2024 4th International Conference on Intelligent Technologies (CONIT)
Karnataka, India. Jun 21-23, 2024

Violence Detection in Automated Surveillance using CNN

Abishek Ghalley
Department of Electronics & Communication Engineering
Delhi Technological University
Delhi, India
abishekghalley01@gmail.com

Abdelrahman Abdelsalam
Department of Electronics & Communication Engineering
Delhi Technological University
Delhi, India
abdelsalamabdelrahman0@gmail.com

Wongani Dombola
Department of Electronics & Communication Engineering
Delhi Technological University
Delhi, India
wongadmbl7@gmail.com

M.S. Choudhary
Department of Electronics & Communication Engineering
Delhi Technological University
Delhi, India
msc_1976@yahoo.com

2024 4th International Conference on Intelligent Technologies (CONIT) | 979-8-3503-4990-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/CONIT61985.2024.10626390

***Abstract*—As the world proceeds into the age of automation, the need of manual surveillance becomes increasingly questionable. This is not only due to excessive hours spent behind security monitors, but their continuous ineffectiveness from lapses of lack of concentration among other factors during manual surveillance. Thus, with the aid of Deep Learning and other acclaimed work related to real-time violence detection in surveillance, a conclusive study has led to proposed autonomous system using artificial intelligence. Hence this paper aims to highlight integration of a proposed Convolution Neural Network model that will not only detect violence but informatively alert authorities to the premise of intrusion. We employ what is referred to as a deep lightweight neural network called MobileNetV2. Preferred due to its unique architecture impressively designated for resource-constrained environments and devices, it is ideal for high performance requirement applications with minimal computational capabilities. The model acts as a feature extractor and classifier in training on various datasets. The model achieved the highest accuracy rate of 99% from one of a datasets.**

***Index Terms*—Convolutional neural network, depth-wise separable Convolution, deep learning, mobileNetV2, surveillance, violence detection.**

I. Introduction

Deep learning is an aspect of machine learning that has revolutionized human activity recognition. This is where acts and objectives of one or more personal are analysed and assessed via an algorithm or more. Hence models are designed to act as a decision makers based on its collective behavioural data. With security being an essential concern of day-to-day activities and livelihood, many research work has gone into violence detection systems using deep learning. Automated deep learning surveillance systems offer an effective substitute to traditional manual surveillance systems [1]. Optimizing the proposed model with various scenarios; both malicious or not, has helped achieve a balanced fast and accurate setup which can be implemented in real-time surveillance systems. This study outlines the effectiveness of deep learning in malicious situations involving two or more people. Previous work has focused models consisting of two parts working cohesively as feature extractor and classification respectively [1], [2], [3].Implementation of the first part consisting of computing optical flow and the latter; convolutional neural networks for feature extraction and classification. These methods have indeed shown results necessary to conclude CNN based surveillance systems are more efficient compared to manual surveillance [4]. By training a proposed CNN model and application of Bidirectional Long Short-term Memories, the model proceeds to sequential previous feature analysis for better decision making on current assignments [1], [2], [5]. In this case, MobileNetV2, a CNN model proposed for its design capabilities to perform complex computations on resource-constraint environments, for feature extraction and classification [6], [7].Hence, the next section of this paper overviews previously proposed methodology in violence detection. Section 3 indicates the computational details and methodology of the proposed model. Next Section illustrates training datasets and their distinctiveness. In Section 5, experimental setup and results are elaborated upon. Finally, the last section concludes the paper followed by References

II. Related Works

Various work centered around automated violence-detection hence significantly progressed with several researchers leaning towards machine learning and deep learning approaches. Muhammad Javed Iqbal et al (2021) [8] have implemented faster R-CNN model for detecting objects in video footage captured by surveillance drones. They comparatively implemented pre-trained neural networks such as SqueezeNet, GoogleNet, ResNet18, and ResNet50 for further classification of the intruder. Results obtained showed a precision rate of 79% on average from the proposed model.

Similarly, Ahmed M et al (2021) [3]reviewed the implementation of InceptionV4 and a sequential CNN keyframe extractor for removal of duplicate frames to improve efficiency

in automated surveillance. After comparisons between various image classification methods such as trajectory-based and non-object centric approaches, deep learning methods based on CNNs achieved higher accuracy rates. Using both accessible and custom datasets, the proposed InceptionV4 model achieved 98% after numerous epochs.

Other authors address the challenges of automating surveillance systems for violence detection in industrial environments uses a limited computational power [9]. This paper introduces a novel AI-enabled framework, called VD-Net, that is specifically designed for Industrial Internet of Things -based surveillance networks. The framework uses a lightweight convolutional neural network (CNN) for initial object detections. VD-Net focuses on humans or suspicious objects like a gun and a knife. Then, when violence is detected, an alert is sent to the IoT network, and only the relevant frames are sent to a cloud server for detailed analysis. The analysis utilizes feature extraction using performance convolutional long short-term memory, which is further processed using gated recurrent units for final violence detection.

III. Research Methodology

The proposed model is based on a deep lightweight neural network called MobileNetv2 for feature extraction and classification. MobileNetv2 is a Depth wise Separable Convolution network that initiates parameter reduction in image classification [10]. The second version architecture includes Depthwise and Pointwise convolutions made separable from the standard convolution through bottlenecks from inverted residuals applied beforehand. This is referred to as Depthwise Separable convolution. Unlike a standard convolution; which is based on N number of kernels/filters being equal to the number of 1 × 1 convolution operations, Depth wise Separable Convolution allows filter application to a single channel per filter in the network [6], [11]. This can be comprehended as follows;

$$N \times D_p^2 \times D_k^2 \times M \tag{1}$$

Which is the total number of parameters attained after multiplication, where $N \times D_p^2$ is the output multiplication per convolution and $M \times D_k^2$ is the filter size applied during convolution. By splitting the convolution into Depth-wise and point-wise convolutions such that M number of filters is applied to M number of channels in Depth wise Convolution;

$$M \times D_k^2 \times D_p^2 \tag{2}$$

This is the total number of convolution operations in Depth-wise convolution, where $D_k \times D_k \times 1$ is the convolutional multiplications applied to all M number of channels and $D_p \times D_p$ is the attained parameters after filtering M number of channels respectively. Point-wise convolution however, 1 × 1 convolution filters are applied M number of channels. This is presented as;

$$M \times D_p^2 \times N \tag{3}$$

where the filters applied is equal to number of channels respectively. Hence the number of total parameters defined in Depth-wise Separable is made up of the number of multiplications attained in Depth-wise and Point-wise convolution and denoted as;

$$M \times D_p^2 \times (D_k^2 + N) \tag{4}$$

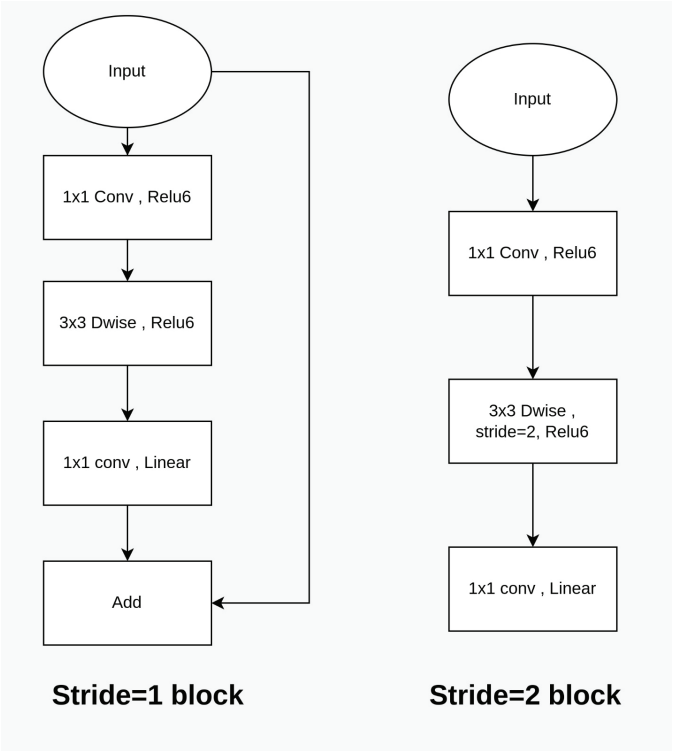

Fig. 1. Architecture of MobileNetv2 CNN model

A. MobileNetV2 Architecture

These separate computations in return allow the reduction of parameters while increasing accuracy by using linear bottlenecks. Inverted residual blocks based on ReLu6 activates input channels expansion to allow processing of more complex attributes initially [7], [10]. Feature responses are then filtered using Squeeze-and-Excitation (SE) blocks allowing the training model to achieve informative analysis.

The architecture of MobileNetV2 contains the initial fully convolution layer with 32 filters, followed by 19 residual bottleneck layers [10]. A kernel size of 3 × 3 convolution is used as a standard for modern networks which utilizes dropout and batch normalization during training. This allows the model a relatively lesser parameter ratio as compared to standard convolutional neural networks [7]. The balance between speed and accuracy is denoted by the proposed framework with up to 16fps (frames per second) of video sequence as input.

The model practically reads the video frames for data augmentation initially. Convolutional pooling layers act as spatial extractors before vectoring out any external features [7], [9], [11]. Informative analysis attained by the model allows an autonomous response which is output based from the input data. Hence, the next section describes various defined datasets used in model training and implementation.

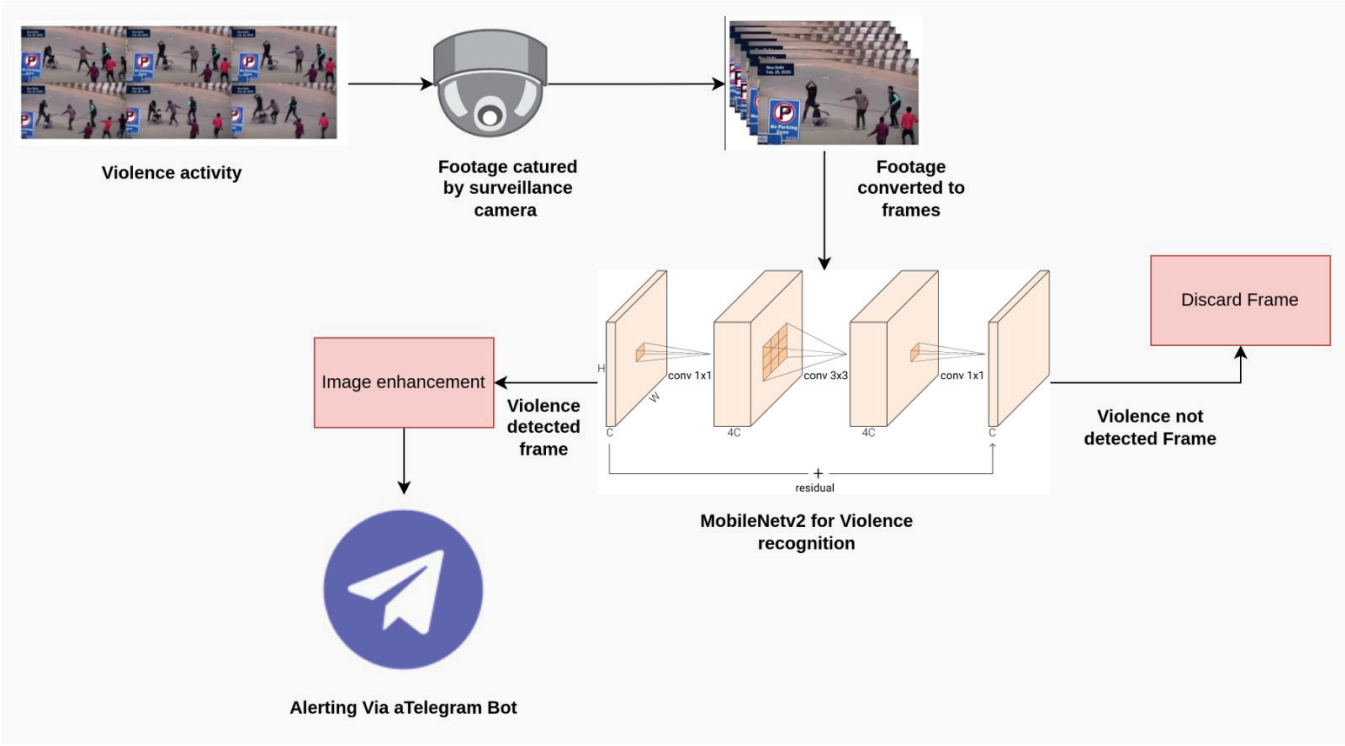

Fig. 2. Proposed model architecture denoting feature extraction and classification in real-time surveillance using MobileNetV2

IV. Datasets

This section of the paper we will discuss about the datasets that we have used to train and test the model and in the later part we will discuss the experimental results that we have obtained from the model. We used 4 types of datasets to train and test the model; hockey fight dataset, real live violence situation dataset, movie dataset all obtained from Kaggle and a custom dataset made from local scenarios. The hockey fight dataset, as the name suggests, includes the 500 violent and 500 nonviolent video clips from the ice hockey match. Each video are 1 to 3 seconds long with the constant frame-sizes. Since the match is occurs on mostly ice, the background in the video frames not likely affectious.

Similarly, the movie dataset contains the two classes of violent and nonviolent videos, 123 each. The variations in scenery of the frames allow the model training on various backgrounds during data augmentation. Real-life violence situation dataset consists of recorded real-life video clips each clip with the duration of 4 to 5 seconds each. There are 1000 violence situation clips and 1000 nonviolence situation clips collected from various time and geographic areas around the world. The frame-size of the varies from video to video and the environment in which it is recorded also varies.

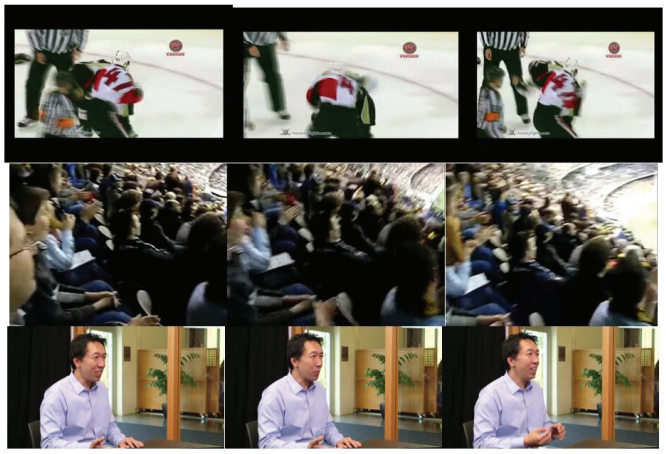

Fig. 3. Videos frames comprising of nonviolence from hockey (top), movie dataset (middle row) and real-life violence situation dataset; an interview (bottom row)

Since the above datasets are accessible from Kaggle, defined violent scenarios from some of the datasets were corrupted. Additionally, the various cinematic environment also led to reduced accuracy. Hence, a proposed dataset was made based on our local environment and practical scenarios which include; street fights from CCTV surveillance, altercations between college students, and other clips from the extremely violent movies. Most of the clips were collected around the college campus taking into account different times of day and nights. The majority include low-light scenarios assuming that the

Fig. 4. Video frames from custom violence dataset attained from local scenarios

most of the violence occur at night compared to daytime. The frame-size of the videos vary and each video is of 2- 5 seconds of duration. We have used 152 violent data and 152 nonviolent data to train the model and 8 clips to test the model.

TABLE I
NUMBER OF DATASETS SAMPLES USED IN MODEL TRAINING

Dataset	Violent videos	Non-violent videos	Total
Custom	152	152	304
Hockey	500	500	100
Movie	123	123	246
Real-life Situation	1000	1000	2000

V. EXPERIMEMENTAL SETUP AND RESULTS

A. Data preprocessing

Before training the MobileNetV2 model, the datasets underwent preprocessing to prepare the data for input into the network. The preprocessing videos in the datasets were processed to extract individual frames using OpenCV library in Python. This step involved iterating through each video file and extracting frames at regular intervals or key frames, depending on the dataset characteristics.

Image augmentations were applied to increase the diversity of the training data and improve model generalization. Augmentations such as random rotations, flips, and shifts were performed using libraries like TensorFlow's ImageDataGenerator. These augmentations helped the model learn robust features and reduce overfitting to specific data characteristics. After extracting frames and applying augmentations, the pixel values of the images were normalized to a range suitable for training the MobileNetV2 model. Normalization ensured that the input data had zero mean and unit variance, which is a common practice to stabilize and accelerate the training process [6], [7], [10].

The training process starts by uploading the dataset labeled as violence and Non-violence. Then, the frames are extracted from these videos and pre-processed to match the requirements of MobileNetV2 CNN model. During the training, the model learns to identify patterns and features indicating violence through iterative optimization on the training dataset. The performance of the model is validated on a separate set of data to ensure it generalizes well and to avoid overfitting. Then, the trained model is tested on an independent set of videos

Fig. 5. Visualization of non-violence detection (False) and violence detection (True) by the model

to verify its accuracy in differentiating between violence and non-violence situations.

B. Results

This subsection describes the main observations the model displayed after runtime on various epochs. As proposed, the system was able to display a visual indicator of analysis and description of outcome obtained. This can be seen in the figure below on the analyzed video frames.

The performance of the MobileNetV2 CNN model was evaluated across various metrics, including accuracy, loss, precision, recall, and F1-score. These metrics will provide a good insight into the model's performance in differentiating between violent and non-violent scenes across different types of datasets.

TABLE II
DETAILED EVALUATION RESULTS OF THE MODEL ON ACCURACY, PRECISION, RECALL, F1-SCORE, AND AUC ACROSS ALL DATASETS

Dataset	Accuracy	AUC	F1-Score	Precision	Recall
Custom	**0.99**	**0.99**	**0.99**	**0.99**	0.99
Hockey	0.92	0.92	0.92	0.92	0.92
Movies	0.89	0.89	0.89	0.89	0.88
Real-life Situation	0.95	0.95	0.95	0.95	0.95

The visual representation in the Figure 6 illustrates the comparison in terms of ROC and AUC for hockey fight dataset, real life violence situations dataset, movies dataset, and our own custom dataset. This comparison in Figure 6 presents a trade-off between the false positive rate(FP-rate) and true positive rate(TP-rate) through ROC curve. ROC curve is made by plotting the true positive rate in the y-axis and false positive rate in the x-axis. Similarly, the Area Under Curve(AUC) is calculated to indicate that a value close to 1 signifies superior classification performance for an ideal model.

$$\text{True Positive Rate} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{False Positive Rate} = \frac{FP}{FP + TN} \tag{6}$$

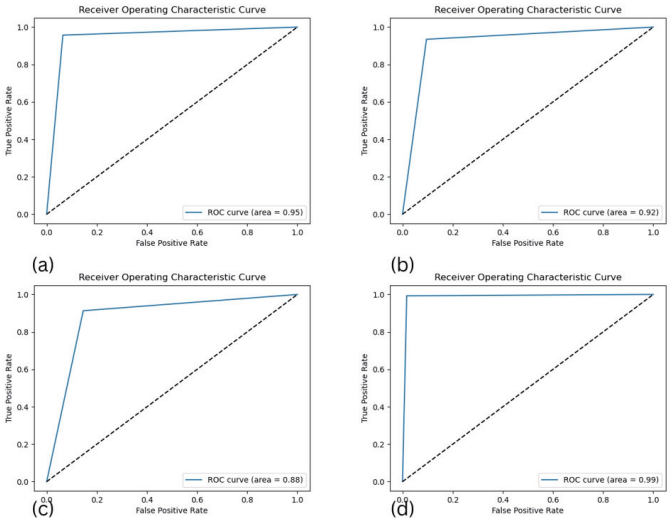

Fig. 6. A In-depth analysis of our model's performances using Area Under Curve and Receiver Operating Characteristic. (a) ROC and AUC for real life violence situations dataset. (b) ROC and AUC Hockey dataset. (c) Results for Movies dataset. (d) Performance of our own dataset.

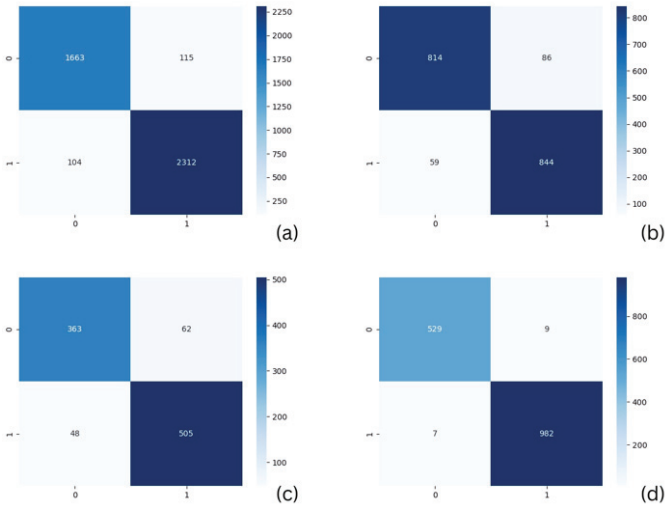

Fig. 7. Representation of the confusion matrix for our model over each dataset. (a) Real life violence situations dataset. (b) Hockey dataset. (c) Movies dataset. (d) Confusion matrix of our custom dataset.

Likewise for True Negative Rate and False Negative Rate respectively, Thus accuracy can be calculated as follows;

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

To conduct a thorough investigation of the correctly and wrongly classified classes of violence and non-violence, we have generated the confusion matrix values such as true negative (TN), true positive (TP), false negative (FN), and false positive (FP). These values also enabled the calculation of precision, recall, and F1 score which are equally important performance measures.

In addition to the evaluation metrics presented in the previous section, the visualizations of the accuracy and loss gives a further insight into the performance of our violence detection model across different types of datasets. The accuracy and loss are one of the fundamental metrics used to assess the efficacy of machine learning models, with the accuracy representing the model's overall predictive correctness and loss indicating the model's learning progress during training.

We will present the plots of the training and validation accuracies and losses for each dataset used in our study: the Real-Life Violence, Movies, Hockey, and Custom datasets. These visualizations aim to offer a deeper understanding of how our model performs on each dataset and it highlights any trends or patterns in our model's behaviour.

The graphs of accuracy and loss plots for each individual

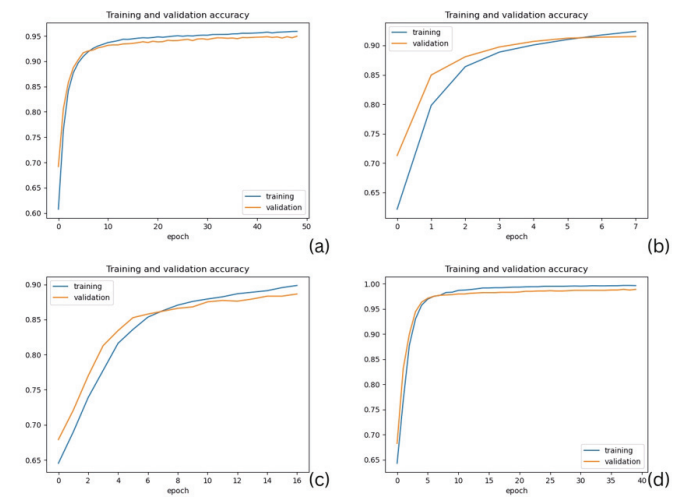

Fig. 8. Visualizations of the accuracy for the proposed model over each dataset. (a) Real life violence situations dataset. (b) Confusion matrix obtained for Hockey dataset. (c) Confusion matrix for Movies dataset. (d) Performance results of our own dataset

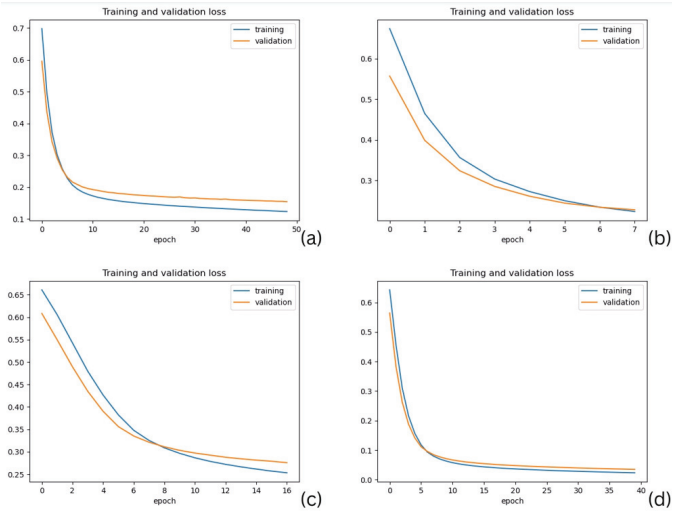

Fig. 9. Loss graphs of the proposed model across 4 different datasets. (a) Real life violence situations dataset. (b) Hockey dataset. (c) Movies dataset. (d) Our custom dataset.

dataset are visualized in Figure (8) and Figure (9), respectively. The accuracy plots show the model's ability to correctly classify instances of violence and non-violence over training epochs, while the loss plots illustrate the convergence behaviour and optimization progress of the model. Together, these visualizations offer a comprehensive understanding of the model's behaviour and performance in violence detection tasks.

VI. Conclusion

The primary objective of this study is to detect the violence activities from the surveillance footage using MobileNetV2 to achieve better accuracy. We trained our model using 4 different datasets among which the fourth one was a custom dataset, which we have collected based on violence in our locality. The custom dataset achieved the best accuracy of 99%, followed by 95% in real-life violence situations dataset. The significant improvement in the accuracy in the custom dataset can be attributed to the improvement in the quality of dataset that we fed to the model.

Hence, the model has proven its stability and efficiency be implemented into a practical hardware system to detect violence in the public areas and alert the security personal instantaneously using the alarm bot. With the escalating rates of violence in the society the traditional human monitoring system is getting inefficient to recognize the violence promptly. With this model, security personal can intervene immediately and prevent malicious acts from escalating contributing towards maintaining the peace and security in the society.

References

[1] Yunpeng Chang and Luo Bin. Bidirectional convolutional lstm neural network for remote sensing image super-resolution. *Remote Sensing*, 11:2333, 10 2019.

[2] Qingshan Liu, Feng Zhou, Renlong Hang, and Xiaotong Yuan. Bidirectional-convolutional lstm based spectral-spatial feature learning for hyperspectral image classification. *Remote Sensing*, 9(12), 2017.

[3] Muzamil Ahmed, Muhammad Ramzan, Hikmat Khan, Saqib Iqbal, Muhammad Khan, Jungin Choi, Yunyoung Nam, and Seifedine Kadry. Real-time violent action recognition using key frames extraction and deep learning. *Computers, Materials and Continua*, 69:2217–2230, 07 2021.

[4] Iftikhar Ahmad, Mohammad Basheri, Muhammad Javed Iqbal, and Aneel Rahim. Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. *IEEE Access*, 6:33789–33795, 2018.

[5] Virender Singh, Swati Singh, and Pooja Gupta. Real-time anomaly recognition through cctv using neural networks. *Procedia Computer Science*, 173:254–263, 01 2020.

[6] Joelton Cezar Vieira, Andreza Sartori, Stéfano Frizzo Stefenon, Fábio Luis Perez, Gabriel Schneider de Jesus, and Valderi Reis Quietinho Leithardt. Low-cost cnn for automatic violence recognition on embedded system. *IEEE Access*, 10:25190–25202, 2022.

[7] H. Nguyen. Fast object detection framework based on mobilenetv2 architecture and enhanced feature pyramid. *J Theor Appl Inf Technol*, 15:5, 2020.

[8] Javed Iqbal, Munwar Iqbal, Iftikhar Ahmad, Madini Alassafi, Ahmed Alfakeeh, and Ahmed Alhomoud. Real-time surveillance using deep learning. *Security and Communication Networks*, 2021:1–17, 09 2021.

[9] Fath U Min Ullah, Khan Muhammad, Ijaz Haq, Noman Khan, Ali Asghar Heidari, Sung Baik, and V.H.C. Albuquerque. Ai-assisted edge vision for violence detection in iot-based industrial surveillance networks. *IEEE Transactions on Industrial Informatics*, PP:1–1, 09 2021.

[10] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. pages 4510–4520, 06 2018.

[11] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, PP, 02 2017.