ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Efficient Violence Recognition in Video Streams using ResDLCNN-GRU Attention Network

Arnab Dey[1], Samit Biswas[2] and Laith Abualigah[3]

ABSTRACT

Detecting Violence in video streams is essential for public safety and security due to the rising frequency of violent incidents. Despite the extensive deployment of CCTV for surveillance, the available human monitoring resources still need to catch up with the need for vigilant supervision. This research presents a new lightweight model to address this gap by accurately identifying and categorizing violent behaviors in various scenarios, including CCTV footage. The proposed method leverages optical flow and RGB data to capture spatiotemporal features in the Violence data. Built on a Residual DLCNN architecture integrated with the Attention mechanism and GRU components, the model effectively handles high-dimensional video data, enhancing accuracy by prioritizing crucial frames containing violent and nonviolent instances. The proposed model's performance was validated on the Hockey Fights (HF), Movie Fights, and SCVD datasets, achieving impressive accuracies of 98.38%, 99.62%, and 90.57%, respectively. Here, we developed the Extended Automatic Violence Detection Dataset (EAVDD), featuring 1530 videos of violent scenes in movies, public spaces, social media, and sports. Testing the model with top fight scenes in rated movies yielded outstanding results. This research supports surveillance systems and advances short video analysis and understanding with applications in public safety, social media, sports, and law enforcement.

Article information:

Keywords: Violence Recognition, Video, Residual Dilated CNN, Attention, Surveillance

Article history:
Received: February 12, 2024
Revised: April 25, 2024
Accepted: June 27, 2024
Published: July 13, 2024
(Online)

DOI: 10.37936/ecti-cit.2024183.255679

1. INTRODUCTION

The increasing instances of Violence in diverse settings have prompted an urgent need for advanced technologies to facilitate timely detection and prevention. Identifying violent behavior within video content is crucial for public safety and security. The widespread adoption of closed-circuit television (CCTV) for surveillance [1] and crime prevention in today's society has surged to unprecedented levels. Despite this widespread use of CCTV systems, there has not been a proportional increase in human resources for vigilant supervision and oversight to match this growth. Automated Violence detection systems are crucial for rapidly identifying violent incidents in recorded CCTV footage, enabling timely alerts to mitigate potential risks and enhance safety measures. Traditional Violence detection methods, which involve manual feature extraction from video footage, face limitations in adapting to dynamic real-world situations [2] and lack resilience against varying installation angles, backgrounds, environments, and video resolutions. Recent advancements in artificial intelligence have led to developing deep learning models that autonomously identify features and patterns, addressing these limitations.

Our research aims to develop a lightweight, vision-based deep learning model for efficient Violence detection, motivated by the need to address the increasing prevalence of violent incidents in public spaces, institutions, and other environments. Identifying violent occurrences in video content relies on assessing the extent of motion, with optical-flow techniques measuring motion between consecutive frames and

[1,2]The authors are with the Computer Science and Technology, Indian Institute of Engineering Science and Technology, Shibpur, India, E-mail: arnabdeycs@gmail.com and samit@cs.iiests.ac.in
[3]The author is with the Computer Science Dept., Al al-Bayt University, Mafraq, Jordan, E-mail: aligah.2020@gmail.com
[3]The author is with the MEU Research Unit, Middle East University, Amman, Jordan.
[3]The author is with the Applied Science Research Center, Applied Science Private University, Amman, Jordan.
[1]Corresponding author: arnabdeycs@gmail.com

RGB frames capturing static images to reveal intricate scene details. Combining optical flow and RGB data improves accuracy and efficiency by providing a more comprehensive understanding of the video content.

This study presents a robust Violence-detection network designed for diverse environments. The proposed approach isolates and examines moving objects within videos, enhancing detection capabilities across benchmark datasets depicting violent situations. We propose a model that integrates Residual Dilated Convolutional Neural Networks (DLCNN), Attention mechanisms, and GRU, referred to as ResDLCNN-GRU Attention. This model's proficiency in deciphering video data enables comprehensive training to identify Violence-related patterns, encompassing dynamic motions and visual surroundings. The proposed method signifies a notable advancement in developing a precise and reliable Violence-detection model by harnessing the strengths of optical flow and RGB data features. The proposed ResDLCNN-GRU Attention model categorizes videos into two distinct classes: *Violence* and *NonViolence*. Notably, our lightweight model surpasses various existing methodologies by adeptly tackling the complexities of the diverse configurations found in real-world CCTV setups, resulting in improved robustness and adaptability. This study marks a significant step forward in creating a sophisticated and efficient system for Violence detection in video streams. Critical applications of this research include identifying violent actions in sports such as football and hockey, analyzing violent content in short movies or fight scenes, recognizing violent scenes in CCTV footage, and detecting Violence in public spaces and transport, thereby enhancing safety and security for the public.

The significant contributions of this research encompass:

(a) Development of an Extended Automatic Violence Detection Dataset (EAVDD) for Violent Activity Recognition with data labeling. This dataset encompasses violent scenes in public spaces, movies, sports, and other contexts.

(b) Recognition of Violent Scenes using the proposed lightweight ResDLCNN-GRU Attention Network.

(c) Evaluation of Modified Binary Cross-Entropy (MBCE) and Categorical Crossentropy Loss.

(d) Evaluation of the proposed model on three standard benchmark Violence video datasets: Hockey Fights, Movie Fights, and SCVD.

(e) Analyzing results with various standard methods based on diverse metrics.

(f) Testing the proposed model with top fight scenes in rated movies, achieving outstanding results.

This research article is structured as follows: Section 2 provides an in-depth review of related works. Section 3 outlines our proposed method, covering data preparation and the network architecture. Section 4 showcases experimental findings employing standard models on our extended violence detection dataset and other benchmark-related data, evaluating performance using diverse metrics. The concluding remarks encapsulate our distinctive contributions, underscore their significance, and explore future research prospects.

2. RELATED WORKS

Numerous researchers have proposed methodologies for Violence detection using both classical computer vision and deep learning techniques. This discussion emphasizes advanced deep learning approaches, particularly pertinent to the method under consideration. Serrano *et al.* [3] advocated encapsulating a video sequence succinctly in a single image. The essence of their approach lies in the feature extraction process, where the objective is to derive a representative image from every input video footage. Leveraging a 2-dimensional Convolutional Neural Network (CNN), they employed a classification framework on the obtained representative image to derive the ultimate decision for the sequence. Remarkably, their methodology demonstrated exceptional performance, achieving 94.6% accuracy and 99% accuracy on the hockey and movie fight datasets, respectively. In another study, Keceli *et al.* [4] utilized 3D CNN combined with transfer learning for violent activity classification. Here, the deep features are extracted using a pre-trained AlexNet model, then reshaped and concatenated to construct 3D feature volumes for classification. This approach demonstrated 92.90%, 98.7%, and 88% accuracy on the hockey fights, movie, and violent-flow datasets. Dai *et al.* [5] employed two-stream CNN to extract features from the RGB frames and dynamic optical flows, aiming to identify instances of Violence. Long Short-Term Memory (LSTM) was incorporated to capture longer-term temporal dynamics, with SVM used in the classification step. Dündar *et al.* [6] introduced a shallow 3DCNN-based network for Violence or fight detection in videos. Zhang *et al.* [7] proposed a 2DCNN-based approach to detect violent behavior in video. Mahmoodi *et al.* [8] have incorporated 2DCNN with an attention module for video Violence detection. Mohtavipour *et al.* [9] have introduced a multistream CNN-based approach to detect Violence in videos utilizing handcrafted features. Huszar *et al.* [10] have proposed an automated, fast, and accurate Violence detection methodology in videos.

Zhenhua *et al.* [11] have proposed a temporal cross-fusion network to detect Violence in video sequences. Garcia *et al.* [12] have proposed an efficient Violence detection method in videos utilizing human skeletons and change detection. Park *et al.* [13] have introduced a Convolution 3D-based Violence detection method in videos utilizing Optical flow and RGB data. Chaturvedi *et al.* [14] introduced a ConvLSTM model with channel-wise attention to identify instances of fights within video. Their approach demonstrated satisfactory results on the datasets of the RWF-2000, hockey fights, and movie fights. Dey *et al.* [15] proposed an Attention-driven DC-GRU Network to recognize umpire actions in Cricket games. Sudhakaran *et al.* [16] harnessed the power of AlexNet to extract spatial features. They combined it with ConvLSTM for temporal features, yielding a remarkable 97.1% and 100% accuracy on the hockey and movie fights datasets. Li *et al.* [17] employed a keyframe-guided video Swin Transformer to recognize violent activity. Our prior research focused on identifying human interactions in images [18] utilizing the AdaptiveDRNet model. Alabid *et al.* [19] pioneered an approach based on interpreting spatial relationships to track objects within video streams effectively. Mekruksavanich *et al.* [20] have proposed a deep residual model based on multi-branch aggregation for sensor-based fall detection. Rendon et al. [21] proposed a ViolenceNet model based on Bi-LSTM architecture to identify violent situations. Li et al. [22] utilized 3DCNN to detect violent action in another study. In recent work, Tang *et al.* [23] introduced an enhanced, faster R-CNN model designed to identify Violence in animation and cartoon videos automatically. Accattoli *et al.* [24] utilized a 3D CNN and SVM for Violence detection in video streams. Ullah *et al.* [25] used a pre-trained C3D model, attaining a remarkable accuracy of 96% for the Hockey Fights data and an impressive 98% accuracy for the Crowd Violence dataset. Khan *et al.* [26] introduced an edge vision-based surveillance system tailored for detecting violent behavior. Bianculli *et al.* [27] introduced a new video dataset for automated Violence detection. However, it is worth noting that the dataset is relatively limited, consisting of only 230 video clips categorized into Violence and Non-Violence classes. Kaur *et al.* [30] introduced an ensemble transfer learning-based methodology for detecting Violence in video content, achieving satisfactory results on the RWF-2000 dataset.

Prominent video datasets like ViF [28], Hockey Fights [29], Movie Fights [29], and SCVD [31] have played a crucial role in advancing Violence behavior recognition. Nevertheless, these publicly accessible datasets contain limited video samples of both the Violence and NonViolence categories and lack diversity in Violence scenarios. To enrich the available resources for researchers and practitioners, we have introduced the EAVDD data, which comprises 1530 video clips collecting samples from various movies and TV episodes, including some samples of the Hockey Fights [29] and SCVD [31] datasets. This dataset aims to enhance violence recognition models' generalization potential and practical utility.

3. PROPOSED METHOD

The method began by collecting EAVDD videos in various scenarios. We applied data augmentation methods to enrich the training data. Next, we extracted frames from the videos, resized them to fixed dimensions, and normalized their pixel values to [0, 1] to optimize them for subsequent model training. Then, we divided the curated dataset into training and validation sets. Fig. 1 illustrates a visual representation of the proposed methodology.

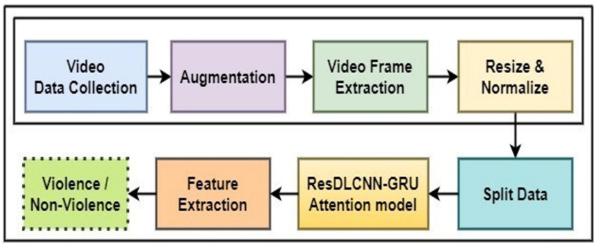

Fig.1: *Representation of the Proposed Methodology.*

Considering the demand for methods capable of operating efficiently within constrained computational resources, this research puts forward the lightweight Residual DLCNN-GRU Attention network to recognize Violence and NonViolence in video streams. The following sections offer comprehensive insights into the subsequent aspects: a) Preparation of the training dataset, b) *ResDLCNN-GRU Attention*: Architecture, and c) Discussion of the utilized loss function.

3.1 Dataset Preparation

The meticulously curated EAVDD dataset comprises top fight scenes from movies, sports events, and real-life videos portraying both violent and non-violent scenarios, providing a comprehensive resource for training and evaluating violence detection models. The collected videos are in RGB. The dataset has two categories. Understanding the critical significance of abundant training data in deep neural networks, we have incorporated the data augmentation technique to increase the overall efficacy of the Residual DLCNN-GRU Attention net. Furthermore, we systematically split the dataset into two subsets, designating 83% of the data for training and 17% for validation. Fig. 2 depicts samples of the collected sports violence data in our EAVDD dataset.

3.2 ResDLCNN-GRU Attention: Architecture

The proposed model architecture integrates a Residual Dilated convolutional neural network (ResDLCNN), Attention mechanism, and Gated Recurrent Units (GRU) for effective feature extraction and temporal sequence modeling. The input data consists of sequences of images with dimensions (SE-

Fig.2: *Sports Violence sample.*

QLENGTH, FRAME_HEIGHT, FRAME_WIDTH, 3) which is $16 \times 64 \times 64 \times 3$. The layers of the proposed model are as follows:

Dilated Convolution Layers: The dilated convolutional layers aim to capture spatial features across the image sequences. In DLCNN, Eqn. 1 mathematically expresses the dilated convolution operation.

$$y_{i,j,c_{out}} = \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{c_{in}=1}^{c_{in}} x_{(i+r.(k-1)),(j+r.(l-1)),c_{in}} \cdot w_{k,l,c_{in},c_{out}} \tag{1}$$

Let us denote the input signal as x with spatial dimensions $H_{in} \times W_{in}$ and C_{in} channels. Here, w denotes the filter (or kernel) with dimensions $K \times K$ and c_{in} channels, r is the dilation rate, and y represents the output of the dilated convolution with dimensions $H_{out} \times W_{out}$ and c_{out} channels. The indices i and j iterate over the spatial dimensions of the output, c_{in} iterates over input channels, and c_{out} iterates over output channels. Fig. 3 depicts the architecture of the proposed ResDLCNN-GRU Attention model.

Initial DLCNN Layers: The model initiates with two Time Distributed (T.D) Dilated CNN (DLCNN) layers with 32 and 64 filters, each using a 3×3 kernel, accompanied by T.D Max Pooling with a (2, 2) pool size. Then, a T.D Dropout layer with a 0.13 dropout rate is incorporated. DLCNN employs a dilation rate; the dilation introduces gaps between the values sampled by the filter, effectively increasing the receptive field without escalating the parameter count. DLCNN helps capture more prominent context information in the input signal. Here, the DLCNN layers utilize dilation rate 1, extracting hierarchical features from the input frames.

The Rectified Linear Unit (ReLU) activation, $f(i) = \max(0, i)$, is applied after each T.D Dilated convolutional operation, introducing non-linearity to the model. ReLU activation promotes the model's ability to learn complex representations. Max pooling, expressed as $y[m, n] = \max_{(i,j)} x[m \cdot s + i, n \cdot s + j]$, follows each convolutional layer, downsampling the spatial dimensions of the feature maps. Here, $y[m, n]$ represents the output of the max pooling operation at spatial location (m, n), s is the size of the pooling window, and $\max_{(i,j)}$ denotes taking the maximum value over a $s \times s$ region, typically called the pooling window or kernel. The model utilizes max pooling with (2, 2) pool size, aiding in retaining essential information while reducing computational complexity.

Residual Block: The residual block comprises two consecutive T.D Dilated 2D convolutional layers having 64 filters and a 3×3 kernel, each with rectified linear unit (ReLU) activation and a specified dilation rate of 1 and 2, respectively. The dilation rate controls the spacing between the kernel elements and helps capture spatial dependencies at different scales. Following each convolutional layer, a dropout layer is applied with a rate of 0.13, introducing a form of regularization to prevent overfitting during training.

The core concept of a residual block lies in incorporating a residual connection. The variable 'res' stores the original input before the convolutional operations. After the convolutional layers and dropout, the output is combined with the original input using the Add() layer. This residual connection facilitates the flow of information directly from the input to the output, mitigating the vanishing gradient issue and enabling the model to learn more efficiently, especially in deep networks. The model can thus focus on learning residual information—what needs to be added or adjusted—rather than relearning the entire representation from scratch. Finally, a T.D 2D max-pooling layer with a (2, 2) pool size is applied to reduce spatial dimensions, contributing to the hierarchical feature extraction process. With its convolutional operations and skip connections, the residual block significantly improves the training dynamics and performance of deep neural networks on complex tasks, particularly when handling sequential or spatiotemporal data.

Next, we incorporate another T.D Dilated convolution layer with 128 filters and a 3×3 kernel, followed by ReLU activation after the Residual block. T.D Max Pooling follows this with a pool size of (2, 2) and T.D Dropout. We then process the output of this layer through a T.D Flatten layer before integrating the Attention mechanism.

Attention Mechanism: The Attention mechanism is integrated into the proposed model using the flattened output, allowing the model to focus on specific segments of the input sequence during prediction. In this context, the input to the Attention mechanism results from the flattened operation on the Residual DLCNN features. Sequence-to-sequence models mainly utilize the dot-product attention mechanism. It operates on three primary inputs: a query tensor, a value tensor, and, optionally, a critical tensor. If the critical tensor is absent, the value tensor acts as the critical tensor. The process starts by calculating attention scores (scores) derived from the dot

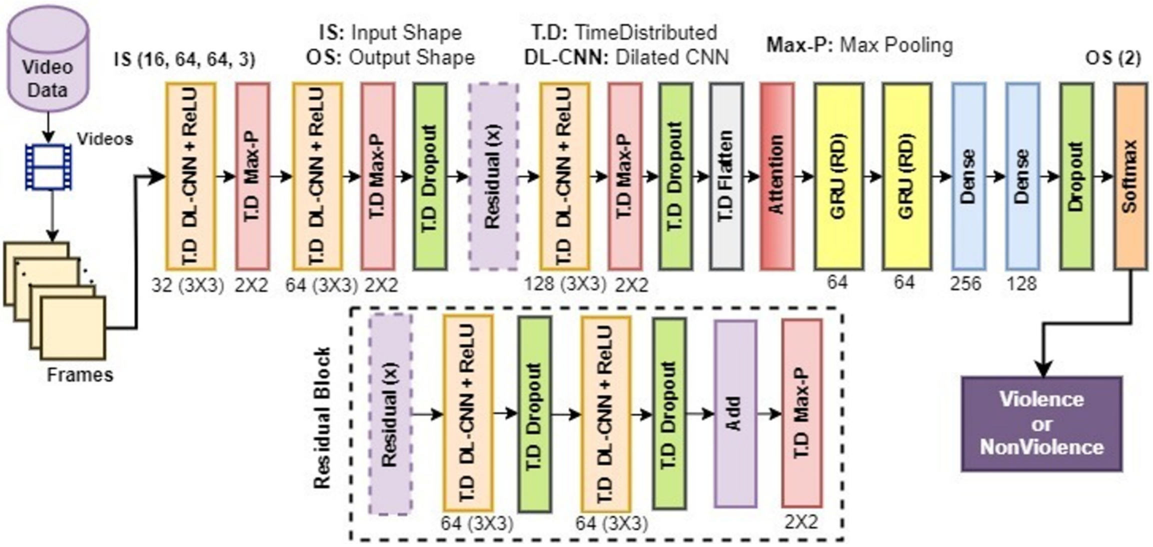

Fig.3: *Proposed ResDLCNN-GRU Attention Network Architecture.*

product of the query and critical tensors, normalized with softmax to obtain attention weights (α). These weights compute the final attention output (FAO) by weighting the value vectors (value). The Attention mechanism includes additional features such as scaling the attention scores, applying dropout to the scores, and using masks to control which elements are considered in the attention computation. The attention mechanism is mathematically defined as follows:

1. Calculate Attention Scores:

$$scores_{ij} = \frac{query_i \cdot critical_j^T}{\sqrt{d}} \tag{2}$$

Here, d is the dimension of the critical vectors.

2. Apply Softmax to Obtain Attention Weights:

$$\alpha_{ij} = \frac{\exp(scores_{ij})}{\sum_{k=1}^{T_v} \exp(scores_{ik})} \tag{3}$$

3. Compute the Attention Output:

$$FAO_i = \sum_{j=1}^{T_v} \alpha_{i,j} \cdot value_j \tag{4}$$

In Eqns. 2, 3, and 4, $query_i$ represents the i^{th} query vector, $critical_j$ represents the j^{th} critical vector, and $value_j$ represents the j^{th} value vector. We use the scaling factor ($\sqrt{d}$) to mitigate the effect of large dot products. The attention output is then passed to subsequent GRU units to process temporal dependencies within the sequence.

GRU Block: We employ Gated Recurrent Units (GRUs) for sequence modeling. Following the attention mechanism, GRUs model the temporal dependencies within the sequences. Weighting the input sequence based on attention scores allows GRUs to prioritize important information, improving their ability to capture and model temporal dependencies effectively. The first GRU layer, with 64 units, returns sequences, capturing temporal dependencies, while the second GRU layer, with 64 units, returns only the final output. We apply recurrent dropout (RD) with a rate of 0.15 for regularization. The GRU equations are defined as illustrated in Eqn. 5, 6, 7, and 8.

$$up_t = \sigma(W_{up}x_t + U_{up}hs_{t-1} + b_{up}) \tag{5}$$

$$rs_t = \sigma(W_{rs}x_t + U_{rs}hs_{t-1} + b_{rs}) \tag{6}$$

$$\widetilde{hs}_t = \tan(W_{hs}x_t + U_h(rs_t \odot hs_{t-1}) + b_h) \tag{7}$$

$$hs_t = (1 - up_t) \odot hs_{t-1} + up_t \odot \widetilde{hs}_t \tag{8}$$

Here, x_t represents the input at a time (t), up_t and r_{st} are update and reset gates, hs_t denotes the hidden state at time t, $\odot$ symbolizes element-wise multiplication, and σ represents the sigmoid activation function. The GRU layers capture sequential patterns within the flattened feature maps, facilitating the learning of temporal dependencies in the input data.

Fully Connected and Classification Layer: The model concludes with two fully connected dense layers with 256 and 128 units, followed by a dropout layer (0.15 rate). The final output layer employs a softmax activation function, rendering it suitable for a classification task with two output classes: Violence and NonViolence. The combination of dilated convolution, attention mechanism, and GRU layers provides a comprehensive approach for spatiotemporal feature learning in sequential data.

3.3 Loss Function Details

The Binary Cross-Entropy (BCE) loss function, also called logistic loss, is generally utilized in binary classification problems. It measures the difference between two probability distributions, typically the anticipated probabilities and the actual labels. Eqn. 9 depicts the mathematical expression of the BCE loss. Here, N represents sample count, y_t^k depicts the actual label for the k^{th} sample, and y_p^k symbolizes the anticipated probability that the k^{th} sample belongs to class 1. This loss function aims to prompt the projected probabilities to be close to 1 for positive instances and 0 for negative cases.

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{k=1}^{N}[y_t^k.\log(y_p^k) + (1 - y_t^k).\log(1 - y_p^k)] \tag{9}$$

Eqn. 12 represents the modified binary cross-entropy loss function with label smoothing. Eqns. 10 and 11 derive the T and U values, respectively. Here, N represents no. of samples in the batch, C signifies the no. of categories (for binary, $C = 2$), y_t^k represents the k^{th} sample's actual label, and y_p^k denotes the predicted probability for the k^{th} sample. This formulation incorporates the label smoothing factor (Lb_s), adjusting the actual labels y_t towards a uniform distribution across classes, helping to improve model generalization and robustness.

$$T = (1 - Lb_s).y_t^{(k)}.\log\left(y_p^{(k)}\right) \tag{10}$$

$$U = \frac{Lb_s}{C}.\log\left(\frac{1}{C}\right) \tag{11}$$

$$L(y_t, y_p) = -\frac{1}{N}\sum_{k=1}^{N}[T + U] \tag{12}$$

This modified loss function enhances the standard binary cross-entropy loss by incorporating label smoothing to improve generalization and robustness. The proposed ResDLCNN-GRU Attention model employs the Adam optimizer, leveraging an initial learning rate established at 0.001 for proficiently minimizing the model's associated error function.

4. EXPERIMENTAL RESULTS

The research presented here is conducted on Google Colab, employing the Python programming language focusing on resource efficiency. This section encompasses a detailed overview of our carefully curated EAVDD dataset, encompassing information on the datasets, a meticulous performance evaluation, and an in-depth analysis of the obtained results.

4.1 Dataset Details

Numerous openly available datasets on Violence recognition encompass Hockey Fights (HF) [29], Movie Fights [29], Violent-flows (ViF) dataset [28], VDD [27], and the Smart-City CCTV Violence Detection dataset (SCVD) [31]. The Hockey Fights and Movie Fights dataset contains data on fights in hockey games and movies. Additionally, we extracted some videos from the Hockey Fights and SCVD datasets. In this study, we present the **Extended Automatic Violence Dataset (EAVDD)**[1], an expansive collection of videos sourced from diverse platforms, including movies, TV episodes, and various social media platforms such as YouTube. EAVDD is distinguished for its comprehensive and unbiased representation of violent activities, spanning various individuals and scenarios, with a commitment to universality that avoids demographic or regional biases. The EAVDD dataset has two categories: 'Violence' and 'NonViolence' comprising 1530 videos, ensuring a broad spectrum of content for robust violence detection research. Fig. 4 depicts some video frames from the EAVDD.

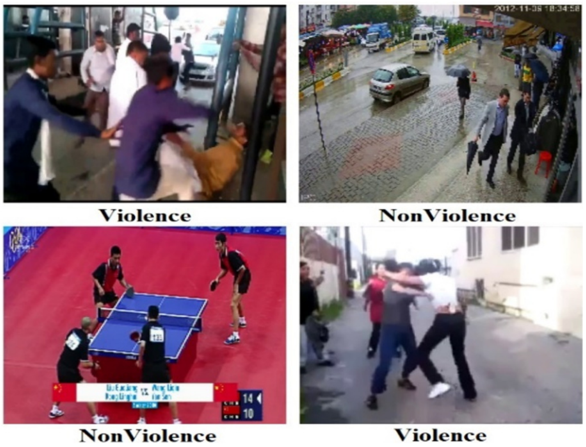

Fig.4: *EAVDD sample data.*

The EAVDD dataset covers scenes such as public Violence, bus Violence, campus Violence, and sports Violence (mainly hockey and football). Some video frames of the Hockey Fights (HF) dataset and SCVD dataset are depicted in Fig. 5 and Fig. 6, respectively.

Fig.5: *Hockey Fights sample data.*

[1] **EAVDD:**kaggle.com/datasets/arnab91/eavdd-Violence/

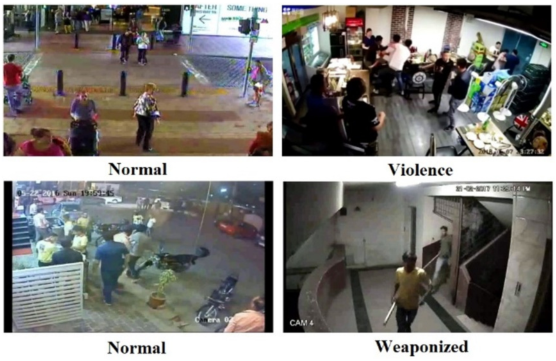

Fig.6: *SCVD sample data.*

4.2 Evaluation Metrics

Assessing the efficacy of a model encompasses a diverse array of metrics. Below are the formulas for the evaluation metrics utilized in this study:

1. **Accuracy:** It furnishes a comprehensive evaluation of the model's efficacy in accurately categorizing both instances of Violence and NonViolence. In Violence detection, true positives (TP) denote instances of Violence correctly identified, true negatives (TN) denote NonViolence instances correctly identified, false positives (FP) represent NonViolence instances incorrectly categorized as Violence, and false negatives (FN) represent violent instances incorrectly categorized as nonviolent. Eqn. 13 expresses the formula for accuracy concerning True Positives (TP), False Positives (FP), and all instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

2. **F1 Score($F1_s$):** It represents the harmonic mean of the precision and recall, wherein precision gauges the ratio of true positives within all positive predictions, and recall examines the ratio of true positives among all actual positives, specifically in the context of Violence detection. Eqn. 14 depicts the expression of the F1 Score.

$$F1_s = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{14}$$

3. **Sensitivity (True Positive Rate):** We refer to this metric as the true positive rate (TPR). It quantifies the proportion of genuine positive instances accurately identified by the proposed model. In Violence detection, sensitivity indicates how well the model captures instances of Violence without missing many (false negatives). Eqn. 15 mathematically expresses sensitivity.

$$Sensitivity = \frac{TP}{TP + FN} \tag{15}$$

4. **Area Under Curve (AUC):** AUC is a performance metric for assessing classification models. It represents the integral of the Receiver Operating Characteristic (ROC) curve, encapsulating the model's capacity to discriminate between positive and negative categories across various thresholds. AUC ranges from 0 to 1, with higher values indicating better performance. It provides a concise summary of classifier performance, independent of threshold choice. Here, TPR denotes the sensitivity, and the false positive rate (FPR) is represented by (1 - specificity) at various thresholds. The integral calculates the area under the ROC curve that encapsulates the model's capacity to discriminate between positive and negative categories across various threshold settings. Eqn. 16 mathematically represents the AUC-ROC score.

$$\text{AUC - ROC} = \int_0^1 \text{TPR(fpr) d(fpr)} \tag{16}$$

These metrics comprehensively evaluate a violence detection model's performance in correctly classifying violent and nonviolent instances.

4.3 Performance Evaluation

The efficacy of the proposed work is comprehensively evaluated across multiple metrics, encompassing validation accuracy score, F1 Score, AUC score, classification performance, and meticulous examination of the confusion matrix. Fig. 7 showcases the ROC curve generated utilizing the proposed ResDLCNN-GRU Attention model. It reveals that the proposed model achieves an AUC score of 0.97 for both the Violence and NonViolence categories in the EAVDD dataset.

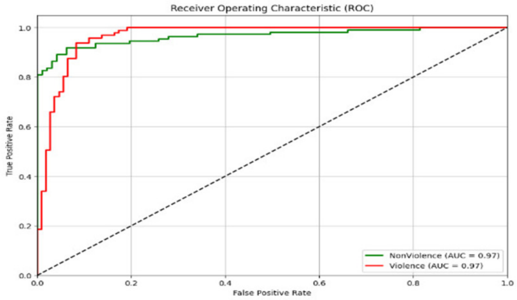

Fig.7: *ROC curve generated on EAVDD dataset.*

Fig. 8 comprehensively evaluates various deep learning models' performance trained from scratch on the EAVDD dataset, focusing on three key metrics: the validation accuracy score, F1 score, and AUC score. These metrics are crucial indicators of a model's efficacy in classification tasks. The models considered for evaluation are ResNet3D+A [13], Two Stream CNN [32], Xp+BiLSTM+A [34], ConvLSTM [37], ResDC-GRU+A [38], and the proposed model on the EAVDD dataset. The proposed ResDLCNN-GRU Attention model excels significantly, achieving an excellent 0.92 validation score, surpassing Xception+BiLSTM+A [34], ConvLSTM [37], ResDC-GRU+A [38], and all other models. The proposed model achieves the highest validation accuracy, indicating its superior capability in generalizing to unseen data compared to other models evaluated. The Xception+BiLSTM+A [34] and ResDC-GRU+A [38] models also demonstrate competitive validation accuracies, showcasing their effectiveness in Violence classification. ResNet3D+A [13], Two Stream CNN [32], ConvLSTM [37], and also demonstrated satisfactory performance, achieving validation accuracy scores of 0.905, 0.863, and 0.897, respectively.

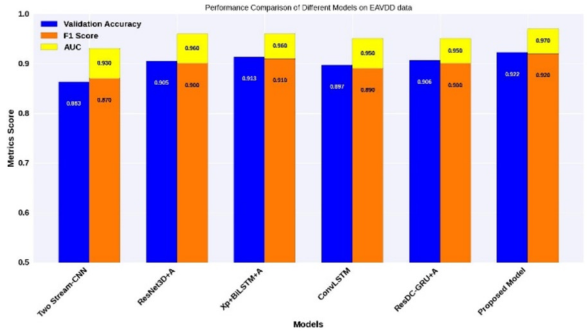

Fig.8: *Assessment of models on the EAVDD dataset.*

Moving on to the F1 Score, a measure that balances precision and recall, the values vary between 0.87 and 0.92. Once again, the proposed model asserts its dominance by achieving a superior F1 Score of 0.92, underscoring its robustness in classifying instances correctly across different classes. The Xp+BiLSTM+A [34] model also exhibits a commendable F1 Score of 0.91, indicating its strong performance. Finally, the AUC score, indicative of the model's capability to discriminate between positive and negative instances within the EAVDD dataset, spans from 0.93 to 0.97. The proposed model achieves the highest AUC score of 0.97, reaffirming its efficacy in distinguishing between classes. Notably, the ResNet3D+A [13] and Xp+BiLSTM+A [34] models also demonstrate the second-highest AUC score of 0.96, suggesting their competence in making accurate predictions across the dataset. Thus, the proposed model outperforms alternative standard deep learning architectures concerning validation accuracy, AUC score, and F1 Score, underscoring its effectiveness in handling Violence classification within the EAVDD dataset.

Confusion Matrix (CM): It is a compact and insightful visual representation summarizing a model's predictions against actual ground truth values. This matrix offers a quick snapshot of a model's performance, enabling a nuanced assessment of its accuracy and effectiveness across different classes. Fig. 9 displays the confusion matrix for the EAVDD data generated using the proposed ResDLCNN-GRU Attention model.

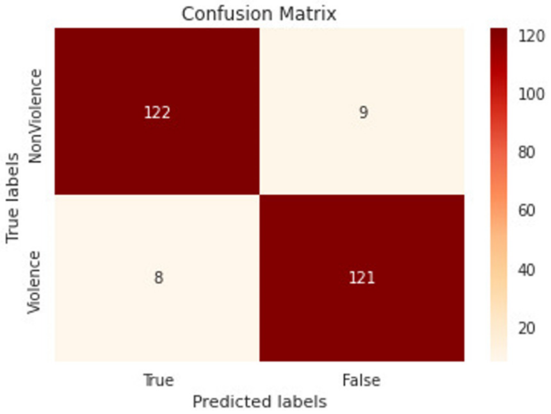

Fig.9: *CM of EAVDD dataset using the proposed model.*

The suggested model showcases exceptional classification accuracy on the Violence dataset. Evaluation outcomes of the suggested model on unseen (not trained) video samples taken from various movie scenes, as illustrated in Fig. 10 and Fig. 11. It took only about 1.02 minutes to identify and analyze the video frame by frame.

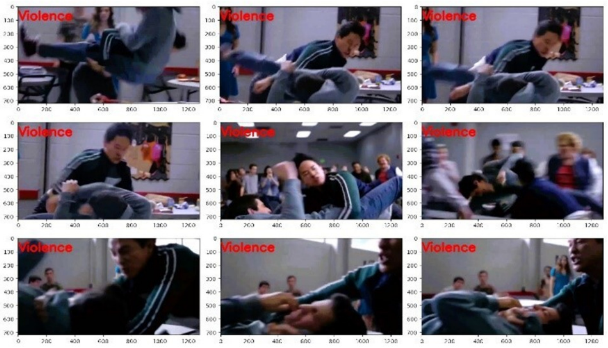

Fig.10: *Test Result of Violence video.*

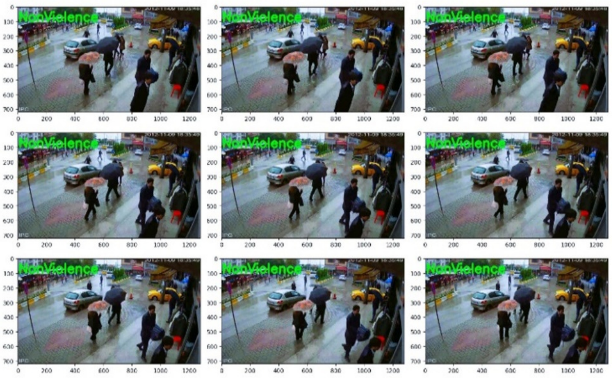

Fig.11: *Test Result of NonViolence video.*

Effect of Loss Function:

Table 1 presents the evaluation results of a proposed model using different loss functions trained over a standard number of epochs with the Adam optimizer. We kept the initial learning rate constant at 0.001 and maintained a batch size of five throughout the loss function analysis. Mean Squared Error yielded a validation accuracy of 89.65% and a 0.88 true positive rate (TPR). Conversely, Binary Cross-Entropy (BCE) achieved slightly higher performance, with a validation accuracy of 91.60% and a TPR of 0.90.

Table 1: *Loss Function Analysis on the EAVDD.*

Loss Function	Optimizer	Val. Acc (%)	TPR
Mean Square Error	Adam	89.65	0.88
BCE	Adam	91.60	0.90
CCE	Adam	92.20	0.91
MBCE	**Adam**	**92.16**	**0.91**

Modified Binary Cross-Entropy (MBCE) emerged as a powerful loss function, boasting a validation accuracy of 92.16% and a TPR of 0.91, showcasing superior performance compared to the other loss functions analyzed. Also, Categorical Cross entropy (CCE) loss achieves 92.20% validation accuracy with 0.91 TPR. These findings underscore the importance of selecting an appropriate loss function, with MBCE and CCE demonstrating significant efficacy in enhancing model performance.

4.4 Result Analysis

Table 2 presents a comprehensive comparative evaluation of various deep learning models utilized for Violence detection across different benchmark datasets, namely Hockey Fights (HF) [29], Movie Fights [29], and SCVD [31]. Each model's performance is measured in terms of classification accuracy, providing valuable insights into their effectiveness in detecting violent content within videos. Among the models evaluated, several notable observations emerge. The ViolenceNet [21] approach achieves impressive results across both HF [29] and Movie Fights [29] datasets, surpassing the majority of other methods with accuracy scores of 99.20% and 100%, respectively. On the Hockey Fights dataset, Efficient 3DCNN [22], Xception+BiLSTM+Attention [34], and ResDC-GRU+Attention [38] attain 98.36%, 97.50%, and 97.72% accuracy, respectively. The AlexNet+LSTM [16], Two Stream CNN [32], and T-MobileNet [35] models attain 97.10%, 92.17%, and 87.0% accuracy, respectively, on the Hockey Fights dataset. Meanwhile, C3D+SVM [24] and ResNet50+POT [33] attained 98.50% and 95% accuracy. On the Movie Fights dataset, AlexNet+LSTM [16], Efficient 3DCNN [22], Xception+BiLSTM+A [34] and ResDC-GRU+A [38] attain 100%, 99.17%, 100%, and 98.65% accuracy respectively. The ResNet3D+Attention [13] model also demonstrates good performance, particularly excelling in the Hockey Fights and Movie Fight datasets with 98.10% and 100% accuracy, respectively. It underscores the significance of incorporating attention mechanisms within convolutional neural networks (CNNs) to effectively capture spatial-temporal features crucial for violence detection.

Table 2: *Comparative Evaluation of Various Methods in Violence Detection.*

Method (s)	HF	Movie Fights	SCVD
TL-3DCNN [4]	92.90	98.70	−−
ResNet3D+A [13]	98.10	100	89.26
AlexNet+LSTM [16]	97.10	100	87.42
ViolenceNet [21]	99.20	100	89.70
Efficient 3DCNN [22]	98.36	99.17	86.40
C3D+SVM [24]	98.50	96.80	85.37
C3D+FC [25]	96.0	97.24	85.82
Two Stream CNN [32]	92.17	94.36	74.38
R+ResNet50+POT [33]	95.0	97.50	−−
Xception+BiLSTM+A [34]	97.50	100	90.14
T-MobileNet [35]	87.0	99.50	−−
ResDC-GRU + A [38]	97.72	98.65	89.76
Proposed Model	**98.38**	**99.62**	**90.57**

On the SCVD dataset, ViolenceNet [21], Xception+BiLSTM+A [34], and ResDC-GRU+A [38] attain 89.70%, 90.14%, and 89.76% accuracy, respectively. Furthermore, the Proposed Model stands out prominently across all datasets, attaining outstanding accuracy scores among evaluated methods-98.38%, 99.62%, and 90.57% on HF [29], Movie Fights [29], and SCVD [31] datasets, respectively. These results underscore its robustness and generalizability in detecting violent content across diverse video sources. The proposed model comprises only 0.65 million (M) trainable parameters. Additionally, models such as ResNet3D+A [13], C3D+FC [25], and R+ResNet50+POT [33] also deliver commendable performance, consistently achieving high accuracy scores across multiple datasets. Thus, the comparative evaluation underscores the effectiveness of various deep learning-based models in Violence detection tasks, with the Proposed Model demonstrating notable performance across all benchmark datasets.

The F1 Score ($F1_s$) obtained on both the HF dataset [29] and SCVD [31] dataset highlights the robust classification capabilities exhibited by a range of standard models, as illustrated in Fig. 12. The SCVD dataset stands as a standard benchmark dataset, encompassing a diverse array of CCTV footage videos. The proposed ResDLCNN-GRU Attention model performs well in identifying violent actions, showcasing remarkable classification performance with 0.91

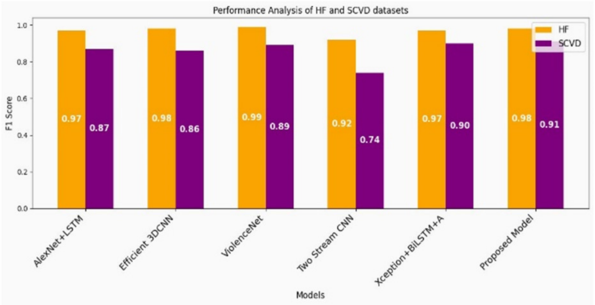

Fig.12: *Performance Analysis on HF and SCVD data.*

$F1_s$ on the SCVD dataset [31] and an impressive 0.98 $F1_s$ on the Hockey Fights dataset.

Some of the top fight scenes in movies are analyzed using the proposed model and presented in Table 3. The Fight Ratio serves as a metric representing the ratio of violent action time to the total duration of the scene. A higher Fight Ratio suggests more intense or sustained action relative to the scene's length. Based on these metrics, IP Man vs. Mike Tyson (IPMAN3) and Bus Fight Scene 1 (Nobody) have the highest Fight Ratio values at 0.73 and 0.72, respectively, indicating a significant portion of the scene is filled with violent action relative to its duration.

Table 3: *Movie Fight Scene Analysis in Violence Detection.*

Scene Name, Movie, and Year	Total duration (min.)	Violence Time (min.)	Fight Ratio
IP Man vs. Mike Tyson, IPMAN3, and 2015	3.42	2.50	0.73
Chung Tin-chi vs. Tony Jaa, Master Z: IP Man Legacy, and 2018	2.25	1.57	0.70
Cliffside Shutdown Scene 10, Mission: Impossible -Fallout, and 2018 [39]	2.25	1.32	0.59
Bus Fight Scene 1, Nobody, and 2021 [39]	4.31	3.11	0.72

Similarly, the Chung Tin-chi vs. Tony Jaa scene (Master Z: IP Man Legacy) and Cliffside Shutdown Scene (Mission: Impossible-Fallout) also have a standard Fight Ratio of 0.70 and 0.59, respectively. These scenes are among the most intense or action-packed based on this analysis.

4.5 Complexity Analysis

The analysis of various models based on the number of trainable parameters (Params) in Fig. 13 provides valuable insights into their complexity and computational requirements. Among the models considered, AlexNet+LSTM [16] and Xception+BiLSTM Attention [34] exhibit the highest complexity with 9.6 million (M) and 9 million trainable parameters, respectively, suggesting a significant demand for computational resources. The Efficient 3DCNN [22] model has 7.4 million trainable parameters. ViolenceNet [21] follows with 4.5 million trainable parameters, indicating a moderate level of complexity.

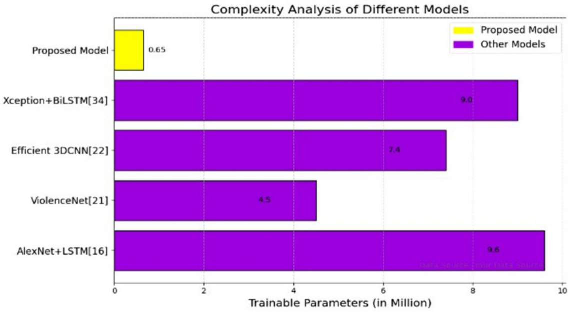

Fig.13: *Model Analysis based on Trainable Params.*

The proposed model has only 0.65 million (M) trainable parameters, significantly reducing the computational burden. This minimal parameter count highlights its suitability for deployment on low-computation machines [36], such as general-purpose computers. Despite its low complexity, the proposed model demonstrates comparable performance across various metrics and proves effective in real-world applications. Fig. 14 depicts some test results generated from the proposed model, correctly identifying fight scene videos as violent.

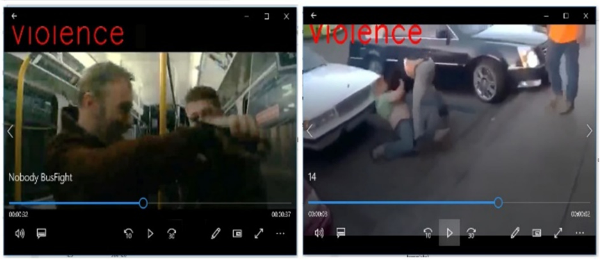

Fig.14: *Test on some Fight Scene video.*

5. APPLICATIONS

The proposed work on violence detection has broad applications, especially in addressing the growing concerns surrounding the spread of violent content online. Primarily, they aid in content moderation on social media platforms by identifying and removing violent material, enhancing user safety. Law enforcement agencies also benefit from these tools, utilizing them to monitor social media for potential threats or acts of Violence and intervening early. These tools ensure compliance with content regulations and provide immediate support to those exposed to Violence, linking them to counseling resources. Beyond social media, violence detection systems are valuable in educational settings, monitoring for bullying, campus Violence, and harassment, allowing timely interventions. In crowd behavior recognition [40], such as concerts and protests, these systems help security personnel de-escalate potentially violent situations, detect weapon-related Violence [41] [45], and ensure public safety. These systems enhance surveillance by swiftly identifying and mitigating violent incidents in public places such as railways [42], airports, schools [43], shopping malls, buses, elevators [44], and streets. These applications demonstrate the potential of violence detection systems to create safer, more secure, and healthier environments worldwide, not only on social media but also in educational institutions, public spaces, and large gatherings.

6. CONCLUSION

This research introduces a novel methodology for efficiently detecting violent activities in video streams. The suggested model leverages a Residual DLCNN to extract spatial features, integrates an attention mechanism to prioritize crucial frames, and employs GRU for temporal features and a dense layer with Softmax for classification, maintaining resource efficiency with just 0.65 million trainable parameters. Training and validation across three datasets—Hockey Fights, Movie Fights, and SCVD—demonstrate the model's strong recognition performance. Moreover, the proposed model can detect Violence in fight scenes, sports, and CCTV footage. Despite its computational efficiency, the model is highly effective, making it particularly beneficial for time-sensitive applications. These advancements would contribute to the ongoing evolution and refinement of Violence detection systems, fostering safer environments in various domains where early detection is paramount. The broad applications of Violence detection, from content moderation in social media to crisis response, underscore the significance of continued research in this field. Future work should focus on expanding the collected EAVDD dataset, incorporating more categories of violent actions, and implementing lightweight transformers to enhance robustness and accuracy, ensuring the continued progress of Violence detection systems.

ACKNOWLEDGEMENT

The authors acknowledge IIEST, Shibpur, for providing support in carrying out the proposed work.

AUTHOR CONTRIBUTIONS

Conceptualization, A.D.; methodology, A.D.; validation, A.D. and S.B.; formal analysis, S.B.; investigation, A.D. and S.B.; data curation, A.D.; writing—original draft preparation, A.D.; writing—review and editing, A.D., S.B., and L.A.; visualization, A.D., S.B., and L.A.; supervision, S.B. All authors have read and agreed to the published version of the manuscript.

References

[1] H. Yao and X. Hu, "A survey of video Violence detection," *Cyber-Physical Systems*, pp. 1–24, Jun. 2021.

[2] G. Kaur and S. Singh, "Revisiting vision-based violence detection in videos: A critical analysis," *Neurocomputing*, vol. 597, 2024,

[3] I. Serrano, O. Deniz, J. L. Espinosa-Aranda and G. Bueno, "Fight Recognition in Video Using Hough Forests and 2D Convolutional Neural Network," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4787-4797, 2018.

[4] A. S. Keceli and A. Kaya, "Violent activity classification with transferred deep features and 3d-Cnn," Signal, Image and Video Processing, 2023.

[5] Q. Dai P*et al.*, "Fudan-Huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with Deep Learning," in *MediaEval*, vol. 1436, 2015.

[6] Naz Dündar, Ali Seydi Keçeli, A. Kaya, and H. Sever, "A shallow 3D convolutional neural network for Violence detection in videos," *Egyptian Informatics Journal*, vol. 26, 2024.

[7] P. Zhang, L. Dong, X. Zhao, W. Lei, and W. Zhang, "An end-to-end framework for real-time violent behavior detection based on 2D CNNs," *Journal of real-time image processing*, vol. 21, no. 2, 2024.

[8] J. Mahmoodi and Hossein Nezamabadi-pour, "A spatio-temporal model for Violence detection based on spatial and temporal attention modules and 2D CNNs," *Pattern Analysis and Applications*, vol. 27, no. 2, 2024.

[9] S. M. Mohtavipour *et al.*, "A multistream CNN for deep Violence detection in video sequences using handcrafted features," *The Visual Computer*, vol. 38, no. 6, pp. 2057-2072, 2022.

[10] V. D. Huszar *et al.*, "Toward Fast and Accurate Violence Detection for Automated Video Surveillance Applications," *IEEE Access*, vol. 11, pp. 18772-18793, 2023.

[11] T. Zhenhua *et al.*, "FTCF: Full temporal cross fusion network for Violence detection in videos," *Applied Intelligence*, vol. 53, no. 4, pp. 4218-4230, 2023.

[12] G. Garcia-Cobo and J. C. SanMiguel, "Human skeletons and change detection for efficient Violence detection in surveillance videos," *Com-*

puter Vision and Image Understanding, vol. 233, p. 103739, 2023.

[13] J. H. Park *et al.*, “Conv3D-Based Video Violence Detection Network Using Optical Flow and RGB Data,” *Sensors*, vol. 24, no. 2, p. 317, 2024.

[14] K. Chaturvedi *et al.*, “Fight detection with spatial and channel wise attention-based ConvLSTM model,” *Expert Systems*, vol. 41, no. 1, p. e13474, 2024.

[15] A. Dey, S. Biswas, and L. Abualigah, “Umpire’s Signal Recognition in Cricket Using an Attention based DC-GRU Network,” *International Journal of Engineering*, vol. 37, no. 4, pp. 662–674, 2024.

[16] S. Sudhakaran and O. Lanz, “Learning to detect violent videos using convolutional long short-term memory,” *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Lecce, Italy, 2017, pp. 1-6.

[17] C. Li, X. Yang, and G. Liang, “Keyframe-guided Video Swin Transformer with Multi-path Excitation for Violence Detection,” *The Computer Journal*, vol. 67, no. 5, pp. 1826-1837, 2024.

[18] A. Dey, S. Biswas, and D.-N. Le, “Recognition of Human Interactions in Still Images using AdaptiveDRNet with Multi-level Attention,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 10, 2023.

[19] N. Alabid, “Interpretation of Spatial Relationships by Objects Tracking in a Complex Streaming Video,” *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 15, no. 2, pp. 245-257, 2021.

[20] S. Mekruksavanich and A. Jitpattanakul, “FallNeXt: A Deep Residual Model based on Multi-Branch Aggregation for Sensor-based Fall Detection,” *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 16, no. 4, pp. 352–364, 2022.

[21] F. J. Rendón-Segador, J. A. Álvarez-García, F. Enríquez, and O. Deniz, “ViolenceNet: Dense Multi-Head Self-Attention with Bidirectional Convolutional LSTM for Detecting Violence,” *Electronics*, vol. 10, no. 13, p. 1601, 2021.

[22] J. Li, X. Jiang, T. Sun and K. Xu, “Efficient Violence Detection Using 3D Convolutional Neural Networks,” *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Taipei, Taiwan, 2019, pp. 1-8.

[23] Y. Tang, Y. Chen, Sagar A.S.M. Sharifuzzaman, and T. Li, “An automatic fine-grained Violence detection system for animation based on modified faster R-CNN,” *Expert systems with applications*, vol. 237, 2024.

[24] S. Accattoli *et al.*, “Violence detection in videos by combining 3D convolutional neural networks and support vector machines,” *Applied Artificial Intelligence*, vol. 34, no. 4, pp. 329-344, 2020.

[25] F. U. M. Ullah *et al.*, “Violence detection using spatiotemporal features with 3D convolutional neural network,” *Sensors*, vol. 19, no. 11, p. 2472, 2019.

[26] M. Khan, A. E. Saddik, W. Gueaieb, G. De Masi and F. Karray, “VD-Net: An Edge Vision-Based Surveillance System for Violence Detection,” *IEEE Access*, vol. 12, pp. 43796-43808, 2024.

[27] M. Bianculli *et al.*, “A dataset for automatic Violence detection in videos,” *Data in brief*, vol. 33, 2020.

[28] T. Hassner *et al.*, ”Violent flows: real-time detection of violent crowd behavior,” *2012 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Rhode Island, USA, June 2012.

[29] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, “Violence Detection in Video Using Computer Vision Techniques,” *Computer Analysis of Images and Patterns*, pp. 332–339, 2011.

[30] G. Kaur and S. Singh, “An ensemble based approach for Violence detection in videos using deep transfer learning,” *Multimedia Tools and Applications*, 2024.

[31] Toluwani Aremu, L. Zhiyuan, Reem Alameeri, M. Khan, and Abdulmotaleb El Saddik, “SSIVD-Net: A Novel Salient Super Image Classification and Detection Technique for Weaponized Violence,” *Lecture notes in networks and systems*, pp. 16–35, 2024.

[32] W. Dai *et al.*, “Two-stream convolution neural network with video-stream for action recognition,” *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, July 2019.

[33] N. Honarjoo, A. Abdari, and A. Mansouri, “Violence detection in compressed video,” *Multimedia Tools and Applications*, 2024.

[34] S. Akti, G. A. Tataroglu, and H. K. Ekenel, “Vision-based Fight Detection from Surveillance Cameras,” *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Nov. 2019.

[35] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, “Cover the Violence: A Novel Deep-Learning-Based Approach Towards Violence-Detection in Movies,” *Applied Sciences*, vol. 9, no. 22, 2019.

[36] J. Wang, D. Zhao, H. Li, and D. Wang, “Lightweight Violence Detection Model Based on 2D CNN with Bi-Directional Motion Attention,” *Applied Sciences*, vol. 14, no. 11, 2024.

[37] Z. Zhang, Z. Lv, C. Gan, and Q. Zhu, “Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions,” *Neurocomputing*, vol. 410, pp. 304-316, 2020.

[38] A. Dey, S. Biswas, and D.-N. Le, "Workout Action Recognition in Video Streams Using an Attention Driven Residual DC-GRU Network," *Computers, Materials & Continua*, vol. 79, no. 2, pp. 3067-3087, 2024.

[39] Top Fight Scenes in Movies [Online Available]: `https://www.timeout.com/film/greatest-fights-in-movies`.

[40] M. Qaraqe *et al.*, "PublicVision: A Secure Smart Surveillance System for Crowd Behavior Recognition," *IEEE Access*, vol. 12, pp. 26474-26491, 2024.

[41] Muhammad Shahroz Nadeem, Fatih Kurugollu, H. F. Atlam, and Virginia N.L. Franqueira, "Weapon Violence Dataset 2.0: A synthetic dataset for violence detection," Data in Brief, 2024.

[42] T. Marteau, Sitou Afanon, D. Sodoyer, and Sebastien Ambellouis, "Violence detection in railway environment with modern deep learning approaches and small dataset," *Transportation research procedia*, vol. 72, pp. 87–92, 2023.

[43] E. Perseghin and Gian Luca Foresti, "A Shallow System Prototype for Violent Action Detection in Italian Public Schools," *Information*, vol. 14, no. 4, 2023.

[44] J. Lei, W. Sun, Y. Fang, N. Ye, S. Yang, and J. Wu, "A Model for Detecting Abnormal Elevator Passenger Behavior Based on Video Classification," *Electronics*, vol. 13, no. 13, 2024.

[45] T. Santos, H. Oliveira, and A. Cunha, "Systematic review on weapon detection in surveillance footage through deep learning," *Computer science review*, vol. 51, 2024.

Arnab Dey is currently a Research Scholar in the Department of Computer Science and Technology at the Indian Institute of Engineering Science and Technology (IIEST), Shibpur, India. His research interests include Computer Vision, Image and Video Processing, Action Recognition, Deep Learning, Intelligent Systems, and Optimization, among others. He has published articles in reputed international journals and conference proceedings indexed by SCIE, ESCI, and Scopus. He also reviews for various international journals and conferences. He is a member of professional organizations, including the IEEE Biometrics Council and IAENG. He was a member of the E-Cell at IIT Kanpur for one year.

Samit Biswas is currently working as an Assistant Professor in the Department of Computer Science and Technology at the Indian Institute of Engineering Science and Technology (IIEST), Shibpur, India. His research interests include Image Processing and Analysis, Computer Vision, Machine Learning, Document Image Processing, Natural Language Processing, and Pattern Recognition, among others. He has published various articles in reputed journals such as Pattern Recognition, IET Image Processing, CMC and IJDAR, as well as in conference proceedings indexed by SCIE, ESCI, and Scopus. He is a reviewer for various international journals and conferences. He is also a member of Indian Unit of the Pattern Recognition and Artificial Intelligence (IUPRAI).

Laith Abualigah is the Director of the Department of International Relations and Affairs at Al Al-Bayt University, Jordan. He is an Associate Professor at the Computer Science Department, Al Al-Bayt University, Jordan. He is also a distinguished researcher at many prestigious universities. He received a Ph.D. degree from the School of Computer Science at Universiti Sains Malaysia (USM), Malaysia, in 2018. According to the report published by Clarivate, He is one of the Highly Cited Researchers in 2021-2023 and the 1% influential Researcher, which depicts the 6,938 top scientists in the world. In addition, the first researcher in the domain of Computer Science in Jordan for 2021-2023. According to the report published by Stanford University, He is one of the 2% influential scholars, which depicts the 100,000 top scientists in the world. He has published more than 500 journal papers and books, which collectively have been cited more than 20900 times (H-index = 64).