability to categorize photographs into 1000 distinct object categories, including but not limited to keyboards, coffee mugs, pencils, and various animals. The network has acquired complex and detailed feature representations for a diverse set of images. The network receives an image as input and produces a label for the object in the image, along with the corresponding probabilities for each object category.

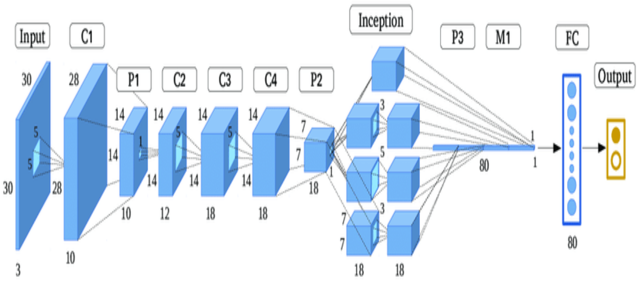


Figure 2: GoogLeNet pretrained CNN Architecture

The primary breakthrough of GoogleNet is in the utilization of a structural design known as Inception. Generally, Inception is a hierarchical network architecture where the ideal local sparse structure of a visual network is spatially replicated from the beginning to the conclusion. Three Inception structures utilized in various situations are presented: commonly, a 1 x 1 convolution is employed in Inception to calculate reductions before to the resource-intensive 3 x 3 and 5 x 5 convolutions.

GoogleNet serves as a source of inspiration for constructing a highly capable architecture. The advancement of identification capabilities mostly depends on novel concepts, algorithms, and enhanced network topologies, in addition to more powerful hardware, huge datasets, and bigger models. In this project, we constructed a structure similar to GoogleNet by studying and drawing inspiration from it. The structure is depicted in Figure 2. We developed the Inception architecture by optimizing the quantity and arrangement of layers and filters.

### *Classifier Training and Testing*

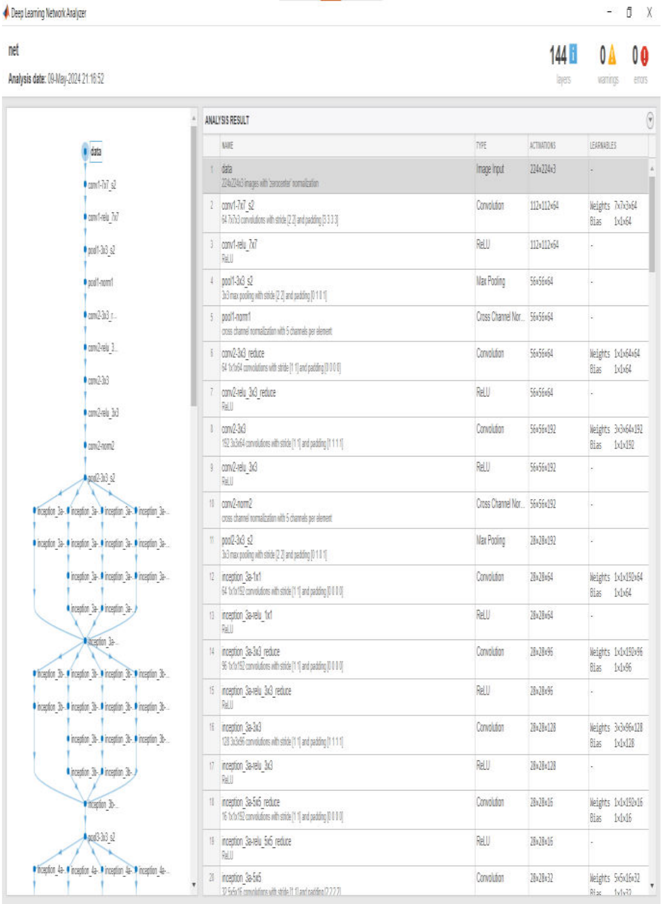
Next, the videos will be classified using a machine learning classification model. The algorithm will ultimately categorize the videos into three groups: 'weaponized', 'violence', or 'normal' occurrences. The feature extraction procedure involves applying image processing techniques, while the classification operation utilizes machine learning algorithms. This enables the development of trained prediction models using the filtered features in a more efficient and rapid manner.

Categorization ECOC is a multiclass learning classifier called error-correcting output codes (ECOC). It is composed of several binary learners, such as support vector machines (SVMs). The Error-Correcting Output Codes (ECOC) is a type of ensemble classifier that is used to solve issues with several classes. It operates by representing each class as a binary code. The ECOC (Error Correcting Output Codes) has served as the primary theoretical basis for output coding approaches. These methods aim to break down a complex multi-class problem into a series of binary problems. Subsequently, the outputs of binary classifiers are reconstructed for each binary problem. The effectiveness of output coding methods is contingent upon the performance of the underlying binary classifiers. There is a need to reexamine the ECOC concept in light of the availability of Support Vector Machines (SVM), which can generate a sophisticated nonlinear decision boundary and exhibit strong generalization performance. SVM can serve as a suitable base classifier for output coding methods.

## IV. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed work utilizes an Intel CORE i5 processor, 16GB of RAM, and operates on the Windows 11 operating system. The programming code in this project was written using MATLAB R2018b software, which utilized the Image Processing, Statistics, Machine Learning, and Deep Learning toolboxes. The train and test movies utilized for testing are sourced from the SCVD dataset, a standard benchmark dataset consisting of real-life recordings depicting violent events [4].

In this experiment, following the suggested block structure, we have two stages of implementation: training and testing. During the training phase, the train set of videos must be preprocessed to match the dimensions of the deep network being used. For the feature extraction procedure of the training photos, we employed an automated feature extraction method based on the GoogLeNet pre-trained deep convolutional neural network. This network comprises a total of 144 levels, including the input, feature, classification, and output layers, as illustrated in Figures 3 and 4. We utilized the 'pool5-7x7_s1' feature layer to extract features from video frames. After extracting the features, the next step is to train the model using an ensemble learning method. The algorithm will be based on the input and output data, where the input is the dataset of video features for training and the output is the corresponding labels. Once the training model was successfully validated, we proceeded to store the trained model.



net

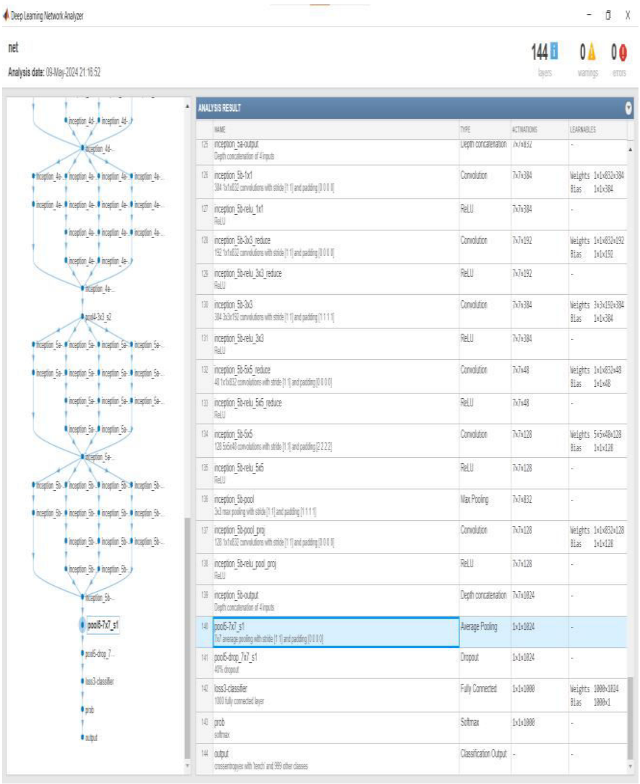
Analysis date: 09-May-2024 21:16:52

144 layers 0 warnings 0 errors

ANALYSIS RESULT

| | NAME | TYPE | ACTIVATIONS | LEARNABLES |
|---|---|---|---|---|
| 1 | data<br>224x224x3 images with 'zerocenter' normalization | Image Input | 224x224x3 | - |
| 2 | conv1-7x7_s2<br>64 7x7x3 convolutions with stride [2 2] and padding [3 3 3 3] | Convolution | 112x112x64 | Weights 7x7x3x64<br>Bias 1x1x64 |
| 3 | conv1-relu_7x7<br>ReLU | ReLU | 112x112x64 | - |
| 4 | pool1-3x3_s2<br>3x3 max pooling with stride [2 2] and padding [0 1 0 1] | Max Pooling | 56x56x64 | - |
| 5 | pool1-norm1<br>cross channel normalization with 5 channels per element | Cross Channel Nor... | 56x56x64 | - |
| 6 | conv2-3x3_reduce<br>64 1x1x64 convolutions with stride [1 1] and padding [0 0 0 0] | Convolution | 56x56x64 | Weights 1x1x64x64<br>Bias 1x1x64 |
| 7 | conv2-relu_3x3_reduce<br>ReLU | ReLU | 56x56x64 | - |
| 8 | conv2-3x3<br>192 3x3x64 convolutions with stride [1 1] and padding [1 1 1 1] | Convolution | 56x56x192 | Weights 3x3x64x192<br>Bias 1x1x192 |
| 9 | conv2-relu_3x3<br>ReLU | ReLU | 56x56x192 | - |
| 10 | conv2-norm2<br>cross channel normalization with 5 channels per element | Cross Channel Nor... | 56x56x192 | - |
| 11 | pool2-3x3_s2<br>3x3 max pooling with stride [2 2] and padding [0 1 0 1] | Max Pooling | 28x28x192 | - |
| 12 | inception_3a-1x1<br>64 1x1x192 convolutions with stride [1 1] and padding [0 0 0 0] | Convolution | 28x28x64 | Weights 1x1x192x64<br>Bias 1x1x64 |
| 13 | inception_3a-relu_1x1<br>ReLU | ReLU | 28x28x64 | - |
| 14 | inception_3a-3x3_reduce<br>96 1x1x192 convolutions with stride [1 1] and padding [0 0 0 0] | Convolution | 28x28x96 | Weights 1x1x192x96<br>Bias 1x1x96 |
| 15 | inception_3a-relu_3x3_reduce<br>ReLU | ReLU | 28x28x96 | - |
| 16 | inception_3a-3x3<br>128 3x3x96 convolutions with stride [1 1] and padding [1 1 1 1] | Convolution | 28x28x128 | Weights 3x3x96x128<br>Bias 1x1x128 |
| 17 | inception_3a-relu_3x3<br>ReLU | ReLU | 28x28x128 | - |
| 18 | inception_3a-5x5_reduce<br>16 1x1x192 convolutions with stride [1 1] and padding [0 0 0 0] | Convolution | 28x28x16 | Weights 1x1x192x16<br>Bias 1x1x16 |
| 19 | inception_3a-relu_5x5_reduce<br>ReLU | ReLU | 28x28x16 | - |
| 20 | inception_3a-5x5<br>32 5x5x16 convolutions with stride [1 1] and padding [2 2 2 2] | Convolution | 28x28x32 | Weights 5x5x16x32<br>Bias 1x1x32 |

Figure 3: Initial layers of GoogLeNet architecture



net

**Analysis date:** 09-May-2024 21:16:52

ANALYSIS RESULT

| | NAME | TYPE | ACTIVATIONS | LEARNABLES |
|---|---|---|---|---|
| 125 | inception_5a-output<br>Depth concatenation of 4 inputs | Depth concatenation | 7x7x832 | - |
| 126 | inception_5b-1x1<br>384 1x1x832 convolutions with stride [1 1] and padding [0 0 0 0] | Convolution | 7x7x384 | Weights 1x1x832x384<br>Bias 1x1x384 |
| 127 | inception_5b-relu_1x1<br>ReLU | ReLU | 7x7x384 | - |
| 128 | inception_5b-3x3_reduce<br>192 1x1x832 convolutions with stride [1 1] and padding [0 0 0 0] | Convolution | 7x7x192 | Weights 1x1x832x192<br>Bias 1x1x192 |
| 129 | inception_5b-relu_3x3_reduce<br>ReLU | ReLU | 7x7x192 | - |
| 130 | inception_5b-3x3<br>384 3x3x192 convolutions with stride [1 1] and padding [1 1 1 1] | Convolution | 7x7x384 | Weights 3x3x192x384<br>Bias 1x1x384 |
| 131 | inception_5b-relu_3x3<br>ReLU | ReLU | 7x7x384 | - |
| 132 | inception_5b-5x5_reduce<br>48 1x1x832 convolutions with stride [1 1] and padding [0 0 0 0] | Convolution | 7x7x48 | Weights 1x1x832x48<br>Bias 1x1x48 |
| 133 | inception_5b-relu_5x5_reduce<br>ReLU | ReLU | 7x7x48 | - |
| 134 | inception_5b-5x5<br>128 5x5x48 convolutions with stride [1 1] and padding [2 2 2 2] | Convolution | 7x7x128 | Weights 5x5x48x128<br>Bias 1x1x128 |
| 135 | inception_5b-relu_5x5<br>ReLU | ReLU | 7x7x128 | - |
| 136 | inception_5b-pool<br>3x3 max pooling with stride [1 1] and padding [1 1 1 1] | Max Pooling | 7x7x832 | - |
| 137 | inception_5b-pool_proj<br>128 1x1x832 convolutions with stride [1 1] and padding [0 0 0 0] | Convolution | 7x7x128 | Weights 1x1x832x128<br>Bias 1x1x128 |
| 138 | inception_5b-relu_pool_proj<br>ReLU | ReLU | 7x7x128 | - |
| 139 | inception_5b-output<br>Depth concatenation of 4 inputs | Depth concatenation | 7x7x1024 | - |
| 140 | pool5-7x7_s1<br>7x7 average pooling with stride [1 1] and padding [0 0 0 0] | Average Pooling | 1x1x1024 | - |
| 141 | pool5-drop_7x7_s1<br>40% dropout | Dropout | 1x1x1024 | - |
| 142 | loss3-classifier<br>1000 fully connected layer | Fully Connected | 1x1x1000 | Weights 1000x1024<br>Bias 1000x1 |
| 143 | prob<br>softmax | Softmax | 1x1x1000 | - |
| 144 | output<br>crossentropyex with 'tench' and 999 other classes | Classification Output | - | - |

Figure 4: Final layers of GoogLeNet architecture

Figure 5 to figure 7 depicts the sample frames of test videos in normal, violence and weaponized violence condition from SCVD dataset. SCVD dataset consists of 2746 number of sample videos while training and 477 number of sample videos while testing among all three classes. Initially, all train videos are trained and validated. After successful validation, all test sample videos are tested to predict the respective output class.

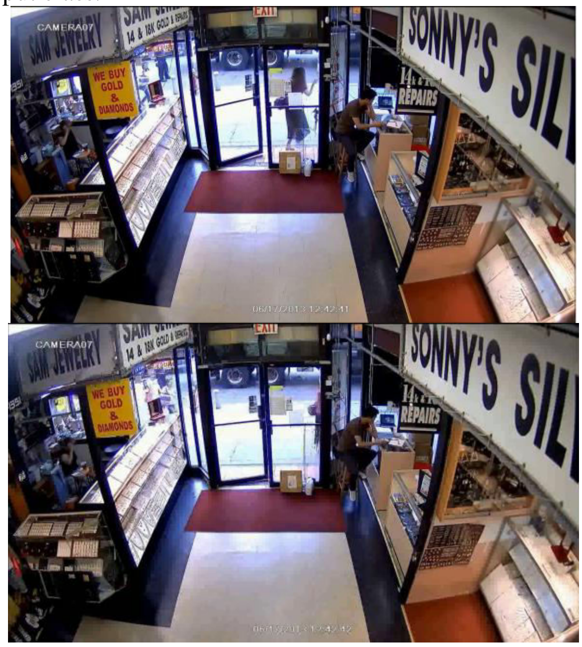


Figure 5: Sample frames of sample test video of Normal condition

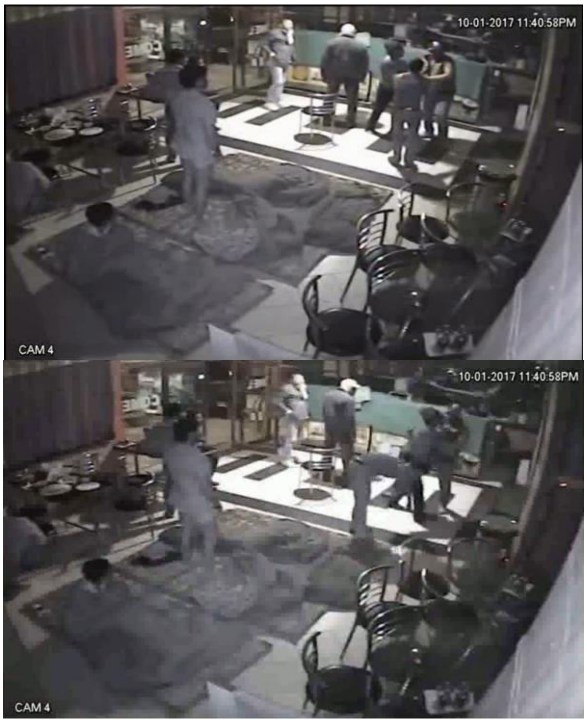


Figure 6: Sample frames of sample test video of Violence condition

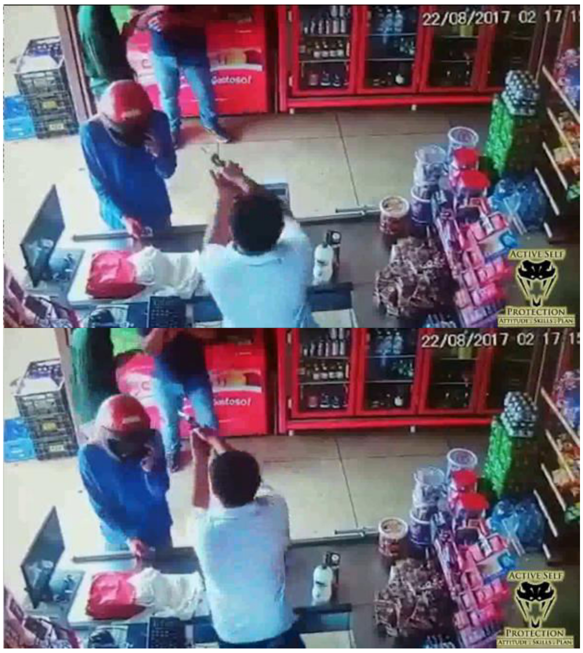
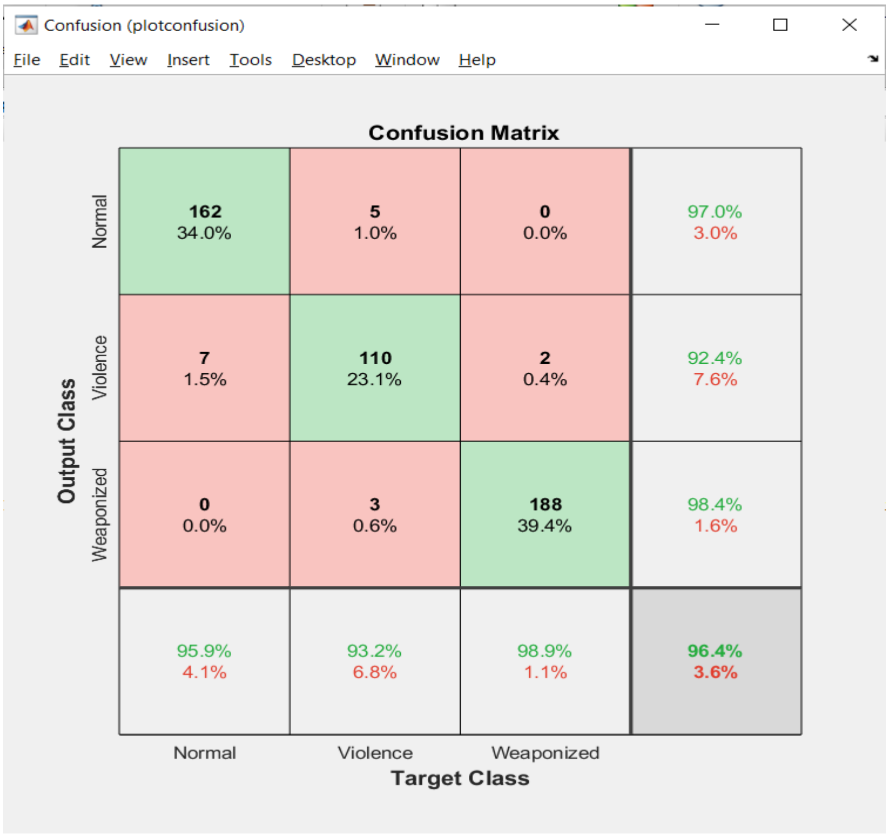


Figure 7: Sample frames of sample test video of Weaponized Violence condition

As per the final testing of all sample test videos, confusion matrix is plotted to display the how much number of samples are correctly classified or mis-classified as per the actual output class of videos. Figure 8 depicts the confusion matrix of all test videos, it shows the relationship between the actual output and predict output for three output classes, normal, violence and weaponized violence. It shows overall accuracy rate of 96.4% for all three output classes.


Confusion (plotconfusion)
File Edit View Insert Tools Desktop Window Help

Confusion Matrix

| Output Class / Target Class | Normal | Violence | Weaponized | |
|---|---|---|---|---|
| Normal | 162<br>34.0% | 5<br>1.0% | 0<br>0.0% | 97.0%<br>3.0% |
| Violence | 7<br>1.5% | 110<br>23.1% | 2<br>0.4% | 92.4%<br>7.6% |
| Weaponized | 0<br>0.0% | 3<br>0.6% | 188<br>39.4% | 98.4%<br>1.6% |
| | 95.9%<br>4.1% | 93.2%<br>6.8% | 98.9%<br>1.1% | 96.4%<br>3.6% |

Figure 8: Confusion matrix of testing phase

The performance of overall system as per the three output classes are calculated based on confusion matrix parameters and shown in table 1 in terms of precision, recall and f-score parameter. It shows that weaponized violence output class shows the best accuracy rate of 98.95% among all three classes with higher precision, recall and f-score rate as defined in figure 9.

Table 1: Performance Evaluation Results

| Classes/Parameters | Accuracy (%) | Precision (%) | Recall (%) | F-score (%) |
|---|---|---|---|---|
| Normal | 97.484 | 97.006 | 95.858 | 96.774 |
| Violence | 96.436 | 92.437 | 93.220 | 92.592 |
| Weaponized | 98.951 | 98.429 | 98.947 | 98.532 |

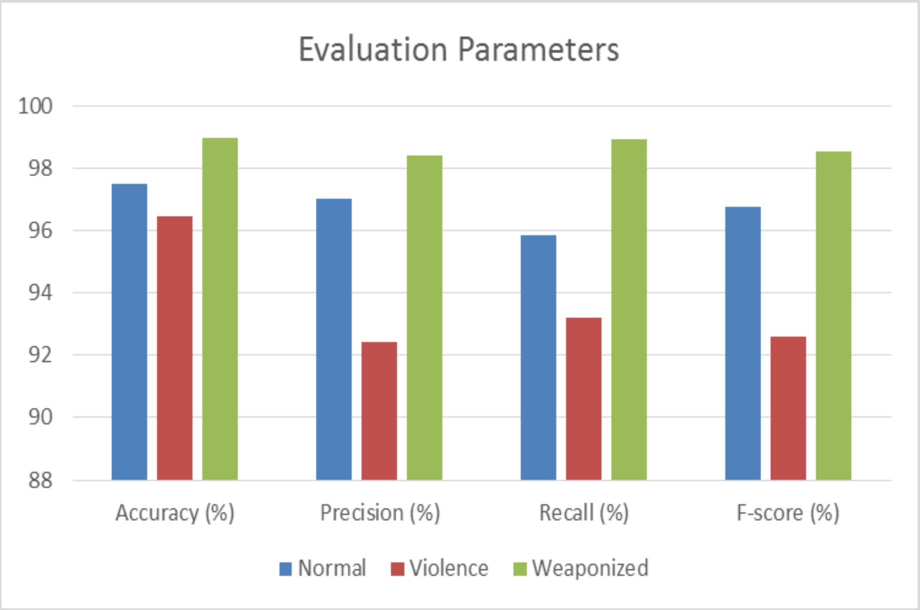


Figure 9: Result performance parameters of system

## CONCLUSION

This article presents a highly effective technique for identifying instances of violent activities including weapons in surveillance films captured in smart cities. This work thoroughly investigates the utilization of extracting prominent characteristics from the frames, which are then employed in the detection of violence in videos. The GoogLeNet deep learning models were tested using convolutional neural networks. The features obtained from each frame have been inputted into a fully linked neural network. The ECOC classifier accurately categorized the events in the movie into three distinct classes: normal, violent, or weaponized violence. This classification was based on the extracted deep features of the test videos. Therefore, the suggested approach demonstrates a superior overall accuracy rate of 96.4% and has the potential to be enhanced even further through the utilization of supervised machine learning or deep learning algorithms.

## REFERENCES

[1] Omarov B, Narynov S, Zhumanov Z, Gumar A, Khassanova M. State-of-the-art violence detection techniques in video surveillance security systems: a systematic review. PeerJ Comput Sci. 2022 Apr 6;8:e920. doi: 10.7717/peerj-cs.920. PMID: 35494848; PMCID: PMC9044356.

[2] Sreenu, G., Saleem Durai, M.A. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. J Big Data 6, 48 (2019). https://doi.org/10.1186/s40537-019-0212-5

[3] E. Ditsanthia, L. Pipanmaekaporn and S. Kamonsantiroj, "Video Representation Learning for CCTV-Based Violence Detection," 2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON), Bangkok, Thailand, 2018, pp. 1-5, doi: 10.1109/TIMES-iCON.2018.8621751.

[4] Aremu, Toluwani & Li, Zhiyuan & Alameeri, Reem & Khan, Mustaqeem & El Saddik, Abdulmotaleb. (2023). SSIVD-Net: A Novel Salient Super Image Classification & Detection Technique for Weaponized Violence. 10.48550/arXiv.2207.12850.

[5] Ali Mansour Al-Madani, et.al. “Real-Time Detection of Crime and Violence in Video Surveillance using Deep Learning”, 2022 Proceedings of the First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022), pp. 431-441, https://doi.org/10.2991/978-94-6463-196-8_33.

[6] Akash, S & Moorthy, R & Esha, K & Narayanaraju, Nathiya. (2022). Human Violence Detection Using Deep Learning Techniques. Journal of Physics: Conference Series. 2318. 012003. 10.1088/1742-6596/2318/1/012003.

[7] Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R. (2011). Violence Detection in Video Using Computer Vision Techniques. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds) Computer Analysis of Images and Patterns. CAIP 2011. Lecture Notes in Computer Science, vol 6855. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23678-5_39

[8] Mumtaz, A. B. Sargano and Z. Habib, "Violence Detection in Surveillance Videos with Deep Network Using Transfer Learning," 2018 2nd European Conference on Electrical Engineering and Computer Science (EECS), Bern, Switzerland, 2018, pp. 558-563, doi: 10.1109/EECS.2018.00109.

[9] Vidhya, J. V. and R. Annie Uthra. “Violence detection in videos using Conv2D VGG-19 architecture and LSTM network.” (2021).

[10] Ramzan, Muhammad & Khan, Hikmat & Iqbal, Saqib & Khan, Muhammad & Choi, Jungin & Nam, Yunyoung & Kadry, Seifedine. (2021). Real-Time Violent Action Recognition Using Key Frames Extraction and Deep Learning. Computers, Materials and Continua. 69. 2217-2230. 10.32604/cmc.2021.018103.

[11] Vijeikis, Romas, Vidas Raudonis, and Gintaras Dervinis. 2022. "Efficient Violence Detection in Surveillance" Sensors 22, no. 6: 2216. https://doi.org/10.3390/s22062216.

[12] Z. Kouzani, "Technological Innovations for Tackling Domestic Violence," in IEEE Access, vol. 11, pp. 91293-91311, 2023, doi: 10.1109/ACCESS.2023.3306022.

[13] J. Su, P. Her, E. Clemens, E. Yaz, S. Schneider and H. Medeiros, "Violence Detection using 3D Convolutional Neural Networks," 2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Madrid, Spain, 2022, pp. 1-8, doi: 10.1109/AVSS56176.2022.9959393.

[14] Elly Matul Imah et.al. “Child Violence Detection in Surveillance Video Using Deep Transfer Learning and Ensemble Decision Fusion Learning”, 2022, International Journal of Intelligent Engineering and Systems, Vol.15, No.3. DOI: 10.22266/ijies2022.0630.38.

[15] Yassine Himeur, Somaya Al-Maadeed, Hamza Kheddar, Noor Al-Maadeed, Khalid Abualsaud, Amr Mohamed, Tamer Khattab, Video surveillance using deep transfer learning and deep domain adaptation: Towards better generalization, Engineering Applications of Artificial Intelligence, Volume 119, 2023, 105698, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2022.105698.

[16] Guillermo Garcia-Cobo, Juan C. SanMiguel, Human skeletons and change detection for efficient violence detection in surveillance videos, Computer Vision and Image Understanding, Volume 233, 2023, 103739, ISSN 1077-3142, https://doi.org/10.1016/j.cviu.2023.103739.

[17] Akti, Seymanur et al. “Vision-based Fight Detection from Surveillance Cameras.” 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA) (2019): 1-6.

[18] Pajon Quentin, Serre Swan, Wissocq Hugo, Rabaud Léo, Haidar Siba, Yaacoub Antoun, “Balancing Accuracy and Training Time in Federated Learning for Violence Detection in Surveillance Videos: A Study of Neural Network Architectures”, https://doi.org/10.48550/arXiv.2308.05106.

[19] Ali Mansour Al-Madani and Vivek Mahale and Ashok T. Gaikwad, “Real-Time Detection of Crime and Violence in Video Surveillance using Deep Learning”, 2023, Proceedings of the First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022), pp. 431-441, https://doi.org/10.2991/978-94-6463-196-8_33.

[20] Li, Ji & Jiang, Xinghao & Sun, Tanfeng & xu, ke. (2019). Efficient Violence Detection Using 3D Convolutional Neural Networks. 1-8. 10.1109/AVSS.2019.8909883.

[21] Balika J. Chelliah, K. Harshitha, and Saharsh Pandey, “Adaptive and effective spatio-temporal modelling for offensive video classification using deep neural network”, International Journal of Intelligent Engineering Informatics 2023 11:1, pp. 19-34.

[22] Magdy, Mai & Fakhr, Mohamed & Maghraby, Fahima. (2022). Violence 4D: Violence detection in surveillance using 4D convolutional neural networks. IET Computer Vision. 17. n/a-n/a. 10.1049/cvi2.12162.

[23] Ullah, Fath U Min, Amin Ullah, Khan Muhammad, Ijaz Ul Haq, and Sung Wook Baik. 2019. "Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network" Sensors 19, no. 11: 2472. https://doi.org/10.3390/s19112472

[24] Vijeikis, Romas, Vidas Raudonis, and Gintaras Dervinis. 2022. "Efficient Violence Detection in Surveillance" Sensors 22, no. 6: 2216. https://doi.org/10.3390/s22062216

[25] Choqueluque-Roman, David, and Guillermo Camara-Chavez. 2022. "Weakly Supervised Violence Detection in Surveillance Video" Sensors 22, no. 12: 4502. https://doi.org/10.3390/s22124502

[26] Wang, Xiang, Jie Yang, and Nikola K. Kasabov. 2023. "Integrating Spatial and Temporal Information for Violent Activity Detection from Video Using Deep Spiking Neural Networks" Sensors 23, no. 9: 4532. https://doi.org/10.3390/s23094532

[27] Sumon, Shakil & Goni, Raihan & Hashem, Niyaz & Shahria, Md Tanzil & Rahman, Mohammad. (2019). Violence Detection by Pretrained Modules with Different Deep Learning Approaches. Vietnam Journal of Computer Science. 7. 10.1142/S2196888820500013.