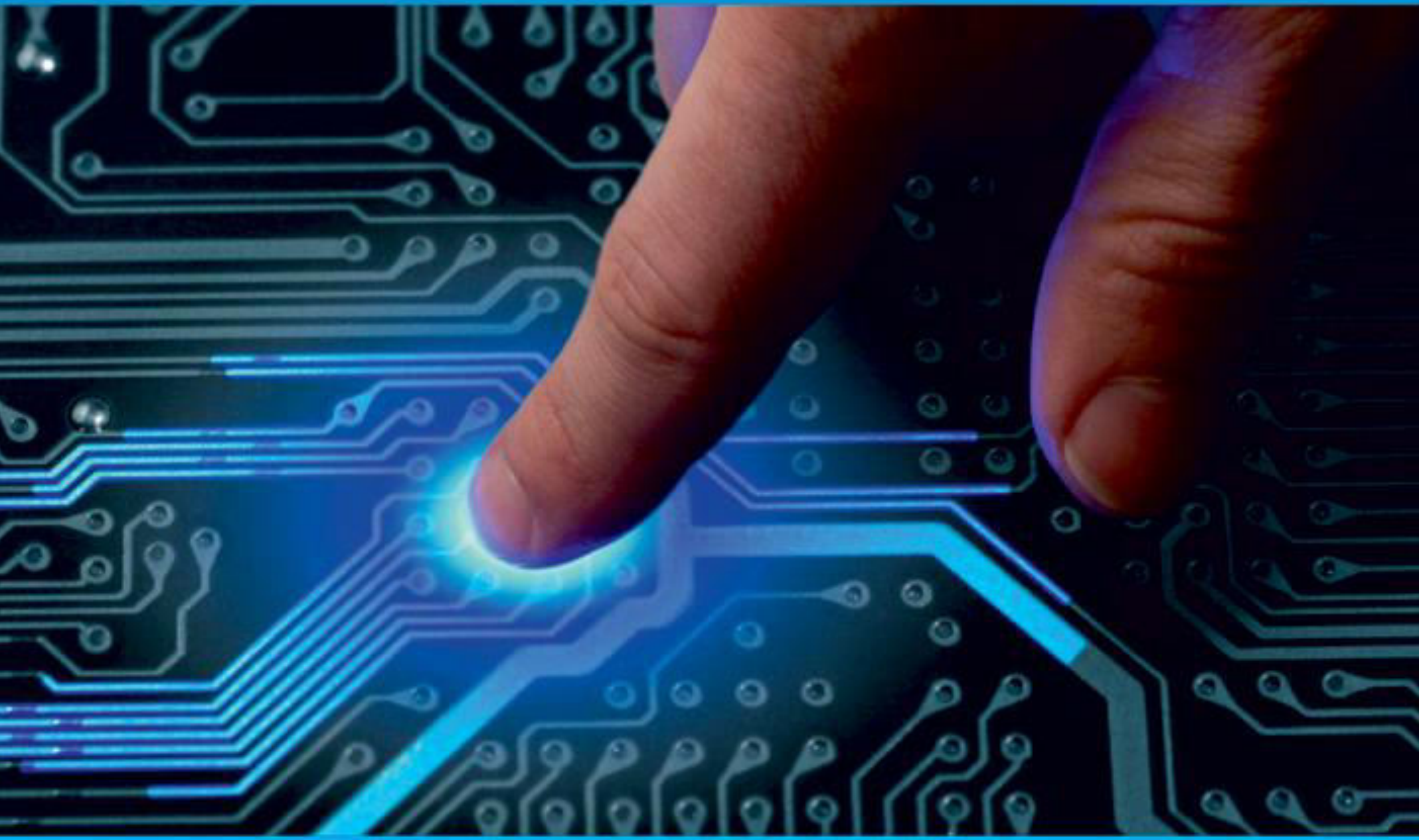
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204030 |

Enriching Violence Detection and Alerting in Public Places using Deep Learning

Anap Prathamesh Rajesh[#1], Ahire Shreyas Milind[#2,] Ghodekar Sujal Santosh[#3,]

Khandagale Anurag Balasaheb[#4,] Prof. Kadlag S.U[#5,] Chaudhari N.K[#6]

Department of Information Technology, Amrutvahini Polytechnic, Sangamner, Maharashtra, India[1-4]

Lecturer, Department of Information Technology, Amrutvahini Polytechnic, Sangamner, Maharashtra, India[5]

Head of Department, Department of Information Technology, Amrutvahini Polytechnic, Sangamner, Maharashtra, India[6]

ABSTRACT: When violence and violent attacks are committed by people with evil and sinister intentions, itcan have a significant impact on the peace and calm of any place. These people intimidate the neighborhood and have the power to seriously hurt both people and property owned by the public. Violence occurs when people locate a gap in the police's patrol and strike while the officers are attending to other matters. These acts are unwanted and can be difficult for law enforcement to address. Because there are not enough officers stationed in a certain region, police forces are able to detect the absence of an automated technique for violence detection. As a result, it has become imperative to integrate the numerous technological breakthroughs to help law enforcement identify violent crimes more quickly and effectively,thereby reducing their frequency. Capturing the live video in an effective manner is the key goal to accomplish the accurate application of deeplearning model like Inception-net for the purpose of detecting violence. With a live video stream, the existing methods are unable to provide an efficient framework for real-time, effective violence identification. Hence, Inception-Net deep learning model is outperformed to provide best result of Violence Detection and to alert the law enforcement agencies for the same via Whatspp Message.

KEYWORDS*:* Deep Learning, Supervised learning,Violence Detection, Inception Net.

I. INTRODUCTION

In recent years, there has been a global increase in the quantity of violent episodes. Since violence causes a great deal of damage and stirs up social dissatisfaction, the rise in violence is undesirable and may even be harmful. The responsibility for maintaining general peace and tranquility across the jurisdiction has fallen on the law enforcement agencies. Violence is committed by people with evil intentions who want to scare the locals and gain control so they can engage in criminal activity. Most of these people commit crimes and are extremely disruptive persons that commit violent crimes.

Additionally, there have been instances where law enforcement has been unable to carry out their duties and reports of illegal acts of force that may be significantly decreased if image processing techniques were used to detect the violence. This can be easily put into practice to improvethe area's safety and security, lighten the workload for law enforcement officers, and support them in situations where they are unable to be physically present.

Consequently, law enforcement agencies benefit from the automation of violence detection through the use of video surveillance that is powered by deep learning technologies. Through the introduction of artificial intelligence protocolsusing deep learning—a tool that can be incredibly useful in a wide range of circumstances—the implementation aims to help police officers and other personnel identify violent scenarios in the surveillance videos.

In [1] the research that is being presented, Anuja Jana Naik et al. tackle the task of a single person violent activity detection system employing selectively pre-trained models. By using the area under the receiver operating characteristics curve, the single person detection was achieved. Ten sample result frames from the videos that were identified using the

[2] Mann B. Patel et al. explored and implemented a variety of deep learning techniques to predict violence in video data. Despite the system's relatively modest GPU power, the implementation performed well on this job. The system discovered that certain training parameters, including CNN network, learning rate, and data augmentation, as well as the clever data preprocessing of the video's frames, play a significant role. As violence scenarios and appliances get more complicated, researchers will need to come up with innovative ways to collect data, develop cutting-edge generalization approaches, and optimize real-time systems. Additionally, the system created a complimentary Flutter app that uses the RTSP Protocol to identify violent activity in CCTV or IP Webcam video streams. Even though distributed computing drastically reduces processing times, the system still operates in pseudo-real time rather than real time when making deductions. Though I'm optimistic that the system will reach the real-time inference milestone in two to three papers while keeping the same level of precision.

[3] Bruno Peixoto et al. discussing the importance of violence detection for numerous applications and how difficultit is to achieve high accuracy in a generalized model due to its high subjectivity. The system's approach is to prioritize violence detection by incorporating more specific ideas about what constitutes a violent setting. This makes it possible to identify the type of violence present in the scene as well as its identification. Additionally, it streamlines the ideas that the neural network must process, which facilitates easier comparisons between the experiments. System found that concepts directly associated to movement, such fighting and explosions, performed better when System used two networks designed to identify elements related to time and movement. More static concepts, like blood and chilly arms, performed similarly well in terms of accuracy. In the future, System can solve this challenging problem even further by merging the features learned in these networks with another network that is better suited for object detection in still photos. System’s studies have already demonstrated that it is more appropriate to specialize a network in detecting violence than to train it to recognize a higher-level notion (Our results of 63% vs. 56%). If necessary, the system can also train various concepts in various network models. The fact that CNN-LSTM had a greater true-positive rate and C3D a better true-negative rate suggests that combining the two networks might be able to attain a higher accuracy. Future research will also focus on refining this combination to increase the accuracy of both the separate concepts and their fusion.

[4] The literature has looked into Nouar AlDahou et al. in great detail. Violence video surveillance powered by IoT has recently been included as a clever feature in smart building security systems. A particular type of detection model called a violence video detector needs to be extremely accurate in order to boost sensitivity and lower false alarm rates. In this paper, a unique end-to-end CNN-LSTM (Convolutional Neural Network - Long Short-Term Memory) model architecture that can operate on inexpensive Internet of Things (IOT) devices, like Raspberry Pi boards, is proposed. The study used CNN to extract spatial information from video frames, which were then fed into an LSTM to classify videos into violent and non-violent categories. For model training andevaluation, a complicated dataset comprising the RWF-2000 and RLVS-2000 public datasets was used. The difficult video footage features fleeting motion, low resolution, small objects at a distance, and throngs of people. In addition, the movies showed people eating, playing basketball, football, tennis, and swimming in a variety of settings, including streets, prisons, and schools. The experimental findings demonstrate the effectiveness of the suggested violence detection model, with average metrics showing 73.35 percent accuracy, 76.90 percent recall, 72.53% precision, 74.01% F1 score, 23.10 percent false negative rate, 30.20% false positive rate, and 82.0 percent AUC. The suggested CNN-LSTM can be installed on an inexpensive IOT node since it balances high performance with few parameters.

This research paper's second section, known as the literature survey, is devoted to a review of earlier works. Additionally, a general description of the implemented approach is provided in the section proposed methodology, which contains three sections. In Section 3, the experiment's results are examined. This study concludes with Section 5, which offers opportunities for future development.

II. LITERATURE SURVEYs

The architectures of three deep learning-based models for violence detection in videos were presented by Paolo Sernani et al. [5] The models were tested on clips from the new AIRTLab dataset, which was created especially to test the robustness against false positives, as well as on the Hockey Fight and Crowd Violence datasets, which are commonly used in literature as benchmarks for violence detection methods. The studies conducted in this paper allow for the derivation of two main conclusions: on all three tested datasets, the proposed transfer learning-based models (C3D combined with an SVM classifier and C3D combined with new fully connected layers) achieve stable accuracy results, often outperforming related studies tested on the Hockey Fight and Crowd Violence.

[6] To create a workable violence detection system, Min-seok Kang et al. suggested spatiotemporal attention modules and the frame-grouping approach. MSM was proposed to derive salient regions from motion boundaries for the purpose of spatial attention. The T-SE block, which the system introduced, could recalibrate temporal aspects with a limited number of extra parameters, so providing temporal attention. Specifically, the technique of grouping three successive channel-averaged images as an input for a 2D CNN was introduced: frame-grouping. It was able to accurately simulate short-term dynamics, which was a necessary component for categorizing aggressive behaviors like striking and kicking. Through a number of studies, the system proved the effectiveness of its suggested modules with effective 2D CNN backbones, and it was able to successfully deploy a real-time violence recognition system in an environment with limited resources. To train a more resilient model, the system will gather additional data in the future and investigate a range of data augmentation strategies. Additionally, for a more flexible application, the system will expand its efforts to handle arange of action recognition jobs.

[7] A suggested end-to-end deep learning neural network for aggression action recognition and detection was made by Mostafa Mohammed Moaaz et al. The preprocessing stage of the suggested approach consists of distributing specific frames throughout the video clip. Convolutional neural networks are used for extraction of spatial features. LSTM is used for temporal feature extraction, and the output of the model is the spatiotemporal features. Lastly, a fully connected neural network is used in the classification step to categorize the videos into violent and non-violent clips. A freshly released RLVS dataset and one of the most popular hockey fighting datasets were used to test the model. The model displayed competitive results when compared to relevant research.

[8] Soheil Vosta et al. discuss how violent behavior is a serious problem that poses a threat to any community. In orderto maintain public safety and lessen potential harm, numerous organizations have employed security cameras to watch over such events. Although it is challenging for human operators to manually watch the large amount of video feed, automated solutions are used to improve the precision and lower mistake rates in violence detection. In this study, the system proposes anovel model called KianNet, which combines the architecturesof ResNet50 and ConvLSTM with a multi-head self-attention layer to efficiently detect violent episodes inside recorded events. Robust feature extraction is made possible by the use of ResNet50, and exploiting the temporal dependencies in the video sequences is made simpler by ConvLSTM. Moreover, the multi-head self-attention layer improves the model's discriminatory power and focus on pertinent spatiotemporal regions. Research studies verify that the suggested model performs around 10% better than its rivals, with a 97.48% AUC for binary classification on the UCF-Crime dataset and a96.21% accuracy on the RWF dataset, which is higher than Violence 4D.

[9] According to Liang Ye et al., campus violence is the most damaging kind of bullying that occurs in schools and is a widespread societal issue worldwide. There are a number of potential ways to identify campus violence as artificial intelligence and remote sensing technology advance, includingmovement sensor-based approaches and video sequence-basedapproaches. Campus violence is detected through the use of sensors and security cameras. The authors of this work employboth auditory and visual cues to detect violence on campuses. Role-playing is used to collect data on campus violence, and 4096-dimension feature vectors are taken out of each of every 16 video picture frames. When features are extracted and classified using the C3D (Convolutional 3D) neural network, an average recognition accuracy of 92.00% is attained. Three speech emotion databases are used to extract mel-frequency cepstral coefficients (MFCCs) as acoustic features. Theaverage recognition accuracies of the C3D neural network, which is used for classification, are 88.33%, 95.00%, and 91.67%, respectively. The authors provide an enhanced Dempster–Shafer (D–S) algorithm to address the evidence conflict issue. The enhanced algorithm raises recognition accuracy by 10.79% in comparison to the current D-S theory, and it can eventually reach 97.00%.

[10] This research presents an innovative and effective strategy that Romas Vijeikis et al. introduced for identifying violent events in real-life surveillance film. The suggestedmodel is a spatial feature-extracting network that resembles a U-Net and employs LSTM for temporal feature extraction and classification after MobileNet V2 as an encoder. Parameters inthe model total 4,074,435. The model has a small and quick computing burden because to its architecture. Three separate datasets—Hockey Fights, Movie Fights, and RWF-2000— were used for the five fold cross-validation. Experiments using a sophisticated real security camera

footage dataset based on RWF2000 revealed an average precision of 0.81 ± 3% and accuracy of 0.82 ± 2%. The suggested model is lightweight and low-cost to compute, yet it nevertheless achieved good accuracy. Utilizing the system's model in edge devices or time-sensitive applications is advantageous.

[11] Muhammad Shahroz Nadeem et al. state that violence identification is becoming more and more important in a variety of contexts, including law enforcement, automatic content filtering, and video surveillance. Current techniques and datasets use highly arbitrary definitions of violence to distinguish between violent and non-violent scenarios. Theavailable datasets, including "Movies" and "Hockey Fight," simply include films of fights versus non-fights; no distinctionis made between weaponry in these datasets. This work presents a novel dataset based on the popular action-adventurevideo game Grand Theft Auto-V (GTA-V), with a specific focus on weapon-based battle scenes. The "Weapon Violence Dataset" (WVD) is the name of this new dataset. The decision to use a virtual dataset is in line with a trend that makes it feasible to create and label as many complex, realistic, high-volume datasets as feasible. Additionally, WVD avoids the potential ramifications and disadvantages of having access to actual data. As far as the system is aware, there isn't a comparable dataset that documents violence involving weapons. The suggested dataset is assessed in the paper using an SVM classifier with local feature descriptors. To classify weapon-based violence movies, the collected features are aggregated using the Bag of Visual Words (BoVW) technique. According to the system's results, SURF performs the best.

[12] Viktor Denes Huszar et al. discuss how the widespread adoption of digital video capturing, storing, and processing technologies has led to an increase in the deployment of surveillance cameras globally. However, due to the massive amount of video data collected, real-time processing by people is challenging, and even manualprocedures may result in events being detected later than intended. In order to overcome this difficulty, automatic violence detection in surveillance footage has drawn a lot of interest from the scientific community. Machine learning algorithms have advanced to the point that it is now possible to do automatic video recognition tasks, such violence detection. In order to capture the temporal and spatial structure of the data, the study's systems investigate the usage of smart networks to model the dynamic relationships between actors and/or objects using 3D convolutions. Additionally, for effective and precise violence identification in surveillance film, the system makes use of the knowledge acquired by a pre-trained action recognition model. The system expands and assesses multiple public datasets with a variety of difficult and demanding video content in order to gauge how well the suggested strategies work. According to the system's results, its approach beats state-of-the-art techniques and achieves a 2% accuracy improvement while requiring fewer model parameters. Furthermore, system studies show how resilient the system's approach is to typical compression artifacts foundin applications using remote server processing.

[13] A brand-new TSODL-VD technique for automatic violence detection in surveillance footage was presented by Ghadah Aldehim et al. It can be used as a preventative step to stop any chaotic circumstances and aid in the automatic and correct recognition of violence. The TSODL-VD technique that is being described utilizes the ResidualDenseNet model for generating feature vectors and the SAE model for classifying events into furious and non-fierce categories. As a hyper parameter enhancer for the residual-DenseNet model, the TSO protocol is used to increase the violence detection efficacy of the TSODL-VD process. Using the benchmark violence dataset, the TSODL-VD procedure's performance validation is investigated. The experimental findings show that, in comparison to the most recent state-of-the-art methods, the TSODL-VD technology achieves accurate and quick detection results.

[14] Mahmoud Abdelkader Bashery Abbass et al. discuss how it can be difficult to detect violence in surveillance footage since it needs to extract spatiotemporal information from a variety of video environments and video perspective scenarios. In light of this, many designs are suggested in this work to carry out this task with excellent performance, utilizing the UBI-Fights dataset as a thorough case study. Convolutional Block Attention Modules (CBAM) are the foundation of the suggested architectures. These modules can be combined with other basic layers, such as ConvLSTM2Dor Conv2D&LSTM. Furthermore, to sharpen the focus on the most crucial features during different architectural training, the Categorical Focal Loss (CFL) is used as a loss function. Performance metrics like as Area Under the Curve (AUC) and Equal Error Rate (EER) are mostly utilized to assess the suggested architectures in order to determine their accuracy in identifying violence with low interaction values between classes. The performance findings demonstrate that thesuggested architectures are capable of outperforming cutting- edge methods in terms of results. The Conv2D&LSTM-based architecture, for instance, beats the state-of-the-artperformance and the majority of the other suggested ones, with an AUC value of 0.9493 and an EER value of 0.0507.

III. PROPOSED METHODOLOGY

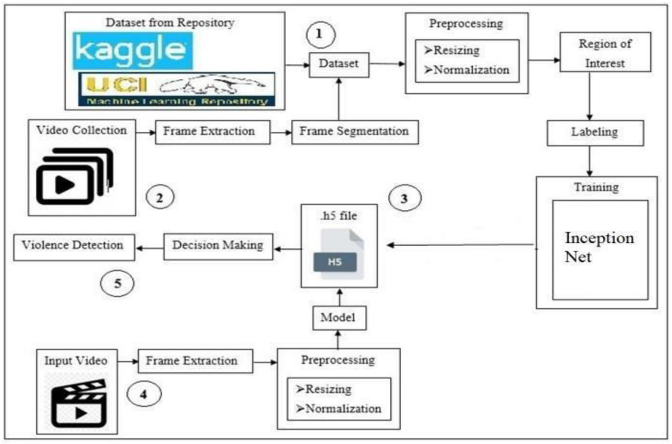
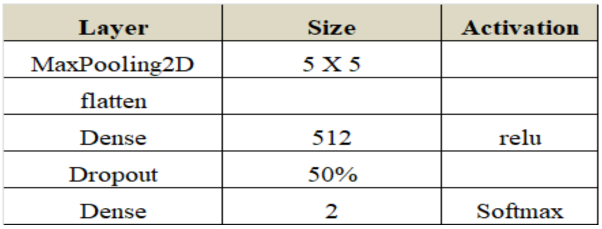


Figure 1: System Overview

Following steps elaborate the process of violence detection through Inception network model as mentioned in Figure 1.
Step 1:Dataset Collection – The proposed model collect the dataset for both violence and non-violence scenarioousing the real time scenario in the different premises like class room, office, park and others. These collected images are stored in the train and test folder with respect to their classes like ‘violence’ and ‘non-violence’.

Step 2: Preprocessing- Once the dataset is being collected it is subjected for the pre-processing process, here the images are being resized to a dimension of 224 X 224. And then they are converted from BGR model to RGB color model to train using the Inception neural network model.

Step 3: Inception Neural network training - In the beginning of the model the dataset images are being categorized as the labels and the image list. This list is then used to partition the data into training and testing splits using 75% of the data for training and the remaining 25% for testing. After this training data augmentation object is created with rotation-range of 30, zoom range of 0.15, width_shift_range range of 0.2, height_shift_range of 0.2. After Image data is generated model is trained with the below architecture mentioned in Figure 2 to obtain the .h5 file containing the trained data. Figure 3 represents the obtained accuracy by the inception net neural network for the given number of epochs



Layer	Size	Activation
MaxPooling2D	5 X 5	
flatten		
Dense	512	relu
Dropout	50%	
Dense	2	Softmax

Figure 2: Inception net Architecture

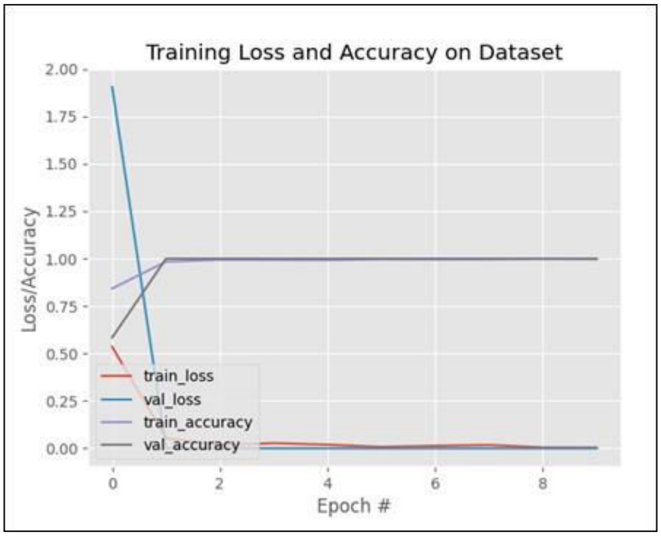


Figure 3: Inception net model Results

Step 4: Violence Detection and alert generation – During the testing phase, the model has been deployed using the Droid Cam application to simulate real-time violence detection. When the suggested model detects violence, it raises a voice alert to put an end to it and sends a Whatsapp message to the appropriate authorities. To evaluate the vulnerability, this message includes the violent image and the map's longitude and latitude coordinates.

III. RESULTS AND DISUSSIONS

Python has been used in development of Inception Net, a deep learning neural network, to accomplish the proposed approach for the purpose of violence detection. The method that is being discussed was developed using the Sypder IDE. The development system has an Intel Core i5 processor, 8GB of RAM, and 500 GB of storage.
Using the Droid Cam application, an external mobile camera was used to conduct the experiment and assess the violence detection system. Droid Cam's dual application model allows for the streaming of mobile camera frames to a laptop. In order to correctly detect the violence scenario using an Inception net neural network, the primary model needs tobe evaluated. A performance evaluation is necessary to identify any mistakes made when putting the model into practice. A description of the assessment process can be foundbelow.

Performance Assessment with the Root Mean Square Approach

A number of investigations have been carried out in order to quantify the error generated by the violence identification system that uses Inception neural networks, as stated. It is easy to examine the performance metrics because of the error that the approach for accurately recognizing the violence scenario accomplishes.

The Root Mean Square Error, or RMSE, is used to make it possible to calculate the error that the method that is being presented achieves. The inaccuracy in the offered wayto violence detection using Inception Net indicates theperformance accuracy of the suggested strategy. The RMSE approach simplifies the error evaluation between two continuously correlated metrics. The degree of Violence identification accuracy and inaccuracy are the factors that are taken into consideration for this procedure. After thesedata have been evaluated, equation 1 is used to calculate the error.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{1,i} - x_{2,i}\right)^2}{n}}$$

Where,
∑ - Summation

$(x_1 - x_2)^2$ - Variations Squared for the total of the differences between the number of violent identifications that were achieved and those that were anticipated

n - Number of Trails

These two variables are measured using ten distinct scenarios in different premises with 5 trails each to detect the violence in the respective location. The mean square error is estimated and tabulated and recorded in table 1.

Premises	No of trails	No Violence Detection	MSE
Office	5	4	1
Class room	5	5	0
Corridor	5	5	0
Park	5	3	4
Campus	5	4	1
Canteen	5	4	1
Play ground	5	5	0
Living Room	5	5	0
Street	5	5	0
Gym	5	3	4

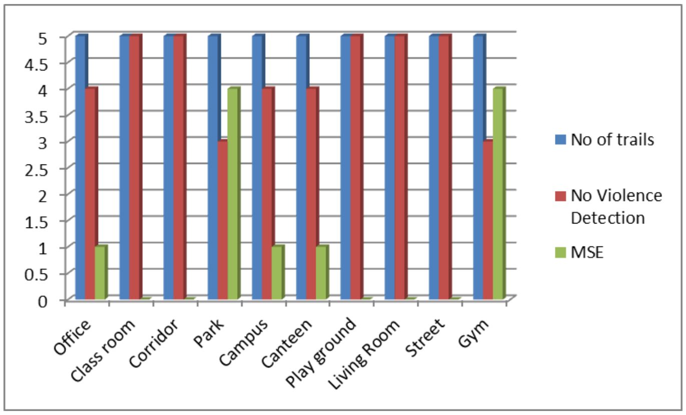


Figure 4: Comparison of MSE in between ExpectedNo of Violence identifications V/s No. of trail conducted

The findings of the method's experimental evaluationhave simplified the graphical depiction of error rate shown in figure 4 above. The graph displays the minimum degree of error the technique experienced when interpreting the violencedetection procedure. This is explained by the precise application of the Inception neural network, which greatlyimproves detection accuracy. The decision-making process also improves the results, as seen by the MSE and RMSE values of 1.1 and 1.048, respectively. This evaluation demonstrates how precisely and accurately a deep learning neural network inception net was used to construct the violence detection model.

IV. CONCLUSION AND FUTURESCOPE

The methodology begins with the realization of the input dataset that will be used to train the Inception net model. The preprocessing module will be applied to the dataset after it has been created. To speed up the Inception neural network model's training process, the preprocessing module resizes and normalizes the input images. Following their preprocessing, these images are used to evaluate regions of interest that will be annotated in order to extract violent interactions from the input images. The region-interested images will be efficiently tagged before being given to the model for training. The Inception network is being used for training, and real-time deployment with external camera gear is being used for testing. Using the RMSE factor, experiments are conducted to assess the legitimacy of the model. After careful examination, the system offers the best RMSE score of 1.048.

For the future deployment this model can be deployed in the real life CC TV cameras for streaming their data via cloud storage to curb the menace of violence in the society.

REFERENCES

[1] Naik, A.J., Gopalakrishna, M.T. Deep-violence: individual person violent activity detection in video. Multimed Tools Appl 80, 18365–18380 (2021).https://doi.org/10.1007/s11042-021-10682-w

[2] Mann B. Patel, "Real-Time Violence Detection Using CNN-LSTM,"April 21, 2021, Mann, GJ https://doi.org/10.48550/arXiv.2107.07578

[3] B. Peixoto, B. Lavi, J. P. Pereira Martin, S. Avila, Z. Dias and A. Rocha, "Toward Subjective Violence Detection in Videos," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 8276-8280, doi:10.1109/ICASSP.2019.8682833

[4] N. AlDahoul, H. A. Karim, R. Datta, S. Gupta, K. Agrawal and A. Albunni, "Convolutional Neural Network - Long Short Term Memory based IOT Node for Violence Detection," 2021 IEEE International Conference on Artificial Intelligence inEngineering and Technology (IICAIET), Kota Kinabalu, Malaysia,2021,pp.1-6,0.1109/IICAIET51634.2021.9573691.

[5] 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/ACCESS.2021.3131315.

[6] M. -S. Kang, R. -H. Park and H. -M. Park, "Efficient Spatio-Temporal Modeling Methods for Real-Time Violence Recognition," in IEEE Access, vol. 9, pp. 76270-76285, 2021, doi: 10.1109/ACCESS.2021.3083273.

[7] Mostafa mohamed moaaz, Ensaf Hussein Mohamed, "Violence Detection in Surveillance Videos Using Deep Learning," in INFORMATICS BULLETIN Access,
Published Online Vol 2 Issue 2, October 2020

[8] S. Vosta and K. -C. Yow, "KianNet: A Violence Detection Model Using an Attention-Based CNN-LSTM Structure," in IEEE Access, vol. 12, pp. 2198-2209, 2024, doi: 10.1109/ACCESS.2023.3339379.

[9] Liang Ye, Tong Liu, Tian Han, Hany Ferdinando, Tapio Seppänen, Esko Alasaarela, "Campus Violence Detection Based on Artificial Intelligent Interpretation of Surveillance Video Sequences ," *Remote Sens.* 2021, *13*(4), 628; https://doi.org/10.3390/rs13040628.

[10] Romas Vijeikis, Vidas Raudonis, Gintaras Dervinis, "Efficient Violence Detection in Surveillance," Sensors 2022, 22(6), 2216; https://doi.org/10.3390/s22062216

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

📱 9940 572 462 ☎ 6381 907 438 ✉ ijircce@gmail.com

www.ijircce.com

Scan to save the contact details