iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59027

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Real Time Violence Detection and Alert System

Manjit Kumar Gautam[1], Prateek Kumar Rajput[2], Yashaswani Srivastava[3], Dr. Anuranjan Kansal[4]
Electronics and Communication Engineering JSS Academy of Technical Education, Noida India

Abstract: ***This paper talks about a précised safety model incorporating elements of Artificial Intelligence for real-time violence detection along with Alert System. This model utilizes the advancements in technology to rapidly respond to any potential violent assault incident. When these technologies are further developed, it also increases the possibilities of how they could be used in future to safeguard schools, public area safety, personal safety, and social stability. Numerous researches and trials have been conducted to counter violence with the passage of time that includes the installations of surveillance systems for warning or alerting to violent activities. Its main objective is to get surveillance systems to automatically annotate the violence activities and issue any Warning/Alerts. For this purpose, first of all the system proceeds with the process of foregrounding a person in each frame, then relevant frames are extracted and irrelevant are ignored, after this violent pattern is identified by the trained model is detected, and end up saving these frames as images. The image is then enhanced and the corresponding details like time and location are transmitted as an alert via Telegram app. The proposed technique is essentially based on deep learning for automatic violence detection using Convolutional Neural networks (CNN). For this diagnosis, a light weight pre-trained model, MobileNetV2, is used to ensure better accuracy than an independent CNN which requires massive computation time and reduced precision.***

Keywords: Real-time violence detection, Alert systems, Surveillance, Convolutional Neural Network (CNN), Image enhancement, MobileNetV2.

I. INTRODUCTION

In our lives we often witness incidents of violence occurring without preventive measures being taken. This leads to injuries, for those involved in conflicts and even bystanders in crowds can be impacted. Traditional surveillance methods may not always catch criminal activities in a manner. That's why it's crucial to update surveillance techniques to keep up with the times. This research project aims to fill this gap by introducing a Time Violent Detection and Alert System, which combines machine learning algorithms and smart alert systems to enhance safety. By using computer vision technology the system analyzes video feeds from security cameras to identify behaviors and potential threats. When a threat is detected the system can quickly alert security personnel or law enforcement for action.

Recent advancements in learning have proven effective in extracting temporal features from videos capturing both movement and detailed spatial information across frames. This study specifically focuses on implementing a Real Time violence alert system utilizing MobileNetv2. After processing the model outputs the enhanced frames are stored in the Firebase database. Face detection is performed using MTCNN and Pyplot on these images. The frames along with details such as the time and location of an incident are then sent as an alert to the police station through the systems module.

In this paper's context, an in-depth into real-time violence detection systems is showcased, covering technologies, uses, challenges, and moral considerations. Through an examination of existing research and emerging patterns in this field, the goal of this publication is to enhance the understanding of the capabilities of real-time violence detection systems in enhancing public safety and security. Furthermore, it aims to address the complex ethical and social issues associated with their deployment.

II. MOTIVATION

The main motivation behind this project is to develop a system that can address the issue of Violent behavior in public places.

Violence is a social problem that has devastating effects over society. Real-time detection of violence is challenging due to the need to go through huge amounts of surveillance video data from various locations and different camera devices, for which violence activity may last for just a few seconds. So, Intelligent Video Surveillance methods are required to detect violence and prevent criminal activities accurately as well as in Real-time. This will aid in quick decision making. Recent researches and studies have also highlighted the accuracy of deep learning approaches to violence detection.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

III. RELATED WORKS

Real-time violence detection has seen a rise in research interest, with different studies, approaches and advances to construct effective processes for recognizing and dodging acts of violence in real time circumstances. A few of strategies that are connected to identify violence are examined below:

A. Effective Two-Dimensional(2D) CNN Modeling Method

Motion Saliency Mapping (MSM) space, 2D Convolutional Neural Networks (CNNs) with frame-grouping, and the Temporal Squeeze-and-Excitation unit is included in this technique. The application of spatiotemporal center strategies and essential CNN structures is the most component of this method. To decrease computational complexity compared to conventional 3D-CNN approaches and increase efficacy and exactness within the identification of savage behaviors, is the main objective.

The Motion Saliency Mapping portion is required in modeling changes that are momentary , whereas the rearrangement in temporal parameters to bring out significant gaps is done in temporal Squeeze-and-Excitation block.

The goal is to highlight human activity in video sequences, strengthening the capacity to identify violent content by including these using video from moving cameras to confirm that the suggested approach works as intended. The study examines many aspects, in order to achieve optimal performance, including suitable 2D CNN designs and frame-grouping lengths. Furthermore, Grad-CAM data analysis highlights the interpretability of the model by emphasizing the importance of geographical points in detecting violence.

Spatio-temporal modeling is illustrated by the method for real-time violence detection in many datasets, such as Real Life Violence Situations(RLVS) and RWF-2000. The evaluation`s findings appeared to be improved.

The method illustrates compelling spatio-temporal modeling for real-time distinguishing proof of viciousness in different datasets, counting RWF-2000 and Real Life Violence Situations (RLVS). The evaluation comes about showing moved forward execution compared to current approaches, with the proposed strategy accomplishing the most elevated accuracy over different video datasets amid a 5-fold cross-validation period. Moving forward, directions are proposed for future study, such as gathering more information and exploring information improvement strategies to boost the model's vigor. The negligible weight of the MSM and T-SE square modules, delivering substantial enhancements in detection capabilities,makes them especially promising for viable usage, while protecting computational productivity.

B. Spatiotemporal Features analysis with 3D Convolutional Neural Network

The recommended strategy here for identifying rough behavior in surveillance video comprises three key stages. This stage comprises the MobileNet CNN model for recognizing people, the 3D-CNN model for capturing spatiotemporal highlights, and the softmax classifier for categorizing exercises. The datasets utilized in this methodology are the violent crowd dataset, the hockey fight dataset, and the violence in movies dataset.

An all encompassing deep learning framework with a three-step approach is suggested for spotting violence in surveillance recordings. At first, a set up MobileNet CNN model analyzes the video feed from the camera to identify individuals within the region. Once people are distinguished, an arrangement of sixteen outlines are extracted and handled employing a 3D convolutional neural network model to extricate spatiotemporal characteristics. These attributes are at that point submitted to a Softmax classifier to recognize activities, especially forceful behavior. To provoke activity and anticipate potential hurt, an alarm is dispatched to the closest security unit when violence is recognized.

By utilizing cutting-edge deep learning algorithms for exact and proficient video examination, the proposed arrangement points to speed up the recognizable proof of savagery in observation film. By amalgamating the individual discovery and spatiotemporal highlight extraction capabilities of CNN models, the system increments the precision of violence detection tasks. Moreover, the execution and proficiency of the model deployment stage are progressed by optimizing the prepared model with the OPENVINO tool compartment.

Utilizing state-of-the-art profound learning models and optimization strategies to upgrade the viability and proficiency of the detection process, the proposed strategy offers a comprehensive approach to recognizing violence in recordings. Results demonstrate that this strategy outperforms elective strategies, accomplishing an exactness of 99.9% on the violent crowd dataset, 98% on the hockey fight dataset, and 96% on the violence in movies dataset. These precision rates emphasize the efficacy of the approach in violence detection, reinforcing security measures and anticipating potential occurrences.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

C. Lightweight mobile network for real-time violence recognition

The paper introduces MobileNet-TSM, a low-weight model intended for violent crime detection in real time. MobileNet-V2 delivers great accuracy while maintaining a small size by combining temporal shift modules. On public datasets, MobileNet-TSM performs competitively when compared to current techniques. The principal reason it is suitable for mobile devices is its lower complexity and parameter count. The study does, however, mention certain incompatibilities with various operating systems. Method 1 extracts spatio-temporal statistics for violence detection using temporal shift modules and depthwise separable convolutions. In order to improve accuracy by strengthening the model's capacity to extract spatiotemporal data, method 2 adds temporal shift modules to the baseline model. Compared to method 1, method 2 exhibits more advanced accuracy improvements. Overall, MobileNet-TSM represents a promising development for efficient violence detection on mobile platforms.

D. Low-Cost CNN for Automatic Violence Recognition

This study looked at how well mobile CNN models could spot violence on a cheap device. They found that MobileNet-v2 was the best, beating other models like SqueezeNet and NASNet, with an accuracy of up to 92.05%. Also, these models could process things quickly, which means they could be useful in real-life situations.

1) *SqueezeNet:* SqueezeNet is a lightweight CNN architecture designed for efficient version inference on resource-restrained gadgets. It pursuits to reduce the variety of parameters even as retaining high accuracy, making it suitable for such systems.
2) *MobileNet:* MobileNet is any other light-weight CNN structure optimized for cell and embedded devices. It makes use of depthwise separable convolutions to lessen computational complexity whilst maintaining overall performance, making it well-desirable for actual-time applications on low-electricity platforms.
3) *NASNet:* NASNet (Neural structure seek community) is a neural network architecture designed through automatic architecture search techniques. It pursuits to locate optimal network structures for specific tasks,but its complexity may hinder efficient deployment on low-cost embedded systems.

E. Face Detection using MTCNN

MTCNN(Multi-Task Protruded Convolutional Neural Networks) is a neural network which detects faces and facial keypoints on filmland. MTCNN is one of the most habituated and accurate face discovery tools. It consists of 3 neural networks connected in a cascade. It uses two Libraries matplotlib, for conniving of images and boxes and MTCNN, which is a perpetration of the MTCNN face sensor for Keras in Python3.4.

It principally uses equals, or the pixel values of a cube where the MTCNN algorithm detected faces. The box value above returns the position of the whole face, followed by a confidence position and other facial milestones, called crucial- points as well.

The exploration introducesMulti-task Protruded Convolutional Networks(MTCNN) as a frame for face discovery and facial corner alignment. It explains the three stages of MTCNN: the Offer Network(P- Net), the Upgrade Network(R- Net), and the Affair Network(O- Net). Each stage employs convolutional neural networks(CNNs) to precipitously upgrade the discovery process and affine facial corner positions. MTCNN's objects include face/non-face brackets, bounding box retrogression, and facial corner localization. The exploration outlines the loss functions employed for each task and provides visual representations of the networks. Likewise, it hints at forthcoming exploration agitating the development of a face discovery operation using MTCNN and OpenCV.

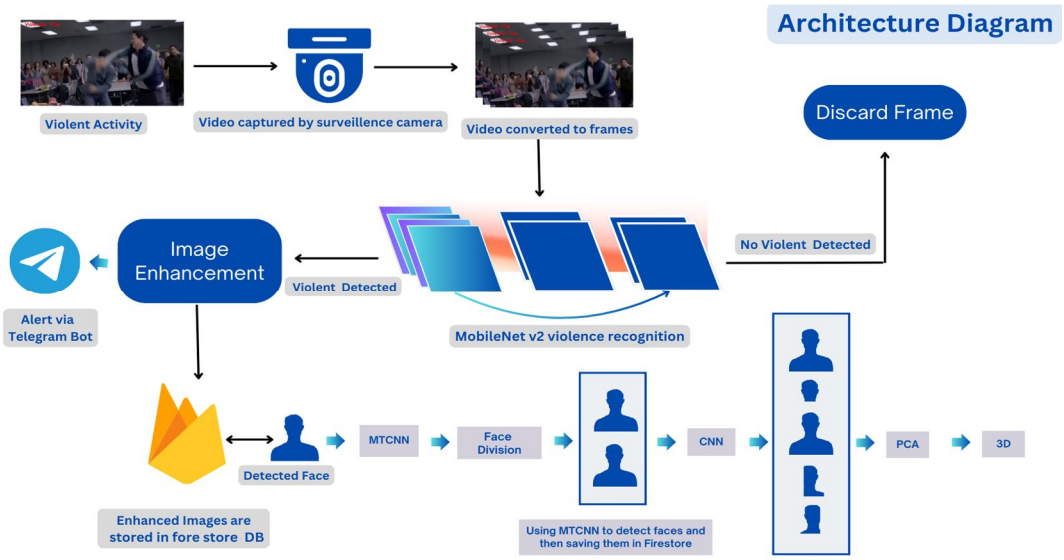
F. Unsupervised spatial–temporal action translation network (STAT)

The method adopts the problem of monitoring and observing human behavior in surveillance systems, especially in the detection of violent actions. This is hindered by the limited data for training deep networks and the complexities of human behavior. To conquer this, the authors recommend an unsupervised Spatial-Temporal Action Translation (STAT) network. The framework contains a person spotting device, motion feature extractor, STAT network, and output interpretation. By effectively discarding unnecessary background information and concentrating on temporal features vital for recognizing rapid changes in violent motion patterns, the framework operates well in different environments. Trained with normal behavior data, the STAT network translates normal motion into spatial frames but struggles to accurately reconstruct violent frames because of their intricacy. Nonetheless, actions are categorized by comparing genuine and reconstructed frames and determining reconstruction mistakes. This unsupervised approach attains comparable precision and surpasses previous works in terms of generality.

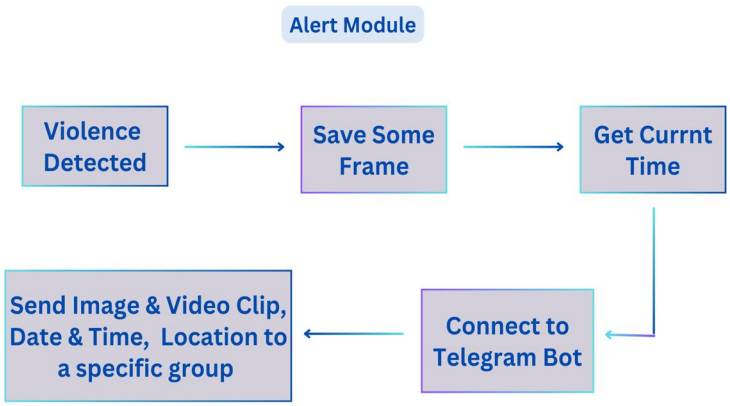
IV. METHODOLOGY

The model is a comprehensive methodology that includes following:

1) *Dataset Collection:* The Dataset taken here is of 1000 videos, which is fed to the model, having an equal number of violent and non-violent videos. It is one kind of unique data having different types of violent and non-violent instances.
2) *Pre-processing:* Methods such as image resizing, normalization, and augmentation to ready the dataset for training purposes are implemented. The dataset is of 1000 video clips partitioned into two groups: violence and non-violence. The video clips last around 5 seconds each on average, mainly sourced from public CCTV cameras. During training, 350 videos from both violent and non-violent categories are selected at each epoch, totaling to 80 epochs, with 67 identified as the most optimal. Employing real-time image pre-processing techniques on incoming video frames.
3) *Libraries used:* The major libraries involved in constructing the model architecture include Tensorflow, Keras, OpenCV, and matplotlib.
4) *Model training:* Employing a MobileNet V2 framework specially optimized for detecting violence, focusing on essential elements like object identification, motion evaluation, and contextual analysis of scenes.



5) *Alert Generation:* Immediately notifying security personnel or relevant authorities upon detecting violent behavior. When a frame is flagged as indicating violence, the counter is activated. If 30 consecutive frames are identified as violent, the event is classified as such, triggering an alert. The alert is transmitted via a bot using the messaging app - Telegram, including a warning, location, time, date, as well as a captured image and short video clip of the incident.



6) *Visualization:* It is the representation by storing all violence alerts along with their enhanced images in the Firebase Database for record-keeping and future review.

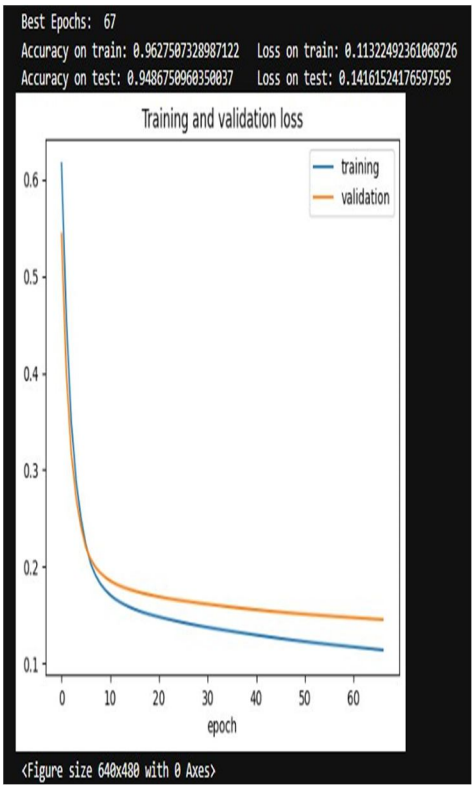


Fig.1. Screenshot of the alert message

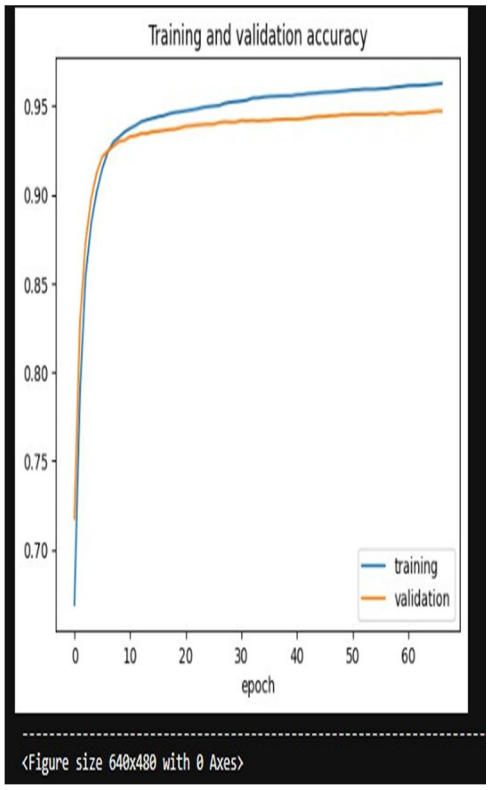
V. EXPERIMENTAL RESULTS

Numerous experiments and tests were carried out to assess the effectiveness of this model in real-world situations. Findings were gauged utilizing standard criteria, including:

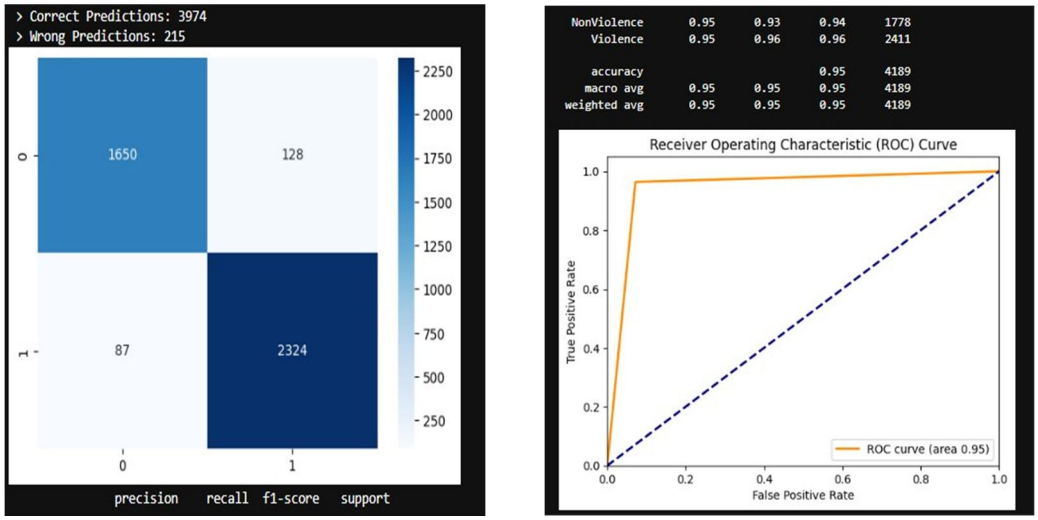
1) *Accuracy:* The system's capability to accurately distinguish between violent and non-violent actions. For the MobileNetV2 model, the accuracy reached 96.2% during training and 94.86% during testing.
2) *Visual Representation:* Figure 2 illustrates the training loss, accuracy, ROC curve, and confusion matrix, showcasing the experimental outcomes that highlight the model's proficiency in correctly detecting various violent activities and excelling in all assessment measures.



(a)



(b)



(c) (d)

Fig.2. Experimental results are demonstrated with the help of following metrics (a) Training and Validation Loss ,(b) Training and Validation Accuracy, (c) Confusion Matrix, (d) ROC Curve

VI. DISCUSSION AND ANALYSIS

The related works for violence detection can be broadly classified into three major groups: visual based methods, audio based methods, and hybrid approach which involves each visual based method and audio based approach.

Visual based method - Violence detection in the visual-based technique is based on the extraction and analysis of visual information, which is represented by crucial features. Those traits may be divided into categories: global features (average speed, region occupancy, relative positional variations, and interactions among objects and the background) and local features such as color, form, position and velocity.

Audio based method - Conversely, the audio-primarily based technique makes use of auditory information to categorize violent acts. This method's strategies often use hierarchical techniques based totally on Hidden Markov fashions and Gaussian aggregate models to differentiate between noises connected to violence, like gunshots, explosions, and automobile braking.

Hybrid method - The hybrid technique enhances violence detection by combining both visual and auditory traits. This method combines techniques from the visual and auditory domain names, together with motion intensity extraction, blood and flame detection in movies, and sound recognition of usual sounds associated with violent events. One instance of a hybrid method is the CASSANDRA system, which analyzes kinematic styles related to human articulation in surveillance motion pictures and recognizes audible cues that resemble screams to identify violence.

When investigation of different models such as 2D CNN, 3D CNN, low-cost CNN, MT-CNN, and MobileNetV2 for real-time savagery location was done it was found that these models have their claim preferences and drawbacks. Comparing these models with a proposed spatio-temporal modeling approach utilizing 2D CNN highlighted the latter's capacity to upgrade violence detection execution while keeping up lower computational complexity compared to 3D CNN-based techniques.

The approach successfully emphasizes human activities, predicts short-term flow, and tunes transient highlights for pertinent interims, coming about in higher real-time viciousness acknowledgment execution. Be that as it may, it isn't reasonable for moving cameras and brings about tall computational costs due to the utilization of three distinctive modules with predefined assignments! The violence detection utilizing 3D CNN with spatial-temporal highlights examination illustrates high exactness on benchmark datasets, beating existing strategies in identifying violent behavior, but the method is truly computationally costly and depends on a lot of parameters, consequently isn't appropriate for real-time situations and for day by day utilization.

In the Low cost CNN approach, MobileNet-v2 was the best, beating other models like SqueezeNet and NASNet, with an accuracy up to 92.05%. Aloso, these models could process things quickly which means they could be useful in real-life situations. MobileNet-TSM, a lightweight model designed for real-time violence recognition, was also analyzed. By integrating temporal shift modules into MobileNet-V2, it achieves high accuracy while maintaining a compact size. Compared to existing methods, MobileNet-TSM demonstrates competitive performance on public datasets. It is particularly suitable for mobile devices due to its reduced complexity and parameter count.

TABLE I. Techniques used in the Deep Learning Model

S.no	Techniques	Methodology used	Advantages	Disadvantages
1	Convolutional Neural Networks (CNN)	Extract facial features, apply convolution layers for hierarchical pattern learning, followed by fully connected layers for emotion classification.	Highly accurate at image recognition tasks, including facial emotion recognition.	Requires a large amount of training data.
2	MobileNet V2	First layer is 1×1 convolution with ReLU6. Second layer is the depthwise convolution and the third layer is 1×1 convolution without any non-linearity.	Lightweight pre-trained model, reduces the Computational cost by using 1x1 convolutions, before applying depthwise convolutions.	It can be 0 after activation due to the small convolution kernel in the depth-separable convolution.
3	3D ConvNet	3D Convolution+Max Pooling	Effective for capturing temporal information, handling volumetric data, and automatically learning hierarchical features.	High computational complexity, potential overfitting due to increased model complexity.
4	SE Layer (Squeeze-and-Excitation Layer)	Squeezing converts convolutional input data to a Single Digit Numerical Value. Excitation scales each of n-channels on an importance basis.	Used to give weights (attention) to each frame of the video.	Introduces additional parameters and computational overhead, leading to increased model complexity and potential training time.
5	Data Pipelining	Storing Pre Processed Data in TF Record files	Tensorflow can read TFRecord files faster than other methods, training will take less time.	Data inconsistency, latency, and error handling challenges persist in data pipe- lining with transfer flow
6	ReLU (Rectified Linear Unit)	An Activation Function, Monotonic Function, returns 0 for negative values, and x for a positive value x.	Introduces property of non-linearity in the model, solves Vanishing Gradient Problem.	Dying ReLU problem may cause Information Loss.

VII. CONCLUSION & FUTURE WORK

A. Conclusion

This study over Real-Time Violence Detection and Alert System marks a significant progression in leveraging deep learning for bolstering security measures.

The system's efficiency, as depicted through the experiments, underscores its effectiveness in identifying violent behaviors across diverse settings. This research not only aids in averting violent incidents but also promotes secure and resilient communities by harnessing machine learning and computer vision capabilities.

B. Future Considerations

Future enhancements could enable the model to function across multiple interconnected cameras simultaneously, integrating audio and other sensor data to enhance detection precision and contextual insight. This advanced system for real-time violence detection contributes to heightened security protocols, allowing for proactive responses to potential threats and enhancing public safety. Embracing emotion data from videos to support violence detection is crucial in distinguishing between harmless actions and genuine threats. Augmenting feature extraction methods to incorporate additional elements into the deep learning framework for a more thorough violence detection approach is recommended. Also, emphasizing ethical deployment and privacy safeguards is essential.

VIII. ACKNOWLEDGMENT

The authors extend sincere gratitude to Dr. Anuranjan Kansal for his mentorship. His expertise played a vital role in shaping the direction and focus of this research.

REFERENCES

[1] Zhang Y, Li Y, Guo S (2022) Lightweight mobile network for real-time violence recognition. PLoS ONE 17(10): e0276939. doi: 10.1371/journal.pone.0276939. PMID: 36315496; PMCID: PMC9621415.
[2] 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.
[3] Ullah FUM, Ullah A, Muhammad K, Haq IU, Baik SW. Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network. Sensors (Basel). 2019 May 30;19(11):2472. doi: 10.3390/s19112472.
[4] J. C. Vieira, A. Sartori, S. F. Stefenon, F. L. Perez, G. S. de Jesus and V. R. Q. Leithart, "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.
[5] P. Sernani, N. Falcinelli, 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.
[6] C. Gu, X. Wu and S. Wang, "Violent Video Detection Based on Semantic Correspondence," in IEEE Access, vol. 8,pp.85958-85967,2020,doi:10.1109/ACCESS.2020.2992617. -May 2020.
[7] Ş. Aktı, G. A. Tataroğlu and H. K. Ekenel, "Vision-based Fight Detection from Surveillance Cameras," 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), Istanbul, Turkey, 2019, pp. 1-6, doi: 10.1109/IPTA.2019.8936070. -February 2020.
[8] B. Jiang, F. Xu, W. Tu and C. Yang, "Channel-wise Attention in 3D Convolutional Networks for Violence Detection," 2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA), Tainan, Taiwan, 2019, pp. 59-64, doi: 10.1109/ICEA.2019.8858306. -August 2019.

Dr. Anuranjan Kansal
Dr. Anuranjan Kansal is a mentor and educator, serving as the Head of the B.Tech ECE Department at JSS Academy of Technical Education, Noida.
Role in the Study:Mr. Kansal, in his capacity as the Chief of the B.Tech ECE Department, played a pivotal role in steering and supervising the study, offering valuable insights and knowledge.

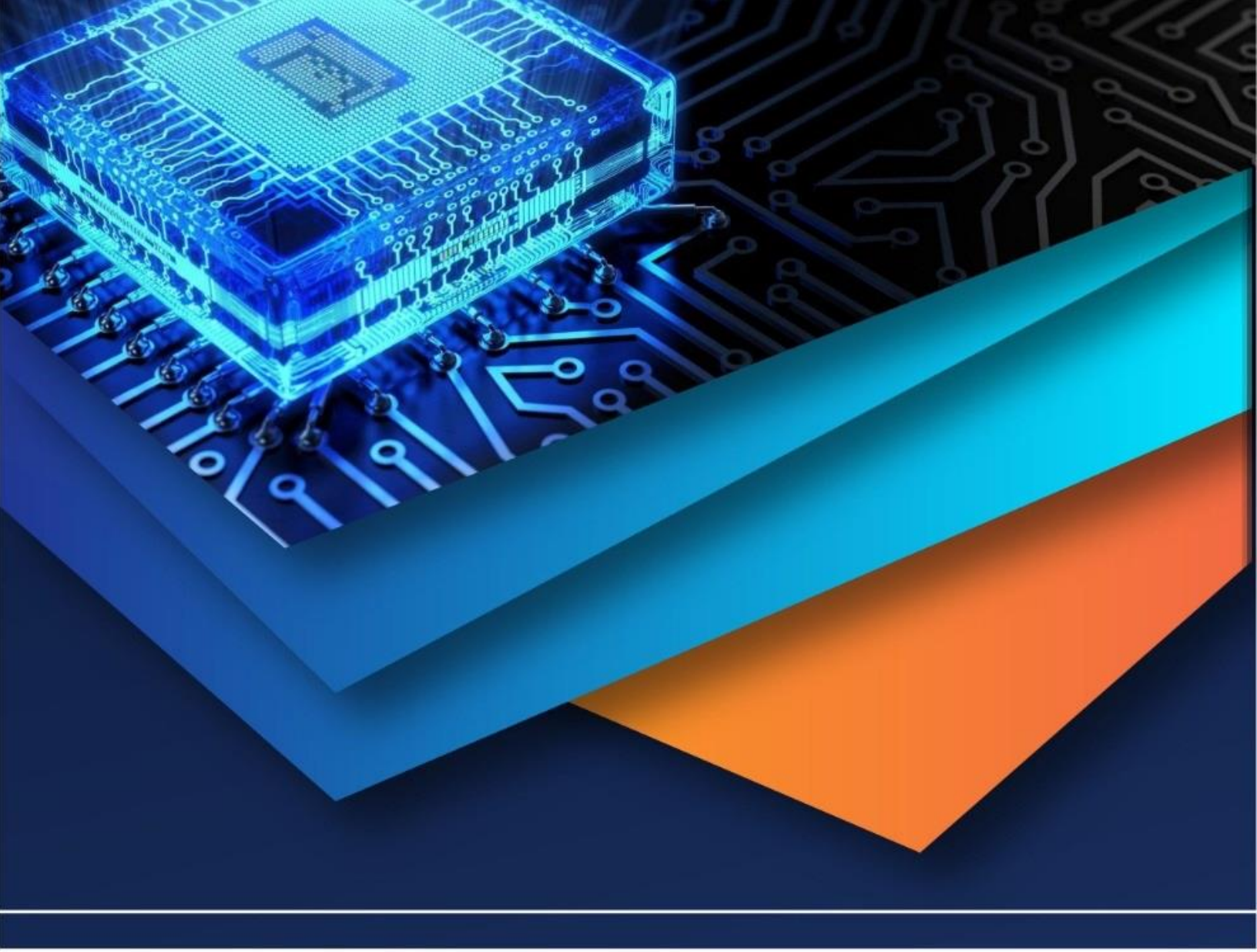
Manjit Kumar Gautam
Manjit Kumar Gautam is currently pursuing his Bachelor's degree in Electronics and Communication Engineering at JSS Academy of Technical Education, Noida, set to graduate in 2024.
Role in the Research: Manjit Kumar Gautam played a significant role in the research project, contributing his skills and knowledge as a B.Tech ECE student at JSS Academy of Technical Education, Noida.

Yashaswani Srivastava
Yashaswani Srivastava is pursuing her Bachelor's degree in Electronics and Communication Engineering (B.Tech ECE) at JSS Academy of Technical Education, Noida, in 2024. She has successfully secured a position at Accenture.
Role in the Research: Yashaswani Srivastava made significant contributions to the research project during her academic tenure, reflecting the practical applicability of her skills developed through the research.

Prateek Kumar Rajput

Prateek Kumar Rajput is currently pursuing his Bachelor's program for Electronics and Communication Engineering at JSS Academy of Technical Education, Noida, with an anticipated graduation year of 2024.

Role in the Research: Prateek Kumar Rajput actively participated in the research as a B.Tech ECE student, offering a fresh perspective and hands-on involvement in the project.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)