iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: IX Month of publication: September 2025

DOI: https://doi.org/10.22214/ijraset.2025.74247

www.ijraset.com

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

A Knowledge Distillation Based Lightweight Deep Learning Framework for Real-Time Violence Detection

Sumedha Arya

Abstract: With the increase of violent incidents in public and private places, it brings and urgent need for the development of advanced and efficient surveillance systems. Traditional methods, rely on manual monitoring, which are impractical and inefficient for violence detection in complex scenarios. This paper presents a lightweight deep learning framework for real-time violence detection, using knowledge distillation to improve public safety and security. We propose a teacher-student model approach, where a large, pre-trained VGG16 model serves as the teacher to transfer knowledge to a significantly smaller, custom-built CNN student model. The methodology involves training the teacher model on a Kaggle based dataset of violent and non-violent images, followed by training the student model using a combined distillation loss function that balances hard and soft targets. Our results demonstrate that the teacher model achieves a high accuracy of 90.96%. The student model, with a remarkable 7.20x reduction in parameters, achieves an accuracy of 83.85%, successfully retaining over 92% of the teacher's performance. This framework offers a convincing trade-off between model size and accuracy, making it an effective, efficient and scalable solution for real-time deployment on mobile devices in smart city surveillance systems.

Keywords: Knowledge Distillation, VGG16, Violence Detection, Convolutional Neural Networks, Loss hard, Loss Soft.

I. INTRODUCTION

In rapid urbanization and technological advancement, the concept of a "smart city" has emerged, offering quality of life for citizens through integrated technologies. A critical component of this vision is the enhancement of public safety and security. Violent incidents, ranging from physical assaults to public fights, pose a significant threat to community well-being and require immediate attention. Manually monitoring the vast and continuous video feeds from surveillance cameras is tough, and often ineffective task due to human limitations.

Recent research has explored various deep learning-based approaches to address this challenge. Several studies have used Convolutional Neural Networks (CNNs) to analyze motion patterns, body poses, and spatial relationships in video streams, achieving high accuracy in real-time detection [1, 2, 4, 6]. Other approaches have incorporated more complex architectures, such as 3D CNNs and CNN-LSTM hybrids, to capture spatio-temporal dependencies, further enhancing performance [5, 8, 9]. The use of multimodal fusion, combining visual and audio data, has also been shown to improve robustness and accuracy [10]. However, many of these high-performing models, like VGG16, are computationally intensive with large number of parameters, making them unsuitable for deployment on resource-constrained devices at the edge of the network. This limitation presents a major hurdle for widespread, real-time implementation in smart city infrastructure.

To overcome the challenge of model complexity and computational demands, this paper introduces a framework that utilizes knowledge distillation (KD). It is a technique to compress a large model to smaller, in terms of parameters. Means, in KD, a smaller, more efficient "student" model is trained to replicate the behavior of a larger, more accurate "teacher" model. This allows for the transfer of valuable knowledge from the complex model to the lightweight one, thereby retaining most of the performance while significantly reducing the model size. This study proposes and evaluates a knowledge distillation framework for real-time violence detection, aiming to create a lightweight, yet highly accurate, model suitable for deployment in smart surveillance systems. As a teacher model, VGG16 is used, while a custom CNN model is utilized as a student model. The VGG16 is a pre-trained model on the ImageNet dataset. It is widely recognized for its robust feature extraction capabilities, which make it a powerful choice for image classification tasks. Therefore, in this research, we preferred VGG16 as our choice been a teacher.

The paper is structured as follows: Section 2 presents a literature review, Section 3 gives the research methodology, Section 4 provides an analysis of the results, Section 5 offers a conclusion and Section 6 gives future work.

II. LITERATURE REVIEW

In this section, we present a detailed and comprehensive review of recent work by various authors on real-time violence detection.

The authors in their research [1], present a comprehensive study on a novel framework for violence detection in video streams. They proposed a CNN based model which analyze motion patterns, body poses, and spatial relationships, achieving high accuracy and efficiency in real-time violence detection. Model performance is evaluated through metrics such as accuracy, precision, recall, and F1-score. This approach offers scalability and adaptability to diverse environments, improving public safety applications. However, there are few limitations, including dependency on annotated data, challenges with extreme conditions such as poor lighting or occlusions, and high computational demands.

The authors [2, 13] in their research, introduced a deep learning-based model for real-time human violence detection in video surveillance systems. They proposed a model which integrates three modules. These modules are: Spatial Motion Extractor (SME), which identifies regions of interest; the Short Temporal Extractor (STE), which captures rapid movement features through temporal fusion and MobileNet V2; and the Global Temporal Extractor (GTE), which enhances precision. The model was evaluated on the RWF-2000, Movies, and Hockey datasets, it achieved state-of-the-art accuracy, further validated on the novel VioPeru dataset.

In similar research, the authors [3, 14] present a deep learning-based approach to detect human violence in real-time CCTV footage. They addressed the global problem of violence, with an annual fatality rate of 7.9 per 10,000 people. In their study, the authors used Inceptionv3 and YOLOv5 models to identify violent actions, with the number of individuals involved, and the weapons used. The proposed model achieved a detection accuracy of 74%. These models are integrated into a local website developed using CSS/HTML as the front end and the FAST API to combine the models. It enables users to upload videos and classify them as violent or non-violent behavior, with face and object detections.

In another research, the authors [4] proposed a framework based on deep learning to detect real-time violence in public places. They used DenseNet121, a pre-trained model to process the Real-Time Streaming Protocol (RTSP) in surveillance systems. The model is trained on a subset of 1,000 labeled frames from the UCF-Crime dataset. It achieved an accuracy of 92 and a 0.91 weighted F1-score, which is supported by data augmentation and class weight balancing to address class imbalance. The system integrates OpenCV for frame capture, Flask for real-time dashboard visualization, MongoDB for metadata tracking, and Dropbox for secure cloud storage.

The authors [5] performed a comprehensive review of 63 articles on AI-based video violence detection. They focused on physical assaults in surveillance systems. This study categorizes 21 unresolved challenges, with camera position being the most prominent, and reviews 28 datasets, noting the Hockey Fights dataset as the most frequently used. It also analyzes 21 keyframe extraction methods and 16 input types of algorithms. The prevalent use of CNN combined with Long Short-Term Memory (LSTM) was highlighted as the most effective technique to detect violence. The review underscores the impact of physical aggression on society and the role of real-time detection in enhancing public safety. In future work, the authors proposed to develop CNN-LSTM-based algorithms and incorporate trustworthy AI techniques to improve decision-making transparency.

The authors [6] explore a state-of-the-art smart surveillance model designed to improve public safety through real-time violence detection using advanced ML, DL and computer vision algorithms. The proposed model uses CNN and other deep learning models for motion analysis, object detection, and behavior recognition. It categorizes violent actions, such as physical assaults and fights, as normal activities in diverse settings, such as public places, businesses, and educational institutions. The model achieves high accuracy and efficiency in controlled datasets, but faces challenges in real-world scenarios due to variable lighting and camera positions.

The authors [7] introduce an innovative approach based on deep learning for the detection and localization of real-time violence in surveillance systems. The research addresses limitations in existing methods by incorporating subgroup analysis. Using the X3D network, the proposed model with an adaptable add-on module tracks multiple subgroups across frames. This approach improves interpretability by identifying groups involved in violent acts. The model achieved an accuracy of 91.3 on the SCFD dataset and 87.2 on RWF-2000. The model also shows generalization on unseen datasets and operates efficiently with eight frames, prioritizing false positives to minimize missed violent events.

In their study, the authors [2, 8] also used deep learning techniques for the detection of violence in industrial surveillance videos. They highlighted the growing importance of IoT-enabled surveillance systems in smart cities and industrial environments. The research proposed a three-stage deep learning framework. Firstly, they used a lightweight CNN for human detection to filter irrelevant frames. Second, they applied a 3D-CNN for spatio-temporal feature extraction from 50-frame sequences. Finally, they used a softmax activation function to classify violence based on different categories. Comparative experiments on four benchmark datasets show superior accuracy and decreased computation time compared to traditional CNN models.

The study by the authors [9] investigated the detection of violence in videos using pretrained models such as VGG16, VGG19, and ResNet50. They perform feature extraction from video frames. A custom dataset of violent and non-violent videos is used and evaluated on multiple techniques. The ResNet50 model, combined with LSTM, achieved the highest accuracy at 97.06, outperforming other models. The study highlights the effectiveness of transfer learning and attention mechanisms in enhancing violence detection.

In another study, the authors [10] presented a multimodal deep learning approach for the detection of violence by integrating audio and visual data from the RLVS dataset. The work used VGGish, Wav2Vec 2.0 and various pre-trained vision transformer-based networks. It focused on combining the VGGish and MobileViT models through late fusion, achieving a classification accuracy of 97.13 and an F1 score of 0.97. Knowledge distillation transfers insights from a fine-tuned ViT (teacher) to a MobileViT (student) model. This enabled efficient deployment on a Jetson Nano edge device at 5–10 frames per second for real-time monitoring. Compared to unimodal approaches, the multimodal framework demonstrated superior accuracy and robustness.

The study by the authors [11] addressed the critical need for advanced security measures due to the increase in violent incidents in public and private places. Inorder to tackle this issue, they proposed a Real-Time Violence Detection and Notification System. This system combines Faster R-CNN and MobileNetV2 architectures for effective and efficient footage analysis in surveillance system. It uses ML and computer vision to accurately distinguish aggressive behavior from non-violent activities. Real-time alerts messages were sent to security personnel through a messenger application; Telegram bot. Comprehensive evaluations demonstrate its superior accuracy and performance compared to existing models. The study emphasizes the limitations of manual monitoring and the need for automated detection to improve public safety.

The research of the authors [12] presents a lightweight deep learning framework for real-time violence detection in smart city surveillance. They also addressed the limitations and issues related to human-operated CCTV systems. The study uses the Smart-City CCTV Violence Detection (SCVD) dataset. The proposed system utilizes MobileNetV2 for the extraction of spatial features from 15-frame video sequences and a stacked LSTM network to capture temporal dependencies in violence and weapon-related activities. The model achieved an accuracy of 99.58, 0.0139 cross-entropy loss, and precision, recall, and F1-scores of 0.99 for non-violent and 1.00 for violent classes.

The existing literature demonstrates the significant progress made in real-time violence detection using various deep learning models. However, due to model complexities there is a clear need for a more lightweight and efficient deep learning framework. This model can maintain high accuracy while being suitable for real-time applications and scalable deployment in smart city surveillance systems.

III. RESEARCH METHODOLOGY

The proposed methodology uses a knowledge distillation framework to train a CNN based model by transferring knowledge from a larger pre-trained model. The dataset used in this study is collected from Kaggle containing images of two categories: *NonViolence* and *Violence*. In image preprocessing, all images were resized to 64×64×3 pixels to standardize input dimensions. The dataset was further divided into training and testing sets with a ratio of 80:20. To improve data handling efficiency, the prepared data were converted into TensorFlow datasets and batched with a batch size of 32.

The teacher–student approach was adopted to reduce model size and computational requirements while retaining most of the predictive accuracy. The teacher model was based on the VGG16 architecture, pre-trained on ImageNet. The fully connected layers of the original architecture were removed, and a Global Average Pooling layer was added to reduce spatial dimensions. This was followed by a dense layer with 1024 neurons present in it with ReLU activation function. A final dense output layer was added with two neurons that uses softmax activation for binary classification. All convolutional layers in VGG16 were frozen to preserve the pre-trained weights, making it non-trainable. The teacher model was compiled using the Adam optimizer, categorical cross-entropy loss, and accuracy as the evaluation metric.

The student model was designed as a lightweight CNN with significantly fewer parameters. It consists of two convolutional layers with filters as 32 and 64, both with 3×3 kernels, 'same' padding, and ReLU activation function, each followed by a max pooling layer. The output was flattened, converting it into 1D array and passed through a dense layer of 128 neurons followed by ReLU units before a final softmax layer with two neurons. The student model was also compiled using the Adam optimizer, categorical cross-entropy loss, and accuracy as a metric.

Knowledge distillation was implemented to transfer information from the teacher model to the student model. A custom distillation loss function was defined as: $\text{Loss\_total} = \alpha \times \text{Loss\_hard} + (1 - \alpha) \times \text{Loss\_soft} \times (T^2)$. Here, Loss_hard is the categorical cross-entropy between the true labels and the student's predictions.

Loss_soft is also a categorical cross-entropy, but it is between the teacher's and student's temperature-scaled softmax outputs. The temperature T was set to 3.0 to produce soft probability distributions, and α was set to 0.7 to balance the contribution of hard and soft targets.

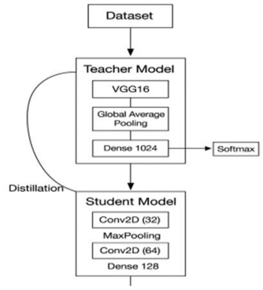


Fig 1: Proposed Teacher Student Model Architecture

The training process was divided into two stages. In the first stage, the teacher model was trained for 10 epochs using the prepared dataset. In the second stage, the trained teacher model provided soft targets for the student model during training. The student was then trained for 10 epochs using the combined distillation loss.

Model performance was evaluated using multiple metrics, including accuracy, loss, confusion matrices, and classification reports with precision, recall, and F1-score. The number of trainable parameters for both models was compared to determine the compression ratio, and the percentage of accuracy retention was calculated to measure how well the student model preserved the teacher model's performance after compression.

IV. RESULTS ANALYSIS

The VGG16 based teacher model, demonstrated strong performance. Over 10 epochs, it achieved a training accuracy of 99.67% by the final epoch. The validation accuracy also showed a consistent increase, peaking at 90.96%. This indicates a bit overfitting in the teacher model, but also showed a positive sign for effective learning in data classification. The model performance was also generalized well to unseen data from the validation set. The final evaluation on the test set confirmed this, with a high accuracy of 90.96% and a loss of 0.4166. The classification report for the teacher model shows balanced precision and recall for both 'NonViolence' and 'Violence' classes, with F1-scores of 0.91 for each, further validating its robust performance.

The training logs for the student model show that its accuracy also improved steadily over 10 epochs. It achieved a final validation accuracy of 83.85%, which is competitive despite its simplified architecture. The final evaluation of the test set resulted in an accuracy of 83.85% and a loss of 0.7609. The classification report for the student model shows a slight drop in precision, recall, and F1-score for both classes compared to the teacher, but the values remain high (0.84 on average), which is a significant achievement for a model with a fraction of the parameters.

TABLE I: Comparative Analysis: Teacher vs. Student Model

Metric	Teacher Model	Student Model
Accuracy	0.91	0.84
Precision (NonViolence)	0.92	0.83
Recall (NonViolence)	0.89	0.85
F1-Score (NonViolence)	0.91	0.84
Precision (Violence)	0.90	0.85
Recall (Violence)	0.93	0.83
F1-Score (Violence)	0.91	0.84
Total Parameters	15,242,050	2,116,930
Compression Ratio	-	7.20× smaller
Accuracy Retention	-	92.18%

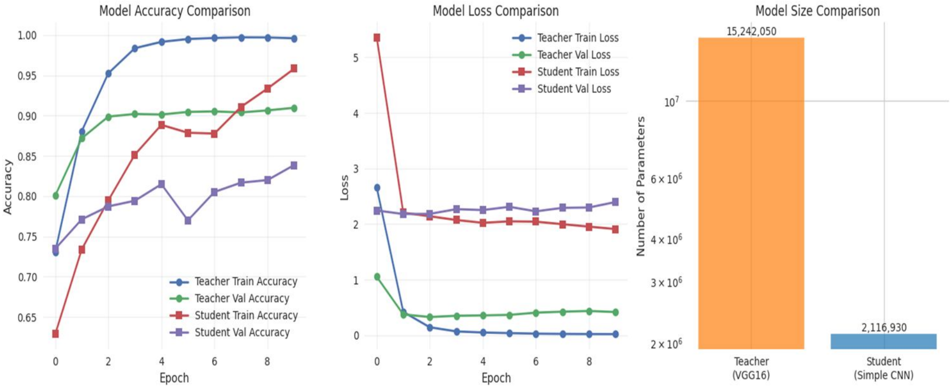


Fig 2: Model Accuray, Loss and Size Comparison

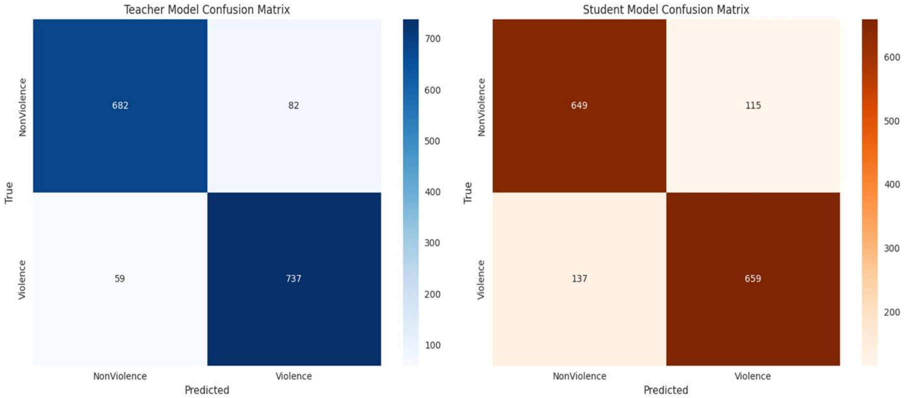


Fig 3: Confusion Matrix of Teacher vs Student Model

The knowledge distillation process was highly successful in creating a compact and efficient model.

1) Model Compression: The teacher model had 15,242,050 parameters, while the student model had only 2,116,930 parameters. This represents a remarkable 7.20x compression ratio, making the student model significantly more light-weight and suitable for deployment on edge devices.
2) Accuracy Retention: Despite the substantial reduction in size, the student model successfully retained a high percentage of the teacher's accuracy. The accuracy retention was calculated to be 92.18%, which is a key indicator of the effectiveness of the knowledge distillation method. This shows that the student model learned not just the final labels, but also the complex decision-making process of the more complex teacher model.

V. CONCLUSION

This research successfully demonstrates the application of a knowledge distillation framework to develop an efficient, lightweight, and accurate real-time violence detection system. The results confirm that the teacher model, while larger and more complex, achieved a high level of accuracy at 90.96%. However, the core finding is the success of the knowledge distillation process. The student model, with a notable 7.20x reduction in the number of parameters, achieved an accuracy of 83.85%, retaining over 92% of the performance of the teacher model. In conclusion, this study shows knowledge distillation as an effective and efficient technique for model compression, offering a practical solution to build real-time surveillance systems.

VI. FUTURE WORK

Based on the research findings and limitations, the following areas are proposed for future work:

1) Identify alternative distillation methods beyond the standard hard and soft loss combination. This could include using attention-based distillation, feature-map distillation, or a different balance of loss weights to potentially improve the accuracy of the student model.

2) The research could be extended by training and evaluating the models diverse dataset with larger in size. This would help to improve the model's generalization to various real-world scenarios.

3) While the current methodology focuses on frame-based analysis, future work could incorporate spatio-temporal models like 3D CNNs or a CNN-LSTM hybrid. This would allow the model to capture motion and temporal dependencies more effectively, which is crucial for accurate violence detection.

4) Conduct further research on optimizing the student model for deployment on specific edge devices, such as the Jetson Nano or Raspberry Pi. This would involve profiling the model for latency and memory usage to ensure it can operate efficiently in real-time.

REFERENCES

[1] K. R. Krishna, V. S. S. Vishak, and V. C. V. Vyshnavi, “Real time violence detection,” Journal of Science and Technology, vol. 9, no. 4, pp. 1–5, Apr. 2024.

[2] H. A. H. Baca, F. d. L. P. Valdivia, and J. C. G. Caceres, “Efficient human violence recognition for surveillance in real time,” Sensors, vol. 24, no. 2, p. 668, Jan. 2024, doi: 10.3390/s24020668.

[3] A. A. S. A. Arun, S. S. M. R. Sri Skandha, K. Esha, and N. Nathiya, “Human violence detection using deep learning techniques,” in Journal of Physics: Conference Series, vol. 2318, no. 1, p. 012003. IOP Publishing, 2022, doi: 10.1088/1742-6596/2318/1/012003.

[4] A. Verma, “Real-Time Violence Detection in Surveillance Streams,” SSRN, New Delhi, India, 2024.

[5] P. Negre et al., “Literature review of deep-learning-based detection of violence in video,” Sensors, vol. 24, no. 12, p. 4016, Jun. 2024, doi: 10.3390/s24124016.

[6] G. Sudeepthi, R. V. A. Reddy, T. Vaishanvi, and C. Swapna, “Smart surveillance for violence detection,” International Journal for Multidisciplinary Research (IJFMR), vol. 6, no. 6, Nov.-Dec. 2024.

[7] E. Veltmeijer, M. Franken, and C. Gerritsen, “Real-time violence detection and localization through subgroup analysis,” Multimedia Tools and Applications, vol. 84, pp. 3793–3807, May 2024, doi: 10.1007/s11042-024-19144-5.

[8] H. Khan et al., “Violence detection from industrial surveillance videos using deep learning,” IEEE Access, vol. 13, pp. 15363-15375, 2025, doi: 10.1109/ACCESS.2025.3531213.

[9] S. A. Sumon et al., “Violence detection by pretrained modules with different deep learning approaches,” Vietnam Journal of Computer Science, vol. 7, no. 1, pp. 19-40, 2020, doi: 10.1142/S2196888820500013.

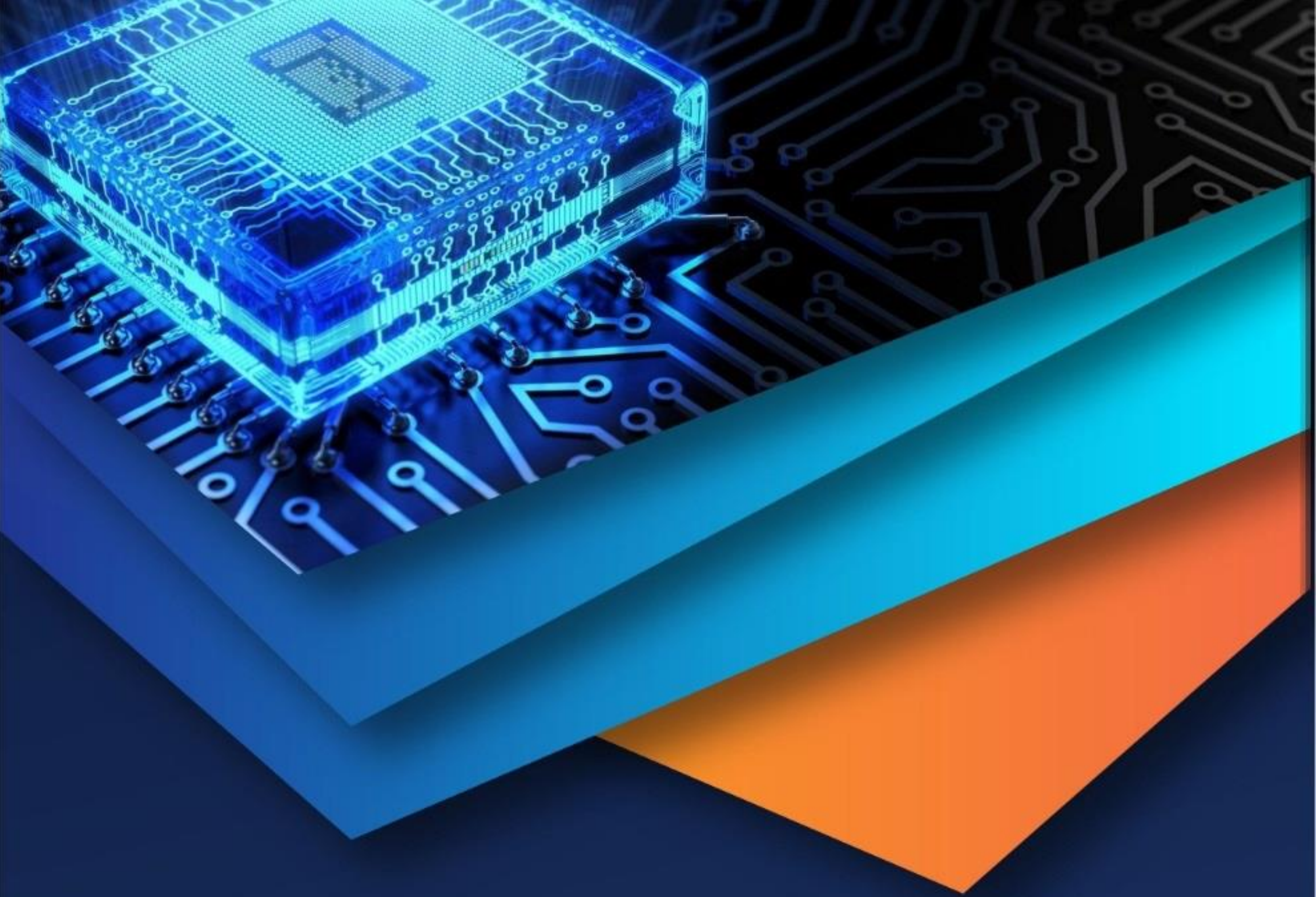
[10] M. R. Khan et al., “Multimodal deep learning for violence detection: VGGish and MobileViT integration with knowledge distillation on Jetson Nano,” IEEE Open Journal of the Communications Society, vol. 6, 2025, doi: 10.1109/OJCOMS.2024.3520703.

[11] R. R. Ojha, H. Chawdary, and S. Saraswat, “Enhancing public safety: Real-time violence detection and notification system,” Procedia Computer Science, vol. 258, pp. 2988–2995, 2025, doi: 10.1016/j.procs.2025.04.558.

[12] M. Ahsan, “Real-time violence detection in smart cities using lightweight spatiotemporal deep learning models,” Journal of Artificial Intelligence and Metaheuristics (JAIM), vol. 9, no. 2, pp. 19-36, 2025, doi: 10.54216/JAIM.090202.

[13] M. A. Soeleman, C. Supriyanto, D. P. Prabowo, and P. N. Andono, "Video violence detection using LSTM and transformer networks through grid search-based hyperparameters optimization," International Journal of Safety and Security Engineering, vol. 12, no. 5, pp. 615–622, Nov. 2022, doi: 10.18280/ijsse.120510.

[14] S. A. A. Akash, et al., "Human violence detection using deep learning techniques," Journal of Physics: Conference Series, vol. 2318, no. 1, p. 012003, 2022, doi: 10.1088/1742-6596/2318/1/012003.



10.22214/IJRASET

INDEX COPERNICUS

45.98

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429