VioNet: An Enhanced Violence Detection Approach for Videos Using a Fusion Model of Vision Transformer with Bi-LSTM and 3D Convolutional Neural Networks

Md. Akil Raihan Iftee, Md. Mominur Rahman, and Sunanda Das*

Department of Computer Science and Engineering,
Khulna University of Engineering & Technology, Khulna-9203, Bangladesh
iftee1807002@gmail.com, mominurrahman229@gmail.com, sunanda@cse.kuet.ac.bd

Abstract. The identification of violence in real-world scenarios is imperative as it enables the detection of aggressive behavior, thereby preventing harm to individuals and communities. This is crucial for ensuring public safety, facilitating effective crime investigation, promoting child safety, safeguarding mental health, and facilitating social media moderation. Various methods, including handcrafted techniques and deep learning algorithms, can be utilized in surveillance or CCTV cameras, as well as smartphones, to enable timely detection of violent behavior and facilitate appropriate action and intervention. In this study, we introduce VioNET, a novel approach that combines a 3D Convolutional Neural Network and a Vision Transformer with Bidirectional LSTM for the purpose of accurately detecting violence in video data. Since video data is inherently sequential, the extraction of spatiotemporal features is essential to accurate detection. The use of these two deep learning methods facilitates the extraction of maximum features, which are then fused together to classify videos with the highest possible accuracy. We evaluate the effectiveness of our approach by employing three datasets: Hockey, Movies, and Violent Flow, for analysis. The proposed model achieved impressive accuracies of 97.85%, 100.00%, and 96.33% on the Hokey, Movie, and Violent Flow datasets, respectively. Based on the obtained results, it is evident that our method showcases superior performance, outperforming several existing approaches in the field and establishing itself as a robust and competitive solution for violence detection in videos.

Keywords: Violence Detection · Vision Transformer · 3D-Neural Network.

1 Introduction

Video has become an integral part of modern society, serving as a primary means of communication and entertainment. According to estimates, YouTube alone

* Corresponding author:
Sunanda Das (orcid: 0000-0002-7164-6859)

receives almost 1 billion hours of video views each day. Similarly, Facebook, Instagram, Tiktok, and Twitter, social media platforms are home to large amounts of video content. However, not all video content is benign. Some videos contain violent or disturbing content that can have negative impacts on individuals and society as a whole. In particular, exposure to violent content can adversely affect the well-being of children and adolescents and disrupt social harmony. In recent times, the proliferation of surveillance cameras and CCTV has captured graphic footage of violent incidents that highlight the alarming levels of aggression and desensitization in both young and older individuals.

Effective violence detection in video data can have a significant impact on society, as it can help identify and prevent violent incidents, protect viewers from disturbing content, and aid law enforcement authorities in reducing crime rates. To achieve this goal, specialized techniques are required for processing and analyzing video data, which differ from those used for static image data.

Initially, traditional approaches are employed for violent video classification. Traditional video classification methods often rely on manually crafted features, such as color histograms, optical flow, or hand-engineered motion features. These features may not be optimal for capturing complex and high-level spatial and motion patterns in videos. These limitations can result in reduced performance and adaptability to diverse video domains and tasks. Moreover, Traditional approaches often involve multiple stages, such as feature extraction, feature selection, and classification, which are typically done separately. This can result in suboptimal performance, as the features and classifiers are not optimized jointly in an end-to-end manner.

Video data classification involves feature extraction and motion analysis from continuous frames, which can be performed using deep learning-based approaches. These methods have shown high accuracy in detecting violence in videos, and their application can promote a safer and more secure environment for individuals and communities. One of the accurate solutions comes from 2D-CNN with RNN which allows for the extraction of both visual and temporal characteristics from video data. CNNs can capture spatial features from video frames, while RNNs can model temporal dependencies between frames, enabling the model to capture both appearance and motion information for improved video classification performance. While RNNs have advantages in modeling sequential data, they also have limitations in capturing long-term dependencies, computational complexity, risk of vanishing or exploding gradients, especially when dealing with long video sequences, limited ability to capture spatial information, and difficulty in handling variable-length sequences in video classification tasks.

The VioNet model, which combines the benefits of Vision Transformers with Bidirectional Long Short-Term Memory (LSTM) networks and 3D-CNN for effective feature extraction, is used in this paper to provide a hybrid technique for the categorization of violent videos. While the Vision Transformer captures fine-grained visual details and the Bi-LSTM models temporal dependencies, the 3D CNN effectively captures spatio-temporal information. The combined feature

finally came to a neural network and analyze whether the video contains any violence or not. Our model can capture intricate spatial and motion patterns. Moreover, it is more reliable than the traditional machine learning approaches for auto feature selection and multi-dimensional feature capturing. By using Bidirectional LSTM can capture both past and future temporal dependencies in a video sequence whereas RNNs can only capture past temporal dependencies. This enables our model to gain a better understanding of the violent context within the current frame and its relationship with the surrounding frames, resulting in its superior performance compared to other current leading methods.

2 Related Works

The violence detection method depends on identifying and categorizing various types of action based on spatiotemporal features from every video frame. Several machine learning and deep learning algorithm with image processing achieve the best accuracy to extract these features.

In previous works, spatiotemporal features had been extracted using handcrafted methods; MoSIFT (Motion Scale-Invariant Feature Transform) [1] and STIP (Space-Time Interest Points) [2]. In terms of image feature extraction, MoSIFT, and STIP are the most popular methods. STIP captures the motion and shape of moving objects, while MosIFT captures appearance and motion information.

In their work [3], P. Zhou et al. employed LHOF and LHOG for the extraction of low-level features from RGB and optical flow images, respectively. To classify images, Das et al. [4] used Histogram of Oriented Gradients(HOG) with Logistic Regression, Logistic Distribution Adaptation, Latent Dirichlet Allocation, and Random Forest.

Image feature extraction using deep learning algorithms has shown impressive results. Convolutional Neural Networks make use of the spatial relationship between pixels to gain insight into images. In [5], Narges Honarjoo et al. used 1D –CNN where spatial features were extracted using pre-trained(VGG) and Resnet. Data efficient video transformer and pre-trained model of 2d-CNN was proposed by Abdali et al. [6]. An LSTM is a special type of RNN that captures the spatial dependencies between regions by using feature vectors as inputs. M. Su et al. [7] had proposed Bidirectional Convolution LSTM. In [7], a 3d-Convulation and Holistic and Localized model for extracting both visual and audio features was implemented. Israel Mugunga et al. [8] used ConvLSTM network as a solution of this classification. Patel [9] proposed, Human Pose estimation using PoseNET with LSTM and CNN using ResNET 50 with LSTM was proposed. To extract spatiotemporal features, 3D-Convulation and 3D-Pooling were applied by Song [10]. Using convolutional neural networks along with LSTMs, Sudhakaran [11] et al. proposed to capture spatiotemporal information.

Bruno Peixoto [12] et al. used C3D, LSTM, and combined them with CNN for visual-based violence classification, as well as four audio feature extractors to generate audio-based features that were applied to statistical methods for

improved accuracy. Theodoros Giannakopoulos [13] et al. used automated processing and analysis of audio and visual signals, followed by meta-classification. Bruno M. Peixoto [14] additionally trained independent audio and visual feature detectors, which were later integrated to form a decision tree neural network.

3 Proposed Method

First, the input video is divided into 16 frames, which are then stacked and sent into the network. In our proposed model, features are extracted using two different approaches. The input tensor is duplicated twice, and each duplicate is processed separately using a distinct architecture. The extracted features from the separate architectures are then fused and passed into a final neural network, which ultimately makes a decision on whether the video contains any violence or not.

Vision transformer with Bidirectional LSTM: This architecture consists of two main parts. One of them involves extracting spatial features from the preprocessed input tensor using a vision transformer model. Each frame of the input tensor, which has a shape of (l, w, c), is split into n square patches. Each patch, with a shape of (h, h, c), is then flattened into a one-dimensional vector of shape $(1, h^2c)$. All patches are linearly flattened to produce a vector X with dimension $[n, h2c]$. These one-dimensional vectors are then passed through a dense layer that outputs D-dimensional embeddings E$[n, D]$, where n indicates the quantity of patches in total. To ensure a E$[n+1, D]$ dimensional embedding, a C$[1, D]$ class embedding is added. Next, a positional embedding tensor of $(n + 1, D)$ dimensional is merged with the patch embeddings.

$$n = \frac{l}{h} \times \frac{w}{h} = \frac{l \times w}{h^2} \tag{1}$$

$$E_{n \times D} = X_{n \times h^2 c} \times W^T_{h^2 c \times D} + b_{n \times D} \tag{2}$$

$$E_{patch} = E_{n \times D} + C_{1 \times D} \tag{3}$$

$$E_{n+1 \times D} = E_{patch} + E_{positional} \tag{4}$$

The second part involves passing the output vector from the first part through a transformer encoder, which consists of a multi-head attention (MHA) layer, a fully connected layer or a dense layer, and a residual connection that links two sub-layers, then a normalization layer. Hidden state dynamics are stabilized via layer normalization which shortens training time and its output, Z, is used to produce queries Q, values V and keys K through a scaled dot product with three weight vectors. The resulting queries vector is multiplied with the transpose of keys vector and then it went via a SoftMax function and multiplied by the values

vector to create a head H of $(n+1, D)$ dimensions. This multi-head attention block is concatenated h times. Finally, the output goes through a dense layer to obtain the final embedded vector of dimension D.

$$Q_{n+1\times D} = E_{n+1\times D} + Wq_{D\times D} \tag{5}$$

$$K_{n+1\times D} = E_{n+1\times D} + Wk_{D\times D} \tag{6}$$

$$V_{n+1\times D} = E_{n+1\times D} + Wv_{D\times D} \tag{7}$$

$$Z_{n+1\times n+1} = Q_{n+1\times D} \times K^T_{n+1\times D} \tag{8}$$

$$H_{n+1\times D} = softmax(\frac{Z_{n+1\times n+1}}{\sqrt{D}}) \times V_{n+1\times D} \tag{9}$$

In our case, the input tensor is $(128, 128, 3)$ in shape and $(16, 16)$ patch size has been taken. (Dimension D, how many multi-head h = 8) The final embedded vector is a 16×2048-dimensional vector produced. The feed-forward neural network is then given this embedded vector to generate a $(16, 300)$ - dimensional vector that contains spatial features. This vector is then treated as a 2D tensor of shape $[16 \times 300]$ and fed into a Bi-LSTM layer. The Bi-LSTM layer captures sequential information as violence occurs as a continuous action, and temporal information needs to be captured. It is well-suited for this task as it can handle variable-length sequences, capture long-term dependencies, avoid vanishing gradients, and extract higher-level features. The Bi-LSTM layer in this architecture has 3 recurrent layers and 256 features in the hidden state. The output of the Bi-LSTM layer is a 40×256 tensor. Finally, a neural network layer with 128 neurons is applied in a time-distributed manner to create 16×128 vectors, which are then sent to the fusion model.

3D ConvNET: Another copy of the input tensor is needed as the stacked frames play the role of the height of the input of this network. A dimensional modification of the input tensor is necessary for this purpose. A 3D convolution is performed on the input tensor to capture both the temporal and spatial characteristics. The time (frame), height, width, and color channel are all convolved in four dimensions via this 3D-Convolution. The convolution operation can be represented as:

$$Y_{i,j,k} = \sum_{u=-1}^{1} \sum_{v=-1}^{1} \sum_{w=-1}^{1} X_{i-u,j-v,k-w} K_{u,v,w} \tag{10}$$

The network has a total of 3 fully connected layers, 6 pooling layers, and 6 convolutional layers. The 3D convolutional kernel size is set at $(3, 3, 3)$. From convolutional layers 1 to 6, there are 32, 64, 128, 256, 512, and 1024 kernels, respectively. In the beginning, the input tensor is transformed into a 32-dimensional feature map using the first $3 \times 3 \times 3$ convolution module. These feature maps are sent to a ReLU activation function to bring nonlinearity, which can be represented as:

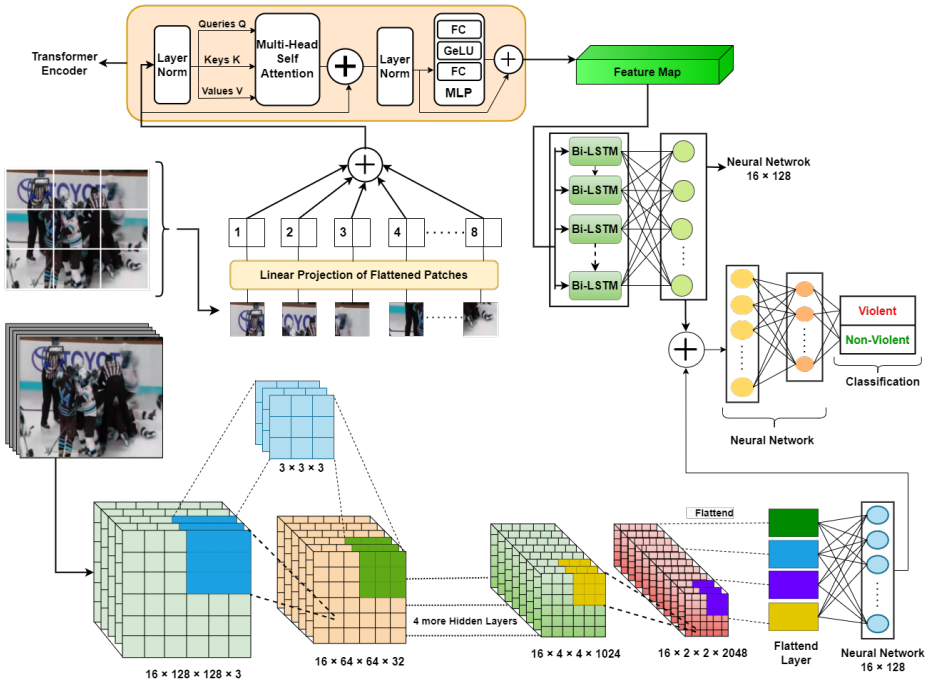


Fig. 1: VioNET: Spatiotemporal Feature Extraction for Violence Detection using a 3D Convolutional Neural Network and Vision Transformer with LSTM

$$ReLU(x) = \max(0, x) \tag{11}$$

The size of the feature maps is cut in half by passing them through a max pooling layer with a $2 \times 2 \times 2$ kernel and $(2, 2, 2)$ stride after the activation function has been applied. The pooling operation can be represented as:

$$Y_{i,j,k} = \max_{u,v,w} X_{2i+u,2j+v,2k+w} \tag{12}$$

The second to sixth convolutional layers have the same kernel and stride, compressing the feature maps but increasing their number. These feature maps are flattened and then run through various neural networks to extract the spatial and temporal characteristics from the video. After that, the fusion model receives the output feature vector.

Feature fusion: In this step, the extracted features from the vision transformer with Bi-LSTM, and 3D-Convolution models are combined and passed to a feed-forward network. The embedded tensor, with a shape of $[16 \times 300]$, is flattened and then processed through three fully connected neural networks. The first two layers have 128 and 64 neurons, respectively, while the last layer has 2 neurons

for classifying into violent or non-violent categories. Nonlinearity is introduced by applying an activation function after each layer. In our situation, though, the last layer was sigmoid.

In conclusion, our proposed model, VioNet combined both the features of a 3D-CNN model and a vision transformer with bi-Lstm. By combining the features of both models, VioNet can capture both spatial and temporal information, leading to a more robust and accurate violence detection system. This also allows the model to generalize better across different video datasets, as it can adapt to different types of violence and variations in video quality.

4 Result And Discussion

4.1 Dataset

In this study, we used three different datasets: Hockey Fight, Violent-Flows, and Movies to assess how well our model performed. Each dataset consists of a different assortment of videos that vary in terms of the recording medium, video caliber, and video duration.

- **Hockey Fight [15]:** This dataset contains 1000 videos, 500 of which show hockey players engaging in physical altercations, while the remaining 500 depict non-violent gameplay. These videos are recorded during actual professional hockey games and feature players engaged in close-body interactions. All these videos are noisy so it is often misclassified. The duration of the videos of this dataset is about 4-5 seconds.
- **Violent-Flow [16]:** This dataset contains 2,000 video clips, 1,000 of which show violent content and the other 1,000 showing non-violent content. The entire video archive contains footage of actual crowd violence. There are a lot of variants in video size so the videos must be preprocessed before fitting into a model. Most of them are between 10-15 seconds.
- **Movies Fight [17]:** There are 200 video clips in the Movies dataset - 100 are violent, and 100 are not. While the non-violent clips show typical movie scenes, the violent clips show fight scenes from movies. The common duration of the videos in the movie dataset is typically 3-4 seconds. The violent videos contain higher movement of objects than the non-violent videos.

4.2 Implementation Details

Our proposed model, VioNET was implemented using PyTorch and trained on a NVIDIA RTX 3080 Ti. We used 0.005 as learning rate and trained the model for 30 epochs and the batch size is 16. Data augmentation techniques, such as random cropping and flipping, rotating, to prevent overfitting and improve generalization.

The hyperparameters of the proposed model has been fine-tuned using a trial and error approach. By systematically exploring various combinations of hyperparameter values and evaluating the model's performance, we determined

the optimal configuration for our violence detection model, as presented in the table 1. The hyperparameters include the input shape of vision transformer and 3D convolution, patch size of the vision transformer, the dimensionality of the token embeddings and the positional embeddings, number of transformer blocks in the model, and total number of attention heads in each multi-head attention layer. Additionally, the number of Bi-LSTM layers and hidden state size are also given to show what parameters used for best fit. Finally, convolutional kernel size, number of 3D convolutional layers, pooling layers follows them, fully connected layers after flattening, learning rate, optimizer, batch size, and number of epochs are also defined.

Table 1: Hyperparameters for the proposed model

Hyperparameters	**Values**
Dimension of patches in Vision transformer	16×16 pixels
Embedding dimension of vision transformer	2048
Number of transformer blocks in Vision Transformer	16
Number of attention heads	16
Number of Bi-LSTM layers	3
Bi-LSTM hidden state size	256
Number of neurons in fully connected layer	128
Convolution kernel size	$3 \times 3 \times 3$
Number of 3D convolutional layers	6
Number of pooling layers	6
Number of fully connected layers in 3D ConvNet	3
Number of fully connected layers in fusion model	3
Learning rate	0.005
Optimizer	Adam
Loss function	Cross-entropy
Batch size	16
Number of epochs	30

4.3 Performance Evaluation

Performance evaluation of VioNet was assessed on the mentioned three datasets: Hockey Fight, Movies Fight, and Violent Flow. Accuracy, precision, recall, and F1-score were used as evaluation criteria. Here, $NV_{correct}$: Number of videos accurately identified as non-violent, $V_{correct}$: Number of videos accurately identified as violent, $NV_{incorrect}$: Number of videos misclassified as non-violent, $V_{incorrect}$: Number of videos misclassified as violent, and $Videos_{total}$: Total number of videos for evaluation. So, the equations of those evaluation criteria are given below:

$$Accuracy = \frac{NV_{correct} + V_{correct}}{Videos_{total}} \tag{13}$$

$$Precision = \frac{V_{correct}}{V_{correct} + V_{incorrect}} \tag{14}$$

$$Recall = \frac{V_{correct}}{V_{correct} + NV_{incorrect}} \tag{15}$$

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \tag{16}$$

For each dataset, we also calculated the precision, recall, and F1 score in order to thoroughly assess the performance of our model. The evaluation results shown in Table 2 demonstrate that VioNet achieved remarkable accuracies, precisions, and recalls on all three datasets, indicating that our approach surpasses numerous cutting-edge methods in the field of violence detection in videos.

Table 2: Evaluation matrices of the proposed model

Datasets	**Evaluation Metrics**			
	Accuracy	**Precision**	**Recall**	**F1-Score**
Hocky Fight	0.9785	0.9868	0.9740	0.9803
Violent Flow	0.9633	0.9701	0.9489	0.9594
Movie Fight	1.0000	1.0000	1.0000	1.0000

In Table 3, we have presented the outputs of our model by showing some extracted frames along with the comparison between the ground truth and its corresponding predicted type. During our analysis, we observed some misclassifications and tried to identify the reasons behind them. It was found that due to the variations in shape of the Violent-Flow dataset, misclassifications can occur at times. Additionally, the model may fail to classify correctly in situations with lower light, where the movements in the data are not clearly understandable.

4.4 Comparison with Other Methods

VioNet, VitBiLSTM, and 3D ConvNET models developed in this study achieve high accuracy rates when detecting violent content in Hockey, Violent, and Movies datasets. They outperform or achieve comparable accuracy rates in each of the three datasets compared to other models presented in the table. A high accuracy rate is achieved by the VioNet model in all categories, followed closely by the 3D ConvNET model and VitBiLSTM model. 1D-CNN + ResNet50, VGG16-NN Classifier, and Pre-trained CNN + LSTM models, with accuracy rates ranging from 89.5% to 96.33%, are the closest competitors to this study's models. However, these models don't achieve the same level of accuracy as VioNet across all three datasets. A HOG model achieves an accuracy rate of 86% in the Hockey dataset, but it hasn't been evaluated in the Violent-Flow or Movies datasets. As well, the Hybrid MOSIFT-BoW model achieves high accuracy rates of 90.9% and 84.2% for the Hockey and Movies datasets, respectively, but has not been

evaluated for the Violent-Flow dataset. Overall, the models developed in this study show significant improvement in detecting violent content in video data and can be a valuable addition to the field of analysis of video content. Table 4 gives a summary of all heading levels. : As the video duration in movie dataset is small, there is less chances of loss of information while extracting features from video to image frames. Movie fight scenes contain over movements than the normal scenes which are easily distinguishable to any model

Table 3: Comparative Analysis: Ground Truth vs. Prediction Type and Assessing Violence Detection Performance with Confidence Values.

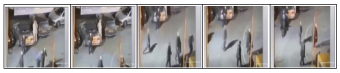
Ground Truth	Dataset	Video Frames	Confidence Value	Predicted Type
Non-Violence	Violent-Flow		0.9956	Non-Violent
Violence	Hockey Fight		0.9439	Non-Violent
Non-Violence	Violent-Flow		0.9859	Non-Violent
Violence	Violent-Flow		0.6732	Non-Violent
Violence	Movie		0.9617	Violent
Non-Violence	Hockey Fight		0.7518	Violent

Table 4: The comparision of accuracy achieved in various dataset

Methods	**Datasets**		
	Hockey(%)	**Violent Flow(%)**	**Movies(%)**
HOG [4]	86	-	-
CNN-LSTM Classifier [9]	89.5	91.4	100
Hybrid MoSIFT-BoW [1]	90.9	-	84.2
3D ConvNets [18]	91	-	-
VGG16-NN Classifier [19]	95.5	96	100
1D-CNN + ResNet50 [5]	96	96	-
Pre-trained CNN + LSTM [20]	96.33	85.71	100
3D ConvNET	**96.80**	94.00	99.00
VitBiLSTM	95.00	92.00	97.00
VioNet (VitBiLSTM + 3D ConvNET)	**97.85**	**96.33**	**100.00**

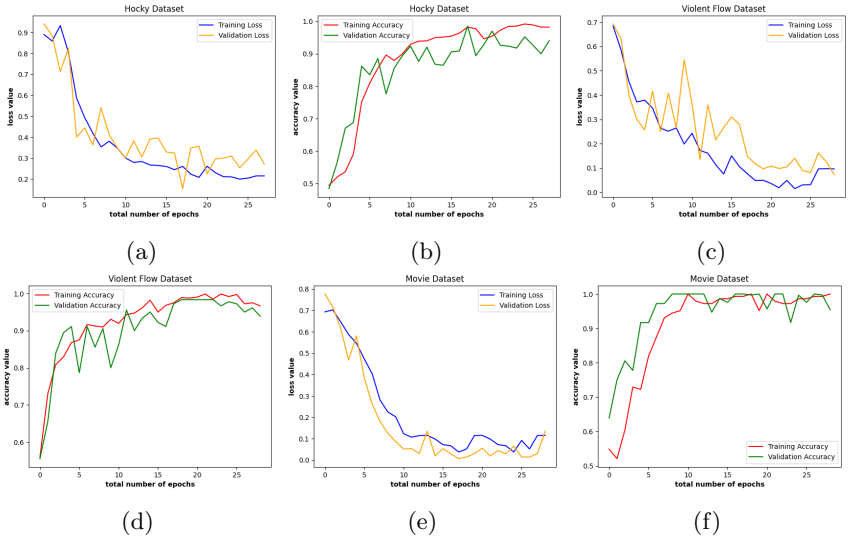


Fig. 2: Different evaluation graphs for the proposed model where {(a), (b)}, {(c), (d)}, {(e), (f)} represent the 'loss vs epochs' and 'accuracy vs epochs' curves for the hocky fight dataset, violent flow dataset and movie fight dataset respectively.

Initially, we utilized a 3D-CNN model to detect violence from the mentioned dataset, and we observed that adding more convolution and pooling layers improved its performance. However, we encountered the problem of overfitting when using more than 10 layers. We then explored another approach, which involved using a vision transformer with bi-LSTM, a new method in violence detection. This approach showed higher accuracy, albeit slightly less than 3D-CNN. Eventually, we combined the extracted features from both 3D-CNN and vision transformer with bi-LSTM and fed them into a neural network. This approach outperformed all other renowned techniques and achieved the best accuracy.

5 Conclusion

In this study, we present VioNET, a violence detection system that combines two of the most efficient deep learning approaches: Vision Transformer with Bi-LSTM and 3D-Convolutional Neural Networks. Our approach involves extracting image features from the same input tensor using both methods and classifying the video using the extracted feature. This allows us to extract spatiotemporal features more accurately than existing methods, making our model more robust and faster. However, we acknowledge that there is still significant room for improvement in detecting violence in video. In our future work, we intend to embed an audio feature extractor into our model. By combining the information from both audio and video modalities, the model can make more accurate predictions

about the presence of violence in a video specially when video quality is poor or the visual cues are not clear.

References

1. Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R.: Violence detection in video using computer vision techniques. In: Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14, pp. 332–339. Springer (2011)
2. De Souza, F.D., Chavez, G.C., do Valle Jr, E.A., Araújo, A.d.A.: Violence detection in video using spatio-temporal features. In: 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images, pp. 224–230. IEEE (2010)
3. Zhou, P., Ding, Q., Luo, H., Hou, X.: Violence detection in surveillance video using low-level features. PLoS one **13**(10), e0203,668 (2018)
4. Das, S., Sarker, A., Mahmud, T.: Violence detection from videos using hog features. In: 2019 4th International Conference on Electrical Information and Communication Technology (EICT), pp. 1–5. IEEE (2019)
5. Honarjoo, N., Abdari, A., Mansouri, A.: Violence detection using one-dimensional convolutional networks. In: 2021 12th International Conference on Information and Knowledge Technology (IKT), pp. 188–191 (2021). https://doi.org/10.1109/IKT54664.2021.9685835
6. Abdali, A.R.: Data efficient video transformer for violence detection. In: 2021 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), pp. 195–199 (2021). https://doi.org/10.1109/COMNETSAT53002.2021.9530829
7. Su, M., Zhang, C., Tong, Y., Liang, B., Ma, S., Wang, J.: Deep learning in video violence detection. In: 2021 International Conference on Computer Technology and Media Convergence Design (CTMCD), pp. 268–272 (2021). https://doi.org/10.1109/CTMCD53128.2021.00064
8. Mugunga, I., Dong, J., Rigall, E., Guo, S., Madessa, A.H., Nawaz, H.S.: A frame-based feature model for violence detection from surveillance cameras using convlstm network. In: 2021 6th International Conference on Image, Vision and Computing (ICIVC), pp. 55–60. IEEE (2021)
9. Patel, M.: Real-time violence detection using cnn-lstm. arXiv preprint arXiv:2107.07578 (2021)
10. Song, W., Zhang, D., Zhao, X., Yu, J., Zheng, R., Wang, A.: A novel violent video detection scheme based on modified 3d convolutional neural networks. IEEE Access **7**, 39,172–39,179 (2019)
11. Sudhakaran, S., Lanz, O.: Learning to detect violent videos using convolutional long short-term memory. In: 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS), pp. 1–6. IEEE (2017)
12. Peixoto, B., Lavi, B., Bestagini, P., Dias, Z., Rocha, A.: Multimodal violence detection in videos. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2957–2961 (2020). https://doi.org/10.1109/ICASSP40776.2020.9054018
13. Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., Theodoridis, S.: Audio-visual fusion for detecting violent scenes in videos. In: Artificial Intelligence: Theories, Models and Applications: 6th Hellenic Conference on AI, SETN 2010, Athens, Greece, May 4-7, 2010. Proceedings 6, pp. 91–100. Springer (2010)

14. Peixoto, B.M., Lavi, B., Dias, Z., Rocha, A.: Harnessing high-level concepts, visual, and auditory features for violence detection in videos. Journal of Visual Communication and Image Representation **78**, 103,174 (2021)
15. Hockey fight vidoes dataset. `https://www.kaggle.com/datasets/yassershrief/hockey-fight-vidoes`. last accessed 2022/05/15
16. Real life violence situations dataset. `https://www.kaggle.com/datasets/mohamedmustafa/real-life-violence-situations-dataset`. last accessed 2022/05/15
17. Movie fights dataset. `https://www.kaggle.com/datasets/frendon/moviefights`. last accessed 2022/05/15
18. Ding, C., Fan, S., Zhu, M., Feng, W., Jia, B.: Violence detection in video by using 3d convolutional neural networks. In: Advances in Visual Computing: 10th International Symposium, ISVC 2014, Las Vegas, NV, USA, December 8-10, 2014, Proceedings, Part II 10, pp. 551–558. Springer (2014)
19. Honarjoo, N., Abdari, A., Mansouri, A.: Violence detection using pre-trained models. In: 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA), pp. 1–4. IEEE (2021)
20. Abdali, A.M.R., Al-Tuma, R.F.: Robust real-time violence detection in video using cnn and lstm. In: 2019 2nd Scientific Conference of Computer Sciences (SCCS), pp. 104–108. IEEE (2019)