ISSN: 2093-5374 / E-ISSN: 2093-5382

Violence Detection in Videos Using Low Complex Convolution Neural Network for Surveillance Applications

L. Abdul Saleem[1*], and Dr. Gowtham Mamidisetti[2]

[1*]Research Scholar, Department of Computer Science and Engineering, Malla Reddy University, Hyderabad, India. skabdulsaleem@gmail.com, https://orcid.org/0009-0000-5697-219X

[2]Professor, Department of Computer Science and Engineering, Malla Reddy University, Hyderabad, India. drmgowtham@mallareddyuniversity.ac.in, https://orcid.org/0000-0002-9411-613X

Received: August 04, 2024; Revised: September 15, 2024; Accepted: October 11, 2024; Published: December 30, 2024

Abstract

Violent incidents in public and private spaces are a growing concern, necessitating more efficient and reliable surveillance systems. Conventional systems, which rely on manual monitoring, are labour-intensive and prone to error, underscoring the need for automated solutions. The proposed research is to use a low-complexity convolutional neural network CNN model for real-time violence detection in videos, which can be used in surveillance settings. Addressing the limitations of existing high-computation methods, which are often unsuitable for real-time detection in resource-constrained environments, a lightweight CNN model incorporating SeparableConv2D layers has been proposed. This architecture reduces computational complexity by decomposing convolutional operations into depthwise and pointwise convolutions, ensuring efficient feature extraction while maintaining accuracy. The model combines a CNN backbone, MaxPooling layers for down-sampling the dimensions, and dense layer with output layer with sigmoid activation using the Keras-TensorFlow framework. Benchmarking against traditional methods like XGBoost, the model achieved a significant accuracy improvement, with computational efficiency suited for deployment on edge devices. Experimental results show the proposed model achieving 95% accuracy in distinguishing violent from non-violent actions, outperforming conventional methods and proving effective for real-time surveillance applications. This highlights the potential of low-complexity CNN architectures in enhancing public safety through timely intervention in high-risk environments, offering an efficient, accurate solution for violence detection.

Keywords: Violence Detection, Low-Complexity CNN, Surveillance Applications, Convolutional Neural Networks, Keras-TensorFlow, Real-Time Detection, XGBoost.

1 Introduction

The prevalence of violence in public and private spaces has become a growing concern in today's world, with an increase in violent incidents necessitating effective surveillance systems (Khan et al., 2024). Most old-fashioned surveillance systems require humans for monitoring which is time taking and not so accurate especially when you need to analyze massive amount of video footage. It is now more important than ever to have an automated system that can recognize and report violent acts in real time (Ullah et

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 15, number: 4 (December), pp. 358-369. DOI: *10.58346/JOWUA.2024.I4.024*

*Corresponding author: Research Scholar, Department of Computer Science and Engineering, Malla Reddy University, Hyderabad, India.

al., 2022). In this context, video surveillance has emerged as a critical tool for ensuring public safety, particularly in areas with high risks such as urban environments, public transportation, and event venues. However, manual analysis of video feeds is not only labor-intensive but also prone to oversight, leading to delayed responses and missed opportunities for intervention.

In response to these limitations, automated violence detection techniques have been developed that process video streams. Old way of thinking has largely depended on the crafted features and rule-based systems which provide limited extent of generalization capacity in various unique environments, camera angles or kinds of violent activities. These conventional methods often struggle with the complexity of real-world situations, leading to high rates of false positives and missed detections. Violence can manifested in various forms such as Physical fights or non-physical aggression so the systems that is capable of detecting violence need to operate under varied conditions (Wajid et al., 2022).

Despite the advancements in surveillance technologies, several challenges still persist in violence detection (Huszar et al., 2023). The complexity of real-time video analysis, combined with the variations in lighting, occlusion, and motion dynamics, makes it difficult to develop a one-size-fits-all solution. In addition, the need for low-complexity solutions becomes evident in the context of deploying such systems on edge devices with limited computational resources. Existing high-accuracy solutions have high computational complexity, and are impractical for deployment in resource-constrained real-time applications (Ando et al., 2012). It must be good enough, but not too good on taking much time to compute the answer step and withstanding so far of real world in results.

Moreover, this makes it necessary to quickly detect such violent activities but also gives the system an advantage over heavy false alarm load generation (Calero et al., 2022). Both a false positive necessitating unnecessary intervention, and a false negative Clinically significant predictive failure of harm. This emphasizes the need to create models that are not just fast enough but also very reliable to classify violent and non-violent activities (Braddock, 2024). Consequently, the abundance of existing violence detection systems does not allow at least all challenges to be addressed effectively while also being suited for practical deployment on surveillance systems (Badidi et al., 2023).

These challenges can be solved using deep learning where the spatial and temporal features from video data can be learned automatically by Convolutional Neural Networks (CNN). Due to very high strength of CNNs on image and video classification related tasks they suggest a good solution for fairly difficult violence detection problem (Mumtaz et al., 2023). Low-complexity CNN architectures enable the development of models with near real-time detection performance but require lower computational resources to use. These types of models can be used in edge devices, making them a potentially optimal choice for maintaining rapid, reliable and efficient violence detection in videos for surveillance applications (Zigui et al., 2024).

2 Literature Survey

Sahay et al., (2022), Introduced A new architecture was presented to detect violence in real time for surveillance systems through deep learning-based networks. Their method involves classification of the videos recorded on public surveillance cameras using spatio-temporal (ST) feature extraction and a Deep Reinforcement Neural Network (DRNN). The approach takes videos as input and extracts features that can be used for classification frame by frame. The system trained and tested on the UCF Crime anomaly dataset, with performance evaluation of high accuracy (98), precision (96), recall (80) and F-1 score (78) and efficient real time detection.

Vijeikis et al., (2022) Presented an innovative architecture for violence detection in video surveillance. This model has a U-Net like structure for extracting spatial features and uses the MobileNet V2 network as the encoder, followed by an LSTM to obtain temporal features and classify them. The model is also computationally efficient but performs well. On a challenging real-world security camera dataset, namely the RWF-2000 dataset, it demonstrated an average accuracy of 0.82 ±2% and an average precision of 0.81 ±3% as highlighted in our experiments.

Srivastava et al., (2022) In their survey, it explicitly reviewed on the deep learning based violence detection methods including transfer learning with hybrid models as well to applied LSTM for this purpose. The need of the hour is helping in identifying individuals from drone captured images which can vary by nature and human facial appearance also varies that is addressed using a CNN along with processing techniques. They created a specialized dataset of videos taken by drones in open fields to test. The hybrid model, which integrates inception modules, residual blocks and LSTM architecture was found to be effective performed better with a high accuracy of 97.33% than other model. The CNN model trained with residual blocks for face recognition also gives the best result of 99.20% accuracy in identifying the person from dataset.

Bakhshi et al., (2023) Presented two different deep learning approaches for speech data classification were to differentiate between violence and non-violence behaviors. The first method handles common deep neural networks and the second one processes light deep neural networks. In both cases, fine-tuned models are applied to Mel-spectrogram images of speech signals as input. Our lightweight model with the highest performance yielded an 8% higher classification accuracy than the previous state-of-the-art result for that benchmark dataset. The lightweight models have comparatively lesser number of parameters and take lesser computation so, it is a vital advantage when deployed on mobile/edge device.

Naidu et al., (2024) Proposed a 3D convolutional ResNet18 model for feature extraction using better spatiotemporal pattern recognition to increase accuracy. A ResNet18 3D model is constructed and serves as the backbone for feature extraction augmented with dropout and activation preprocessing layers to enhance non-linear representation of features and prevent overfitting. We follow that up with a fully connected layer to reduce the feature space, then finish off with a classification binary output layer. The model is effective for intricate patterns identification of videos data, which helps to detect the action regarding violence efficiently.

Ehsan et al., (2024) Proposed to identify behaviors while overcoming the challenge of limited violence-related data, an unsupervised Spatial Temporal Action Translation (STAT) network. The framework consists of a person detector, motion feature extractor, STAT network and output interpretation which efficiently filters out distraction background features. Recognizing that violent actions involve rapid movements, temporal features are essential and are fed into the STAT network. Trained on normal behavior, the STAT network translates typical motion into spatial frames but struggles to accurately reconstruct violent actions due to their complexity. Consequently, violent actions are identified by measuring reconstruction errors between actual and reconstructed frames in the output phase.

Haque et al., (2024) Explained about BrutNet, a model for detecting and classifying violent videos. Combining a Deep Convolutional Neural Network (DCNN) with Gated Recurrent Unit (GRU), this approach also outperforms existing methods. Temporal features are learned by a GRU layer and a 1D vector is generated for classification the violent content as binary violent/non-violent. BrutNet was run on an NVIDIA Tesla K80 GPU in Google Colab and attained test accuracies of 97.62%, 100%, 97.22% and 86.43% for the hockey fights and movie fights datasets, respectively BrutNet exhibits promise in

civic applications, requiring just 3.416 million parameters for public safety, content moderation, investigations and law enforcement.

Honarjoo et al., (2024) introduced a compressed domain technique where residual information from partial decoding is used as spatiotemporal features, offering a faster alternative to traditional raw-domain methods. By leveraging residual data, the time required for feature extraction, a critical factor in real-time applications, is significantly reduced. The authors also proposed a new method for clustering similar adjacent residuals, which dramatically reduces the number of frames to travel, in order to detect. Experimental comparison with the recent state-of-the-art algorithms shows that our approach is very efficient

3 Proposed Method

The proposed model is a CNN to classify images especially suited for binary-class problems. It uses the TensorFlow Keras Sequential API to build the model as a linear stack of layers. By passing external images of the shape 256x256 and RGB containing three channels through multiple convolutional, pooling, and dense layers, the model classifies them into one-out-of-two categories (Gyamfi et al., 2022). Convolutional layers extract useful features from the images, like edges, textures, and shapes; pooling layers reduce the spatial dimensions of the feature maps, making it computationally efficient. The Separable Convolution layers optimize the model further, breaking down standard convolutions into physical operations that are cheaper to compute but accurate. Finally, the last dense layers bring together the gathered information to classification. The final layer uses a sigmoid activation function since it is a binary classification problem. Here is the full breakdown of each of the layer in the model. Proposed Low Complex CNN Model shown in Figure 1.

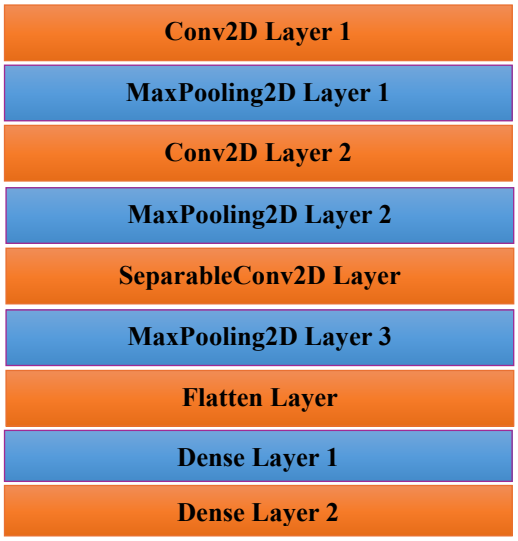

Figure 1: Proposed Low Complex CNN Model

- **Conv 2D Layer 1**

  The first convolutional layer which uses 16 filters with the size of 3 ×3 applied to the original image. The stride is one, which means that the filter slides across the image by moving one pixel at a time. The Rectified Linear Unit (ReLU) activation function is used to introduce non-linearity, ensuring the network can learn complex patterns. The output of this layer will be a feature map with which the model has started learning low level features like edges.

- **Max Pooling 2D Layer 1**

  The first layer is a MaxPooling2D Layer which applied right after the first convolution. This pooling operation down-samples the dimensions of the feature maps, usually by obtaining maximum values of 2x2 blocks of input feature map. By discarding less important spatial details, it discards some of less important information which can reduce computational load and chances of overfitting and also helps to down sample the data.

- **Conv 2D Layer 2**

  It has 32 filters, so the second convolutional layer extracts more complex features and learns advanced properties of the image (David Winster Praveenraj et al., 2024). This is performed using a 3x3 kernel, and like the previous convolution, at a stride of 1. In turn the ReLU activation still applies non-linearity. With the application of more filters, a network subsequently learns a representation of an image with finer detail, identifying complex structures.

- **Max Pooling 2D Layer 2**

  Similar to the previous pooling layer, MaxPooling is used again to down sample feature maps by reducing length and width size which reduces resolution and will help emphasize the most notable features. The purpose of this layer is to maintain the model which should not be sensitive to the changes in input, like shift or distortion.

- **Separable Conv 2D Layer**

  Separable Convolution is a less costly variant of traditional convolution. It breaks up the convolution into two parts; a depth wise and pointwise convolution, where the depth wise applies a single filter to each input channel) and the pointwise is just 1 x 1 convolution which mixes the outputs of the depth wise. This minimizes the number of computations but still helps the network learn crucial features. We also apply ReLU activation here.

- **Max Pooling 2D Layer 3**

  After SeparableConv2D, another MaxPooling2D is applied to further downsize the spatial dimension of the output. This ensures the model continues to focus on the most critical features while maintaining computational efficiency.

- **Flatten Layer**

  The Flatten layer transforms the 2D feature maps into a 1D vector. This step is necessary because the next layers are fully connected dense layers, which require a 1D input. Flattening takes the spatial structure learned by the convolutional layers and translates it into a format that can be processed by the dense layers for classification.

- **Dense Layer 1**

  This fully connected layer has 256 neurons, each receiving input from all the features of the previous layer. Having ReLU in an activation function helps the network to capture and model the non-linear relationship between features and output. This layer acts as a high-level feature integrator, combining all the learned features from the convolutional layers to make predictions.

- **Dense Layer 2**

  The final layer is a single neuron that outputs a value between 0 and 1, using the Sigmoid activation function. This value represents the probability of the input image belonging to one of

the two categories (binary classification). A threshold (typically 0.5) can be used to decide whether the image belongs to class 0 or class 1.

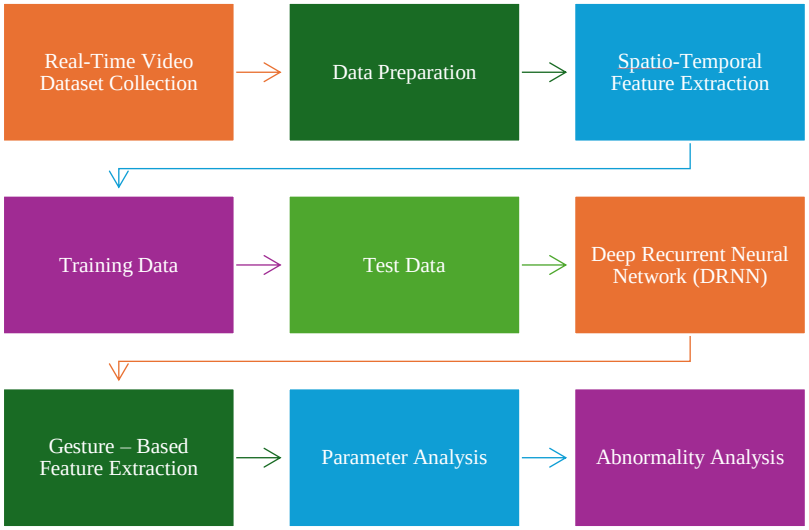

Figure 2: Architecture of Violence/ Non- Violence Detection

Figure 2 presents the architecture of a proposed model for detecting violence and non-violence in real-time video feeds.

- **Real-Time Video Dataset Collection**

  Real-time video is recorded in order to create a dataset of violent and non-violent actions. The first step is collecting this dataset, which will be used to train and test the model.

- **Data Preparation**

  Collected video data is pre-processed, including steps like resizing, normalization, and segmentation, to ensure consistency and suitability for feature extraction and model training.

- **Spatio-Temporal Feature Extraction**

  The video frames are extracted through the two dimensions of spatio-temporal features. Spatial features are appearance descriptors and temporal features are motion descriptors, both of which are critical to violent behaviour recognition.

- **Gesture-Based Feature Extraction**

  This module focuses on extracting body language and gesture-related features that are commonly associated with violent actions, adding another layer of specificity to the feature set.

- **Deep Recurrent Neural Network (DRNN)**

  A The sequential data of video frames is added to input the DRNN, where the model can capture complex dependencies and temporal patterns indicating whether or not a piece of action is violent.

- **Data Splitting**

  The data, which has been extracted and prepared, is split into two parts Training Data & Test Data. The its DRNN gets trained on the training data and tested the input data to be predicted by utilizing the test Data.

- **Parameter Analysis**

  After training, the model's parameters are analyzed to optimize performance. This step helps in refining the model to improve its accuracy and efficiency.

- **Abnormality Analysis**

  This module identifies and analyzes unusual or abnormal behaviors that may indicate potential violent actions. It measures deviations from normal patterns to enhance detection accuracy.

The proposed model achieves low complexity by integrating Separable Convolutional layers and using fewer filters in the convolutional layers, particularly in the early stages. Standard convolution operations, which are computationally intensive due to dense weight matrices, are replaced by Separable Convolutions. This method essentially separates the convolution into depth wise and pointwise convolutions, drastically reducing the number of parameters while still allowing for learning of important features. The model achieves a compromise between efficient processing and accurate output by limiting the number of filters while remaining lightweight for on-edge devices or real-time systems which are often resource-constrained in surveillance applications.

Although it was a simplified version, this model is able to detect violence in videos leveraging spatial and temporal features. Convolutional layers can fall purely on spatial feature extraction, allowing for learned patterns to recognize more motion-oriented features as well as object boundaries and textures that lead into violent action classes. MaxPooling layers serve the purpose of down-sampling the feature maps, thus also containing important information. At the end of the network, dense layers turn learned features to make precise binary predictions about whether an activity is violent or not. The final sigmoid activation function gives the probability of violence between 0–1 that can trigger alerts instantly. This is followed by a series of discrete processes that facilitate both high quality detection of violence in each frame and efficient learning, which when together support fast classification times in real-world applications.

4 Experimental Results

This section is dedicated to the thorough justification of the results achieved with application of proposed framework in current simulations. For these simulations the dataset used is a public Kaggle dataset (Mustafa, 2019). This study is guided by the general approaches for the processing of datasets used in this kind of research.

Violence Detection

The dataset used in this work is created to fill the gap for publicly available datasets focusing on interpersonal violence mainly aimed at video-based violence recognition. These action videos are violent and non-violent which can be categorized into collection 1000, collected from one youtube channel. The variety of street fights for different conditions: room, park, and others types of environments added the diversity in a dataset. On the other hand, non-violent videos include several activities such as sports, eating, walking etc. to make it a balanced dataset for training deep learning models. The dataset serves as an important resource and ensemble for research on violence detection from video data, enabling engineering and scientific-focused developments in the area.

(a) Violence (b) Non-violence

Figure 3: Sample Images in the Dataset

Figure 3: Example of violence and non-violence detection samples from the dataset. Image (a) titled "Violence" shows outdoors violence between two men with all the motion and aggression found in common violent acts. An image of this kind is useful to train models in learning concepts related to violence e.g. speed, physical confrontation. The second image (b), labeled as "Non-violence", depicts two people seemingly calmly talking indoors, which plays against the visual cues of violence. Moreover, the dataset covers a wide range of setting, activity, and behaviour Figure 3, as shown in these samples ensuring that the model can identify violence action/non-violence actions in different contexts. Classification Report shown in Table 1.

Table 1: Classification Report

Index	precision	recall	f1-score	support
Non-violence	0.96	0.95	0.95	170
Violence	0.95	0.96	0.95	170
accuracy			0.95	340
macro avg	0.95	0.95	0.95	340
weighted avg	0.95	0.95	0.95	340

The classification report shows the model performance for every single class "Non-violence" and "Violence" in terms of precision, recall and f1-score. Precision the fraction of true positives over all positive predictions made by the model for a specific class. For "Non-violence class" the precision score is 0.96 which shows that while predicting Non-violent instances model is making very few false positives. Similarly, the precision score for the "Violence" class is 0.95, indicating the model is almost as effective in minimizing false positives for this class as well.

Recall also known as sensitivity or true positive rate. It is a measure of how well the model can find all instances of the given class. Here recall for Non-violence is 0.95 which means model was able to identify 95% of actual non-violence items correctly. The recall of 0.96 for 'Violence' class implies the model is good at identifying violent instance and has failed to miss only a small percent of them. The different balance of precision and recall for both classes indicate a well-calibrated model for the specific task.

F1 score is a harmonic mean of precision and recall which gives an overall view on how the model is performing taking as a whole say false positives or negative rates. The model has shown high

consistency amongst both the classes Non-violence and Violence, having f1-score of 0.95 for each of them. This means that out of 340 instances, the model's overall prediction accuracy was 95%. The macro average and weighted average for precision, recall, and f1-score is 0.95 which shows that the model gives equal performance to classes and it does not bias one class over another. These metrics demonstrate the model's robustness and its potential to be a reliable tool for tasks involving violence detection. Confusion Matrix shown in Figure 4.

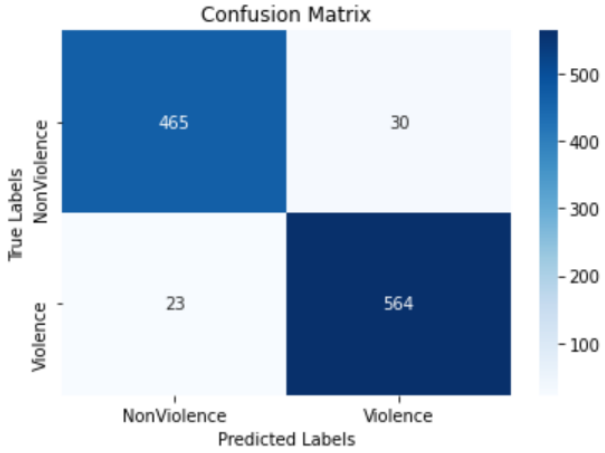

Figure 4: Confusion Matrix

The confusion matrix shown above represents the performance of a binary classifier predicting between "Non-Violence" and "Violence". Here, 465 "Non-Violence" was correctly classified as non-violent true negatives, and 30 instances was incorrectly classified as violence false positives. Conversely, in the "Violence" category, we correctly recognised 564 instances as violent true positives but also misclassified 23 instances as non-violent false negatives. The diagonal components are large while the off-diagonal are small in this matrix therefore shows a good ability of the model to identify violent and non-violent instances separately. It is a reasonably good model with more true positives and true negatives as opposed to misclassifications. However, 23 false negatives suggest that some violent events are going unnoticed, which could be a concern in real-time or high-risk applications such as security or surveillance. Conversely, the 30 false positives indicate that a small portion of non-violent instances is incorrectly flagged as violent, which may result in unnecessary alerts. Despite these misclassifications, the model demonstrates a good balance between precision and recall, effectively handling both categories while still leaving room for further tuning to minimize the false negative and false positive rates.

Table 2: Comparative Analysis

Methods	Accuracy
XGBoost (Negre et al., 2024)	85%
Mobile Net (Khan et al., 2024)	89%
ResNet-v2 (Nayak et al., 2024)	90%
Proposed Model	95%

Table 2 presents a comparative analysis of various methods for violence detection, focusing on their accuracy in classifying violent and non-violent actions. The methods reviewed are popular models, including an XGBoost model, a MobileNet model, and a ResNet-v2 model along with the present model. The 85% accuracy is achieved with the widely known gradient-boosting algorithm XGBoost. Although

XGBoost is very effective in most of the machine learning tasks it may not be able to capture complex spatial and temporal interactions from video data, which are important for detecting violence. MobileNet is more of a lightweight convolutional neural network (CNN) that achieves 89% accuracy which implies better accuracy edge but enhances it for mobile and edge devices so, exemplifying efficiency with effectiveness.

Another model, ResNet-v2 which is even deeper CNN with well-known concept of residual connections reaches the accuracy of 90%. Due to its architecture, it is able to extract more detailed features which makes it a better at identifying violent actions compared with other simple models like XGBoost and MobileNet. Yet still, ResNet-v2 is outperformed by the proposed model which achieves a state of 95%. The boost in accuracy is a strong indication of the efficiency of the proposed model architecture design. Then particularly with respect to utilizing Separable Convolutional layers and reducing filter number for a memory efficient computation without loss in accuracy. The accurate 95% of the proposed model demonstrates its suitability for violence detection in real time, specifically for resource constrained settings like embedded surveillance systems running on edge devices. This model has a potential as a simple solution to decrease violence while still effective since it retains generalisability over the traditional models, per-sequence computational cost and in addition to its robustness between sequences by determining whether there is presence of violent behaviour or otherwise. By means of the relative results, we emphasize on architectural variations which partially avoid the well-studied between high efficiency and effectiveness. Augment to world-wide effort in violence detection with a novel applicable solution for deployment that is both utilize without cost and guarantee scalability effectively at low computational expense whilst maintaining very high accuracy.

5 Conclusion

The final research discussed in this review developed a low-complexity model able to detect violence in real time, overcoming the difficulties of traditional high-compution approaches that cannot be applied in environments with limited resources. This model uses Separable Convolutional layers and a smaller filters number to be computationally effective by reducing the amount of operations needed, whilst maintaining accuracy. This methodology enables the model to effectively extract spatio-temporal features relevant for classifying violent and non-violent events, despite being resource limited. The model has proven effective in experiments where it reached an accuracy of 95%, it has surpassed traditional violence detection models like XGBoost and Mobile Net. The above-mentioned findings hint that the model is robust enough for edge deployment to deliver real-time surveillance solutions with a contribution towards public safety. Hence, this model makes use of feature extraction dimensionality reduction and binary classification unit in a single framework. It is a viable step toward an efficient violence detection model for surveillance areas. The study results also reinforce the idea that low-complexity neural networks provide a higher level of performance and efficiency for securing environments with high attack risk.

References

[1] Ando, R., Takahashi, K., & Suzaki, K. (2012). Inter-domain Communication Protocol for Real-time File Access Monitor of Virtual Machine. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 3*(1/2), 120-137. https://doi.org/10.22667/JOWUA.2012.03.31.120

[2] Badidi, E., Moumane, K., & El Ghazi, F. (2023). Opportunities, applications, and challenges of edge-AI enabled video analytics in smart cities: a systematic review. *IEEE Access*, *11,* 80543-80572. https://doi.org/10.1109/ACCESS.2023.3300658

[3] Bakhshi, A., García-Gómez, J., Gil-Pita, R., & Chalup, S. (2023). Violence detection in real-life audio signals using lightweight deep neural networks. *Procedia Computer Science*, *222*, 244-251. https://doi.org/10.1016/j.procs.2023.08.162

[4] Braddock, K. (2024). Using Deep Learning Neural Networks to Predict Violent vs. Nonviolent Extremist Behaviors. *Terrorism and Political Violence*, 1-23. https://doi.org/10.1080/09546553.2024.2376639

[5] Calero, J. A. M., Rituerto-González, E., Luis-Mingueza, C., Canabal, M. F., Bárcenas, A. R., Lanza-Gutiérrez, J. M., ... & López-Ongil, C. (2022). Bindi: affective internet of things to combat gender-based violence. *IEEE Internet of Things Journal*, *9*(21), 21174-21193. https://doi.org/10.1109/JIOT.2022.3177256

[6] David Winster Praveenraj, D., Prabha, T., Kalyan Ram, M., Muthusundari, S., & Madeswaran, A. (2024). Management and Sales Forecasting of an E-commerce Information System Using Data Mining and Convolutional Neural Networks. *Indian Journal of Information Sources and Services, 14*(2), 139–145. https://doi.org/10.51983/ijiss-2024.14.2.20

[7] Ehsan, T. Z., Nahvi, M., & Mohtavipour, S. M. (2024). An accurate violence detection framework using unsupervised spatial–temporal action translation network. *The Visual Computer*, *40*(3), 1515-1535. https://doi.org/10.1007/s00371-023-02865-3

[8] Gyamfi, N. K., Goranin, N., Čeponis, D., & Čenys, H. A. (2022). Malware detection using convolutional neural network, a deep learning framework: comparative analysis. *Journal of internet services and information security, 12*(4), 102-115. https://doi.org/10.58346/JISIS.2022.I4.007

[9] Haque, M., Nyeem, H., & Afsha, S. (2024). BrutNet: A novel approach for violence detection and classification using DCNN with GRU. *The Journal of Engineering*, *2024*(4), e12375. https://doi.org/10.1049/tje2.12375

[10] Honarjoo, N., Abdari, A., & Mansouri, A. (2024). Violence detection in compressed video. *Multimedia Tools and Applications*, 1-14. https://doi.org/10.1007/s11042-024-19478-0

[11] Huszar, V. D., Adhikarla, V. K., Négyesi, I., & Krasznay, C. (2023). Toward fast and accurate violence detection for automated video surveillance applications. *IEEE Access*, *11*, 18772-18793. https://doi.org/10.1109/ACCESS.2023.3245521

[12] Khan, M., El Saddik, A., Gueaieb, W., De Masi, G., & Karray, F. (2024). VD-Net: An Edge Vision-Based Surveillance System for Violence Detection. *IEEE Access*, *12*, 43796-43808. https://doi.org/10.1109/ACCESS.2024.3380192

[13] Khan, N. F., Amin, S. U., Jan, Z., & Yan, C. (2024). The Detection of Violent Scenes in Cartoon Movies Using Deep Learning Approach. *IEEE Access, 12,* 154080-154091. https://doi.org/10.1109/ACCESS.2024.3480205

[14] Mumtaz, N., Ejaz, N., Habib, S., Mohsin, S. M., Tiwari, P., Band, S. S., & Kumar, N. (2023). An overview of violence detection techniques: current challenges and future directions. *Artificial intelligence review*, *56*(5), 4641-4666. https://doi.org/10.1007/s10462-022-10285-3

[15] Mustafa, M. (2019). *Real Life Violence Situations Dataset*. https://www.kaggle.com/datasets/mohamedmustafa/real-life-violence-situations-dataset/data

[16] Naidu, T. P., Sekhar, P. C., & Boya, P. K. (2024). Violent Human Behaviour Detection in Videos Using ResNet18 3D Deep Learning. *SN Computer Science*, *5*(7), 874. https://doi.org/10.1007/s42979-024-03243-z

[17] Nayak, R., Pati, U. C., Das, S. K., & Sahoo, G. K. (2024). YOLO-gtwdnet: A lightweight YOLOv8 network with ghostnet backbone and transformer neck to detect handheld weapons for smart city applications. *Signal, Image and Video Processing*, *18*(11), 8159-8167. https://doi.org/10.1007/s11760-024-03458-w

[18] Negre, P., Alonso, R. S., González-Briones, A., Prieto, J., & Rodríguez-González, S. (2024). Literature Review of Deep-Learning-Based Detection of Violence in Video. *Sensors*, *24*(12), 4016. https://doi.org/10.3390/s24124016

[19] Sahay, K. B., Balachander, B., Jagadeesh, B., Kumar, G. A., Kumar, R., & Parvathy, L. R. (2022). A real time crime scene intelligent video surveillance system in violence detection framework using deep learning techniques. *Computers and Electrical Engineering*, *103*, 108319. https://doi.org/10.1016/j.compeleceng.2022.108319

[20] Srivastava, A., Badal, T., Saxena, P., Vidyarthi, A., & Singh, R. (2022). UAV surveillance for violence detection and individual identification. *Automated Software Engineering*, *29*(1), 28. https://doi.org/10.1007/s10515-022-00323-3

[21] Ullah, F. U. M., Obaidat, M. S., Muhammad, K., Ullah, A., Baik, S. W., Cuzzolin, F., ... & de Albuquerque, V. H. C. (2022). An intelligent system for complex violence pattern analysis and detection. *International journal of intelligent systems*, *37*(12), 10400-10422. https://doi.org/10.1002/int.22537

[22] Vijeikis, R., Raudonis, V., & Dervinis, G. (2022). Efficient violence detection in surveillance. *Sensors*, *22*(6), 2216. https://doi.org/10.3390/s22062216

[23] Wajid, M. S., Terashima-Marin, H., Paul Rad, P. N., & Wajid, M. A. (2022). Violence detection approach based on cloud data and Neutrosophic cognitive maps. *Journal of Cloud Computing*, *11*(1), 85. https://doi.org/10.1186/s13677-022-00369-4

[24] Zigui, L., Caluyo, F., Hernandez, R., Sarmiento, J., & Rosales, C. A. (2024). Improving Communication Networks to Transfer Data in Real Time for Environmental Monitoring and Data Collection. *Natural and Engineering Sciences, 9*(2), 198-212. https://doi.org/10.28978/nesciences.1569561

Authors Biography

L. Abdul Saleem, is a graduate with M.E in Computer Science Engineering from Arulmigu Meenakshi Amman College of Engineering, Kanchipuram, India. He is currently pursuing his Ph.D. degree in Computer Science engineering at Malla Reddy University. He is currently working as an assistant professor at Computer Science Engineering and AIML department in Malla Reddy College of Engineering and Technology, Hyderabad, India. His research interests are in fields of Computer Vision, Machine Learning.

Dr. Gowtham Mamidisetti, received his Ph.D. in Computer Science and Engineering from Acharya Nagarjuna University, M.Tech in Computer Science from University College of Engineering, JNTU Kakinada and B.Tech in Computer Science and Engineering from JNTU Kakinada. He is GATE Rank holder in GATE 2010. He Published 17 papers in International Journals and presented 4 papers in International Conferences.