**DE Grp-6**

**Team Members:** 1.Devesh Chandak

2.Vibhor Surana

3.Fawazz

**REAL TIME VIOLENCE DETECTION &**

**ALERT SYSTEM**

**DE PROJECT REPORT**

**INTRODUCTION**

Violent behavior in public places is an issue that has to be addressed. Communities are also eroded by violence, which reduces productivity, lowers property values, and disrupts social services. Across the world, violence is a severe public health issue. It affects people at various phases of life, from infants to the elderly.

Recognizing violence is challenging since it must be done on real-time videos cap- tured by a large number of surveillance cameras at any time and in any location. It should be able to make reliable real-time detection and alert corresponding authorities as soon as violent activities occur.

Public video surveillance systems are widespread around the world and can provide accurate and complete information in many security applications. However, having to watch videos for hours reduces your ability to make quick decisions. Video surveil- lance is essential to prevent crime and violence. In this regard, several studies have been published on the automatic detection of scenes of violence in video. This is so that authorities do not have to watch videos for hours to identify events that only last a few seconds. Recent studies have highlighted the accuracy of deep learning approaches to violence detection.

Indeed, deep learning methods have proven effective for extracting spatiotemporal features from videos. A function that represents the motion information contained in

a series of frames in addition to the spatial information contained in a single frame. In this work we will be discussing about the implementation of a Real-Time violence alert system using MobileNetv2. The frames obtained as output from the model is enhanced. Then these frames along with time and location of the recorded incident are send to the nearby police station as an alert via the alert module of the proposed system.

The remaining report is categorized as follows, Section II presents related studies and their detailed comparisons. Section III discusses the selected approach in a de- tailed manner explaining the proposed architecture and the dataset, whereas section IV provides details of experimentation and discusses the evaluation of the approach. Finally, section V concludes the report and discusses future innovations that could be brought into our project.

**Problem Statement**

The primary problem addressed in this project is the automatic detection of violent activities in real-time surveillance video footage. The challenge lies in processing and analyzing large volumes of video data to identify brief instances of violence. The dataset used consists of 1,000 video clips categorized into violence and non-violence classes, with an average duration of 5 seconds per clip. The proposed technique employs the MobileNet v2 architecture, a deep learning model, for detecting violent activities in the video frames. This model is optimized for real-time applications, balancing accuracy and computational efficiency.

**Objective**

The objective of this project is to develop a real-time surveillance system that can accurately and efficiently detect violent activities in video footage and send alerts to the appropriate authorities. The system should:

1. **Process video data in real time**: Handle continuous video streams and analyze them frame by frame.
2. **Detect violent activities with high accuracy**: Identify and classify actions as violent or non-violent with minimal false positives and false negatives.
3. **Send automated alerts**: Provide timely notifications to law enforcement or other relevant authorities, including visual evidence (enhanced frames) and contextual information (time and location).

**LITERATURE REVIEW**

Recently proposed methods for violence detection can be roughly classified into three categories visual based approach, audio-based approach and hybrid approach

Visual Based Approach : Visual information is retrieved and represented as rele- vant features in this approach. Local features and global features are two types of fea- tures. Position, velocity, form, and color are examples of local features, while average speed, region occupancy, relative positional fluctuations, and the interactions between objects and backdrop are examples of global features.

Audio Based Approach : Audio data is used to classify violence in this approach. It uses a hierarchical technique based on Gaussian mixture models and Hidden Markov models to distinguish gunshots, explosions, and automobile braking in audio.

Hybrid Approach : The emphasis in the hybrid method is on merging visual and audio characteristics. Some techniques recognize violent incidents in videos utiliz- ing flame and blood detection and recording the degree of motion, as well as the typ- ical sounds of violent occurrences. The CASSANDRA system, detects aggression in surveillance videos using motion features associated with articulation in video and scream-like cues in audio.

* 1. **SPATIO-TEMPORAL MODELLING METHOD WITH 2D CNN**

This approach mainly consists of 3 modules, Motion Saliency Map(MSM),2D CNN’s with frame-grouping, Temporal Squeeze and Excitation Block. MSM can effectively highlight moving objects, to magnify prominent regions connected to violence in videos. Attention maps are generated by dilation of the motion boundaries, which can repre- sent the boundaries of moving objects. Then, using the 2D CNN backbones, convert each three-channel frame into a single-channel frame in the middle of the forward path and group three consecutive frames to learn Spatio-temporal representation in a video. This technique is known as frame-grouping. Temporal Squeeze and Excitation Module is a module that is used to classify a target event with less computational cost.

This approach cannot be used in moving cameras as a sequence of frames in a fixed position is necessary to recognize violence. The computational cost of this approach is high as it requires lots of calculations using 3 different modules with a predefined set of tasks.

* 1. **SPACE TIME INTEREST POINT (STIP)**

Interest points are detected using the Space Time Interest Point (STIP) approach, which analyses spatial and temporal differences. A 3 dimensional volume is extracted around each space time interest point; the volume depicts how a 2D image segment evolves over time. The scale of the identified characteristics determines the size of the 3D volume. The discovered interest points have a significant degree of intensity variation in space and non-constant motion in time. These important points can be found on a number of spatial and temporal scales.A Histogram of Oriented Gradients is used to identify the appearance of the 3D cube, and a Histogram of Oriented Flows is used to define the motion. These traits may be utilised to recognise motion events with high

accuracy, and they are resistant to changes in pattern scale, frequency, and velocity.

Despite promising results with a high accuracy rate for this task, the computational cost of extracting such features is extremely high for practical applications, such as surveillance and media rating systems.

* 1. **VIOLENCE DETECTION USING LOW-LEVEL FEATURES**

The motion region is segmented according to the distribution of the optical flow field. In the motion domain, two types of low-level traits are extracted to express the appear- ance and dynamics of aggressive behavior. The low-level functions are the Oriented Gradient Local Histogram (LHOG) descriptor extracted from the RGB image and the Optical Flow Local Histogram (LHOF) descriptor extracted from the optical flow im- age. The extracted features are encoded using a Bag of Words (BoW) model to remove redundant information and obtain a constant-length vector for each video clip. Finally, we use a support vector machine (SVM) to classify the video layer vectors.

Detection of violence in a crowded scene was a challenge due to the serious occlu- sion and moving crowd. It has low detection rate and high false alarm.

* 1. **SPATIO-TEMPORAL FEATURES WITH 3D CNN**

For this purpose, the framework of triple staged end to end deep learning violence detection is proposed. Firstly in the surveillance video streams, persons are detected using lightweight CNN model to overcome and reduce the huge processing of unusable frames. Secondly, an order of 16 frames with detected individuals is passed to 3D CNN, where the spatiotemporal features of these sequences are extracted and fed to the Softmax classifier. Then, the 3D CNN model is optimized using a neural networks

optimization toolkit and an open visual inference developed by Intel. Trained model is converted into an intermediate illustration and changes it for execution at the end platform for the final detection of violence

This process is really computationally expensive and depends on a lot of parame- ters, hence is not suitable for real-time environments and for daily use.

* 1. **SENSOR-NETWORK APPROACH**

Deep neural network images are extracted from an MPEG encoded video input stream. Then the DNN is trained to recognise frames which correspond to violent behavior.

Similar to other methods, lots of computational power is required which is not suitable for real -time usage / environments or for daily usage.

**Techniques and Algorithms**

The system developed in this project is based on the MobileNet v2 architecture, a lightweight convolutional neural network (CNN) optimized for mobile and embedded vision applications. The key features of the MobileNet v2 architecture include:

1. **Depthwise Separable Convolutions**: This technique decomposes standard convolutions into a depthwise convolution followed by a pointwise convolution, significantly reducing the number of parameters and computational cost while maintaining accuracy.
2. **Inverted Residual Blocks**: These blocks reverse the traditional residual connections by connecting bottleneck layers (narrow-wide-narrow structures) instead of the standard wide-narrow-wide approach. This design allows for efficient memory usage and faster computations.
3. **Linear Bottlenecks**: Linear bottleneck layers expand the channel dimension of the input before applying nonlinear functions, reducing the risk of information loss and enhancing the model's ability to learn complex features.

In addition to the MobileNet v2 architecture, the system incorporates image enhancement techniques to improve the clarity of detected frames. The Python Imaging Library (PIL) is used to enhance the brightness, contrast, and color balance of the output frames, ensuring that the visual evidence provided to authorities is of high quality.

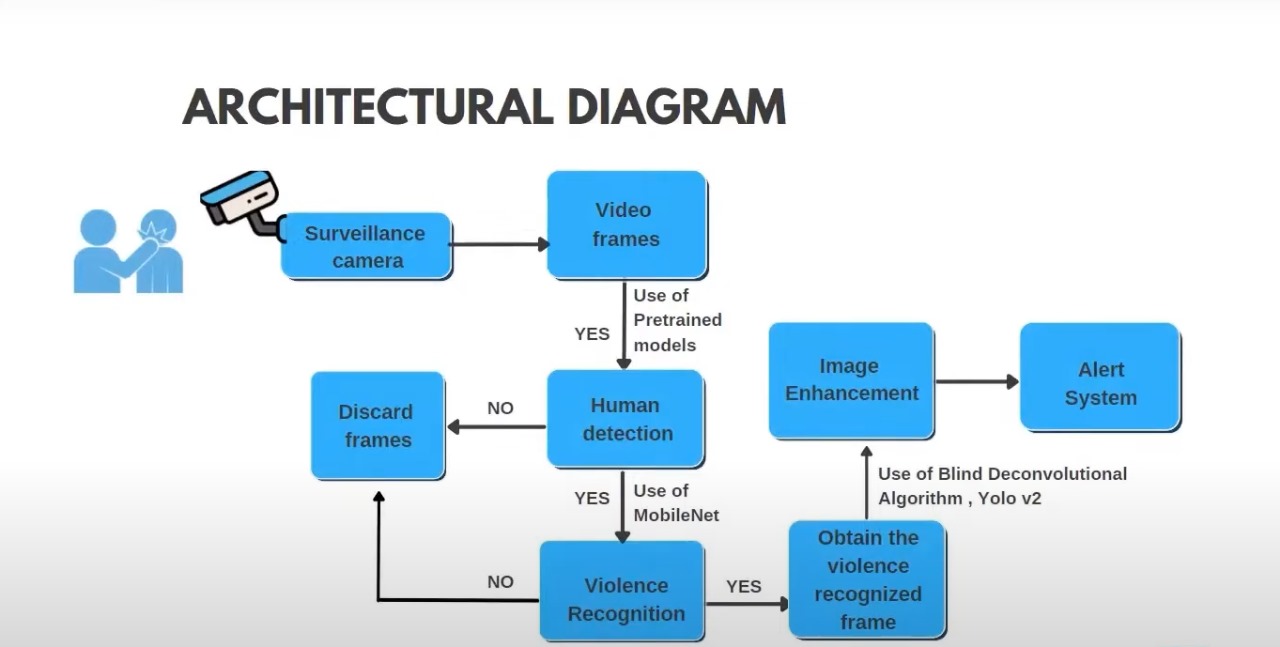
## ****Dataset & Description****

The dataset used for training and testing the violence detection system consists of 1,000 video clips, each approximately 5 seconds long. The clips are divided into two categories: violence and non-violence. The majority of the videos are sourced from CCTV footage, providing a realistic representation of the types of scenes the system will encounter in real-world applications.

* **Violence Class**: This class includes 500 video clips depicting various forms of violent activities, such as physical fights, assaults, and other aggressive behaviors. These clips are used to train the model to recognize patterns and features associated with violence.
* **Non-Violence Class**: This class includes 500 video clips of non-violent activities, such as walking, talking, or standing still. These clips help the model learn to distinguish between violent and non-violent behaviors, reducing false positives.

During each training epoch, 350 videos from the violence class and 350 videos from the non-violence class are used to train the model. This balanced approach ensures that the model can generalize well to new, unseen footage.

**FLOW DIAGRAM**



## USE CASE DIAGRAM

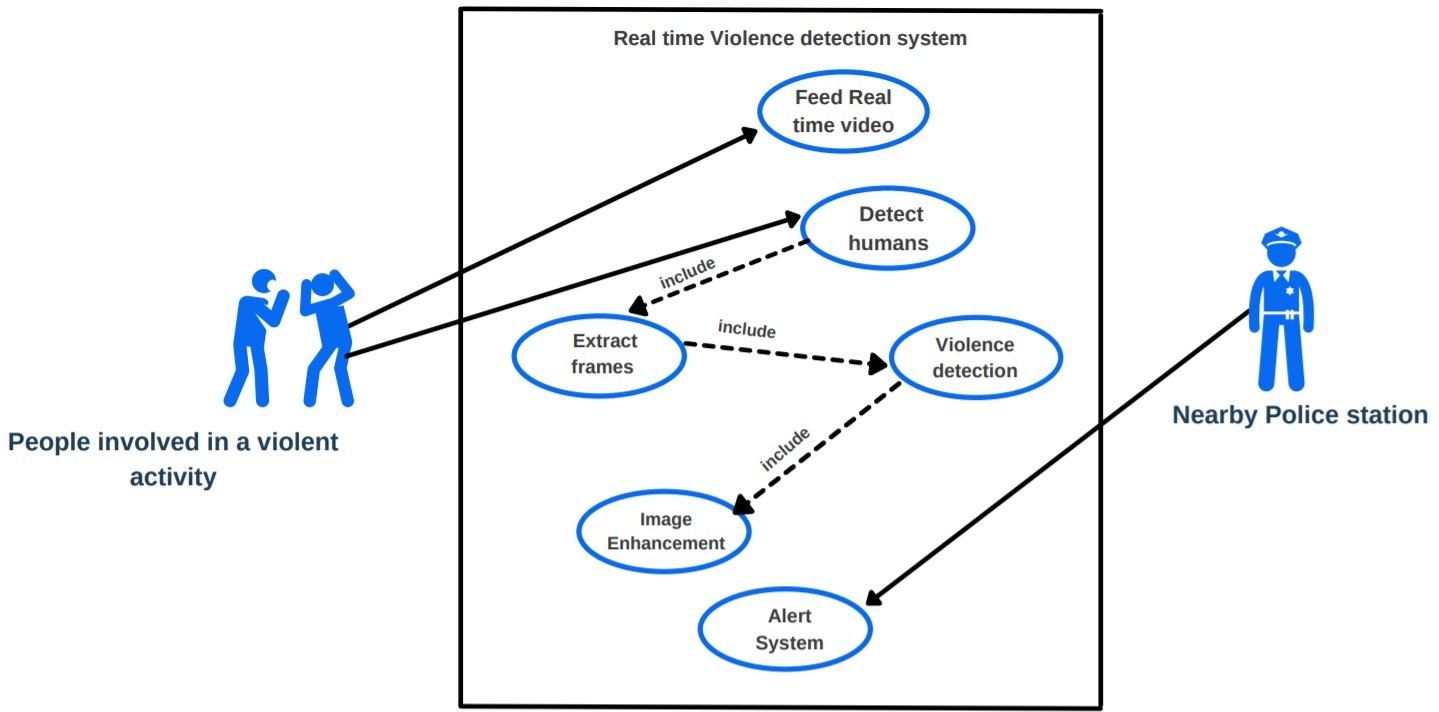


Figure 3.1: Use Case diagram

The rectangle helps to define the scope of the proposed architecture. Anything happens within the rectangle happens within the system. There are 2 actors in this scenario, People involved in a violent activity and a nearby police station. People involved in a violent activity are primary actors as their actions will be detected by a real-time system. Use cases are represented by oval shape and they represent an action that accomplishes some sort of task within the system. Whenever a real-time video is given as input, the system will try to detect humans. Every time humans are detected the given included use cases are executed as well. They are Frames Extraction, Violence Detection and Image Enhancement. After these three steps alert system will send an alert to a nearby police station.

## ARCHITECTURE

Footage from the surveillance camera is broken down into frames. The frames are given as input to MobileNet v2 classifier for detecting violent activities in the given sequence of input frames. If no violent activity is recognized the respective frames are discarded. The violence detected frame is obtained and it is enhanced for better clarity. That frame, along with the location are sent to the nearest authorities using Telegram bot.

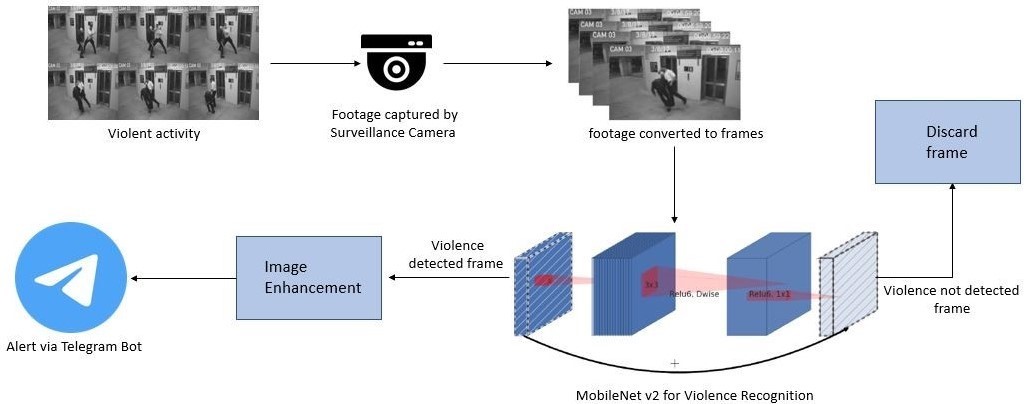


Figure 3.2: High level architecture diagram

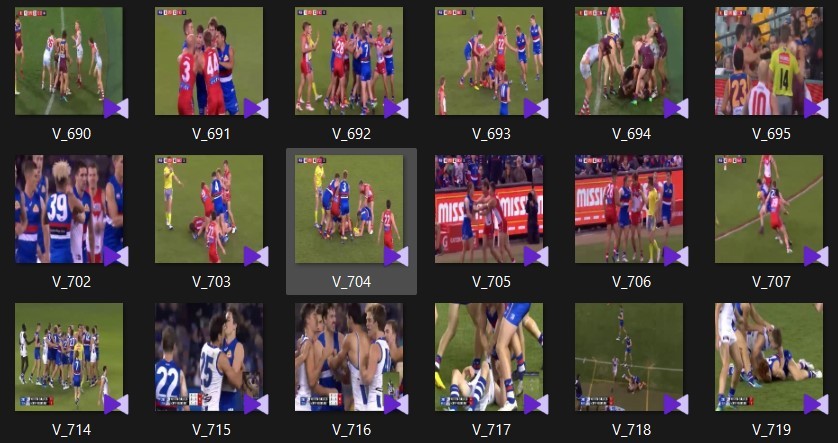
**Dataset:** The dataset contains 1000 video clips which belongs to two classes, violence and non-violence respectively. The average duration of the video clips is 5 seconds and majority of those videos are from CCTV footages. For training, 350 videos each from the violent and non-violent classes are taken at each epoch.

Figure 3.3: Video clips from the violence dataset

**MobileNet V2:** The MobileNet architecture is primarily based on depth wise sep- arable convolution, in which factors a traditional convolution into a depth wise convo- lution followed by a pointwise convolution.

The module presents a residual cell (has a residual/identity connection) with stride of 1, and a resizing cell with a stride of 2. From Figure 3.3, "conv" is a normal con- volution, "dwese" is a depth wise separable convolution, "Relu6" is a ReLu activation function with a magnitude limitation, and "Linear" is the use of the linear function.

The main strategies introduced in MobileNetV2 were linear bottleneck and inverted residual blocks. In the linear bottleneck layer, the channel dimension of input is ex- panded to reduce the risk of information loss by nonlinear functions such as ReLU. It stems from the fact that information lost in some channels might be preserved in other

channels. The inverted residual block has a ("narrow" -"wide"-""narrow") structure in the channel dimension whereas a conventional residual block has a ("wide" - "narrow"- "wide") one. Since skip connections are between narrow layers instead of wider ones, the memory footprint can be reduced.

**Image Enhancement:** Image Enhancement is performed on the frames that are ob- tained as output. This is performed using the inbuilt functions provided by the Python Imaging Library(PIL). PIL offers extensive file format support, efficient presentation, and fairly powerful image processing capabilities. The Core Image Library is designed to provide quick access to data stored in several major pixel formats. It provides a solid foundation for common image processing tools. The brightness and colour of the ob- tained output frames is increased by a factor of 2.

**Alert Module:** The alert module sends alert message to the specified authority. Figure 3.4 describes the architecture of the implemented alert system. When a frame is detected true for violence, the system initialises a counter variable to one. Then it checks the subsequent 30 frames, whether if they too have violence detected true. The counter is incremented at each consecutive frame that is true for violence. If a frame is false for violence, the counter variable is set to 0 and starts checking the consecutive frame respectively checking whether violence is recognized. On the other hand, if the violence is detected true for the 30 consecutive frames, the current time is obtained using an inbuilt python function and an alert is sent to a Telegram group that consists officials of higher authorities. The Alert message comprises of an image of the detected violent activity, current timestamp and the location where the camera is placed.

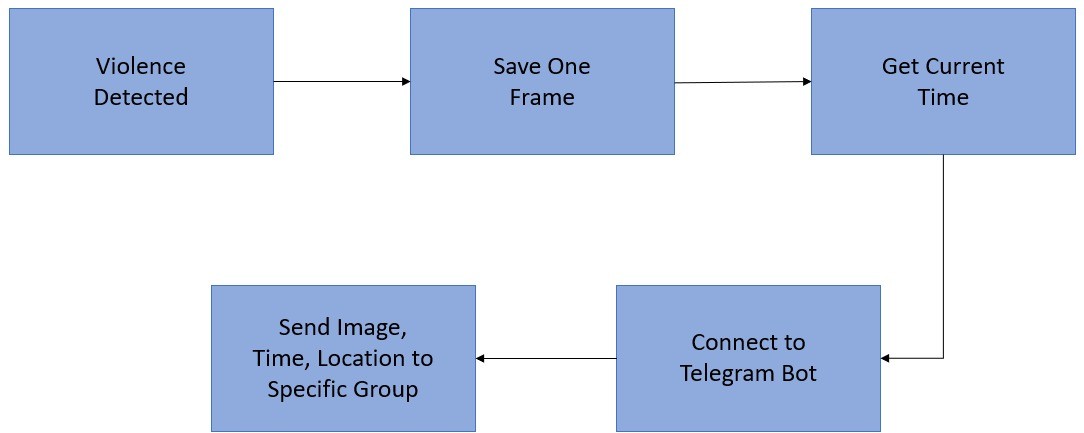


Figure 3.5: Architecture diagram of the Alert System

Figure 3.5 shows the alert message that is sent to the telegram group by the telegram bot. The concerned authorities can view the alert and take necessary actions.

**Pros and Cons**

**Pros:**

* **High Accuracy**: The system achieves high accuracy in detecting violent activities, ensuring that incidents are identified promptly and correctly.
* **Real-Time Processing**: The use of the MobileNet v2 architecture allows for real-time processing of video streams, enabling immediate detection and response.
* **Low Computational Cost**: The system is optimized for deployment on low-end devices, such as embedded systems, making it accessible for a wide range of applications.
* **Scalability**: The system can be scaled to monitor multiple cameras simultaneously, providing comprehensive coverage of large areas.

**Cons:**

* **Limited to Visual Detection**: The system is primarily based on visual data, which means it may not detect violence in scenarios where the visual evidence is unclear or obscured.
* **Dependency on Video Quality**: The performance of the system may degrade if the video quality is poor, such as in low-light conditions or with low-resolution footage.
* **Computationally Intensive Training**: While the system is optimized for real-time use, the initial training of the model requires substantial computational resources, which may be a limitation for some users.

**RESULTS AND DISCUSSION**

In this section testing and training accuracy are displayed in the below given graphical representation. Fig. 4.1 displays the training and testing accuracy and loss for the MobileNet v2 model when a dataset containing 1000 videos of average duration 7 seconds is given as input. For each epoch 350 videos from the violence class and 350 videos from the non-violence are trained. 81% accuracy was obtained on training and a respective accuracy of 80% was obtained when a CCTV footage that was not included in the dataset was given for testing. The obtained output video frames are shown in Figure 4.3

In the graph in Figure 4.1 the accuracy and loss comes to a constant level of incre- ment and decrement after approximately 5 epochs. The obtained confusion matrix and other evaluation parameters are shown in Fig. 4.1

A video with violence is given as input to the system. Figure 4.3 shows one frame in the video that was labeled to have violent activity. Another video clip without violent activity was given as input. Figure 4.4 shows one frame of that video which is rightly labelled as false or violence.

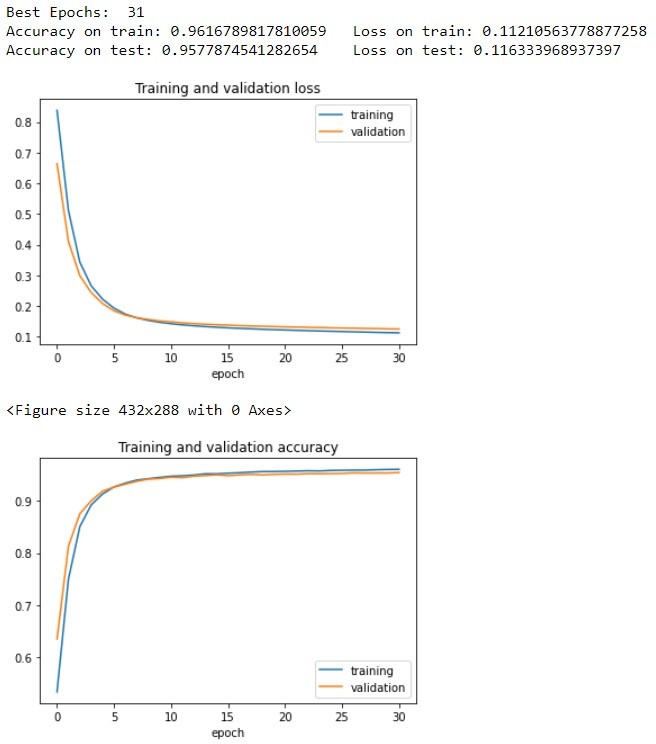


Figure 4.1: Accuracy and Error of the training set

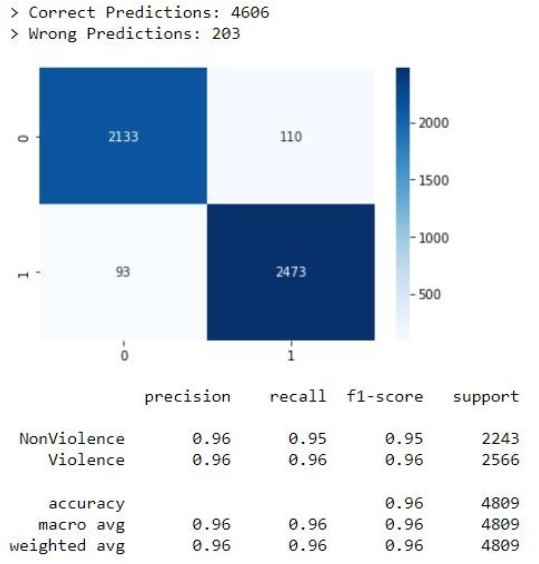


Figure 4.2: Confusion matrix of the trained model



Figure 4.3: Output frame that recognized violence



Figure 4.4: Output frame that did not recognize violence

# CONCLUSION

Violence scene detection in real-time is a challenging problem due to the diverse con- tent and large variations quality. In this research, we use the MobileNet v2 model to offer an innovative and efficient technique for identifying violent events in real-time surveillance footage. The proposed network has a good recognition accuracy in typical benchmark datasets, indicating that it can learn discriminative motion saliency maps successfully. It’s also computationally efficient, making it ideal for use in time-critical applications and low-end devices. Here, we had also shown the working of an Alert system that is integrated with the pretrained model. In comparison to other state-of- the-art approaches, this methodology will give a far superior option.

**Future Scope:** This model could be upgraded to work in multiple cameras con- nected by a single network in a concurrent fashion. A short video of the violent activity could be incorporated along with the alert message.

# References

* + 1. Khan SU, Haq IU, Rho S, Baik SW, Lee MY. Cover the Violence: A Novel Deep- Learning-Based Approach Towards Violence-Detection in Movies. Applied Sci- ences. 2019; 9(22):4963. https://doi.org/10.3390/app9224963J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
    2. Gopalakrishna, MT. (2016). Violence Detection in Surveillance Video-A survey. International Journal of Latest Research in Engineering and Technology (IJL- RET). NC3PS - 2016. 11-17. K. Elissa, “Title of paper if known,” unpublished.
    3. M. -S. Kang, R. -H. Park and H. -M. Park, "Efficient Spatio-Temporal Model- ing Methods for Real-Time Violence Recognition," in IEEE Access, vol. 9, pp. 76270-76285, 2021, doi: 10.1109/ACCESS.2021.3083273.
    4. Ullah FUM, Ullah A, Muhammad K, Haq IU, Baik SW. Violence Detection Us- ing Spatiotemporal Features with 3D Convolutional Neural Network. Sensors (Basel). 2019 May 30;19(11):2472. doi: 10.3390/s19112472.
    5. A. -M. R. Abdali and R. F. Al-Tuma, "Robust Real-Time Violence Detection in Video Using CNN And LSTM," 2019 2nd Scientific Conference of Computer Sciences (SCCS), 2019, pp. 104-108, doi: 10.1109/SCCS.2019.8852616.
    6. J. C. Vieira, A. Sartori, S. F. Stefenon, F. L. Perez, G. S. de Jesus and V. R. Q. Leithardt, "Low-Cost CNN for Automatic Violence Recognition on Embedded

System," in IEEE Access, vol. 10, pp. 25190-25202, 2022, doi: 10.1109/AC- CESS.2022.3155123.

* + 1. Sandler, Mark Howard, Andrew Zhu, Menglong Zhmoginov, Andrey Chen, Liang-Chieh. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. 4510-4520. 10.1109/CVPR.2018.00474.
    2. J. Wang and Z. Xu, "Crowd anomaly detection for automated video surveillance," 6th International Conference on Imaging for Crime Prevention and Detection (ICDP-15), 2015, pp. 1-6, doi: 10.1049/ic.2015.0102.
    3. P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo and A. F. Dragoni, "Deep Learning for Automatic Violence Detection: Tests on the AIRTLab Dataset," in IEEE Access, vol. 9, pp. 160580-160595, 2021, doi: 10.1109/AC- CESS.2021.3131315.
    4. M. Ramzan et al., "A Review on State-of-the-Art Violence Detection Tech- niques," in IEEE Access, vol. 7, pp. 107560-107575, 2019, doi: 10.1109/AC- CESS.2019.2932114.