BrutNet: A Novel Approach for Violence Detection and Classification using DCNN with GRU

Mahmudul Haque[1], Hussain Nyeem[2], and Syma Afsha[3]

[1]Independent University
[2]Military Institute of Science and Technology
[3]Universitat de Girona

April 04, 2024

Abstract

Automatic Violence Detection and Classification (AVDC) with deep learning has garnered significant attention in computer vision research. This paper presents a novel approach to combining a custom Deep Convolutional Neural Network (DCNN) with a Gated Recurrent Unit (GRU) in developing a new AVDC model called BrutNet. Specifically, we develop a time-distributed DCNN (TD-DCNN) to generate a compact 2D representation with 512 spatial features per frame from a set of equally-spaced frames of dimension 160×90 in short video segments. Further to leverage the temporal information, a GRU layer is utilised, generating a condensed 1D vector that enables binary classification of violent or non-violent content through multiple dense layers. Overfitting is addressed by incorporating dropout layers with a rate of 0.5, while the hidden and output layers employ rectified linear unit (ReLU) and sigmoid activations, respectively. The model is trained on the NVIDIA Tesla K80 GPU through Google Colab, demonstrating superior performance compared to existing models across various video datasets, including hockey fights, movie fights, AVD, and RWF-2000. Notably, our model stands out by requiring only 3.416 million parameters and achieving impressive test accuracies of 97.62%, 100%, 97.22%, and 86.43% on the respective datasets. Thus, BrutNet exhibits the potential to emerge as a highly efficient and robust AVDC model in support of greater public safety, content moderation and censorship, computer-aided investigations, and law enforcement.

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ORIGINAL RESEARCH PAPER

BrutNet: A Novel Approach for Violence Detection and Classification using DCNN with GRU

Mahmudul Haque[1] | **Hussain Nyeem**[2] | **Syma Afsha**[3]

[1]Department of Computer Science & Engineering (CSE), Independent University, Bangladesh (IUB), Dhaka, Bangladesh

[2]Department of Electrical, Electronic & Communication Engineering (EECE), Military Institute of Science & Technology (MIST), Dhaka, Bangladesh

[3]Erusmus Mundus Joint Master in Intelligent Field Robotics System (IFRoS), Universitat de Girona, Girona–17003, Spain

Correspondence
Hussain Nyeem, Department of EECE, MIST.
Email: h.nyeem@eece.mist.ac.bd

Present address
Mirpur Cantonment, Dhaka-1216, Bangladesh.

Abstract

Automatic Violence Detection and Classification (AVDC) with deep learning has garnered significant attention in computer vision research. This paper presents a novel approach to combining a custom Deep Convolutional Neural Network (DCNN) with a Gated Recurrent Unit (GRU) in developing a new AVDC model called BrutNet. Specifically, we develop a time-distributed DCNN (TD-DCNN) to generate a compact 2D representation with 512 spatial features per frame from a set of equally-spaced frames of dimension 160×90 in short video segments. Further to leverage the temporal information, a GRU layer is utilised, generating a condensed 1D vector that enables binary classification of violent or non-violent content through multiple dense layers. Overfitting is addressed by incorporating dropout layers with a rate of 0.5, while the hidden and output layers employ rectified linear unit (ReLU) and sigmoid activations, respectively. The model is trained on the NVIDIA Tesla K80 GPU through Google Colab, demonstrating superior performance compared to existing models across various video datasets, including hockey fights, movie fights, AVD, and RWF-2000. Notably, our model stands out by requiring only 3.416 million parameters and achieving impressive test accuracies of 97.62%, 100%, 97.22%, and 86.43% on the respective datasets. Thus, BrutNet exhibits the potential to emerge as a highly efficient and robust AVDC model in support of greater public safety, content moderation and censorship, computer-aided investigations, and law enforcement.

KEYWORDS
BrutNet, DCNN, GRU, violence detection, activity recognition, spatio-temporal model

1 | INTRODUCTION

Automatic Violence Detection and Classification (AVDC) of video content has recently emerged as a critical problem[1,2], leading to an increased focus on using Deep Learning (DL), computer vision, and image processing methods for flagging unsuitable content and detecting violent behaviour[3]. Automatic and real-time video analysis is imperative to identifying violent offenders and maintaining the safety of cities[4,5]. Detecting violent behaviour is also a critical requirement in video surveillance applications for facilities like prisons, mental or elderly care facilities, and camera phones[6]. Besides, as visual materials become more abundant on online platforms such as YouTube, Facebook, Twitter, and Netflix, their open availability and the lack of a suitable monitoring or certification body also demand an automated classification of sensitive content[7,8,9].

However, accurately characterizing violence in multi-modal video content is a formidable task, yet essential for the development of an AVDC system. Violence in video content is mainly characterized by conflict aggressiveness, damaging conduct, psychological instability, and the deliberate use of physical strength or force against another person or group[10]. These diverse attributes necessitate the spatial and temporal analysis of the content to identify violent or hostile behaviour. Furthermore, given the broad range of events and activities that can be captured in the multi-modal video, such as surveillance at correctional and elderly-care facilities, traffic, sports, and social and online media, detecting violent behaviour in videos has proven to be a challenging endeavour[11]. Addressing this challenge has led to the development of various methods using DL and Support Vector Machines (SVMs).

As a promising DL approach, DCNNs are recently used to identify violent situations[12,13]. While DCNN-based models exhibit impressive image categorization and object recognition abilities, they may encounter computational inefficiencies

when applied to video analysis (see Sec. 2). Particularly, such a model can analyse only one video frame at a time and cannot identify visual data in a time series [14]. To overcome this limitation, sequential learning models such as RNN and LSTM are used [15]. LSTM, which contains three gates (input, output, and forget gates) and a memory cell that stores previous sequence information, is computationally expensive, which makes the activity recognition system inefficient. In contrast, GRU, a simpler RNN variant with two reset and update gates but no memory cell, has a better ability to learn long-term sequences.

In this paper, a novel approach is proposed to utilize time-distributed DCNN (TD-DCNN) with GRU in developing a new AVDC system. Specifically, we developed BrutNet, a customized TD-DCNN model combined with GRU, utilise their strengths for AVDC with greater accuracy and efficiency. BrutNet is designed to operate on video segments of 3-5 seconds having a resolution of 160×90. We consider reducing the number of convolutional layers while increasing the number of hidden layers, to minimize the overall number of model parameters without making it susceptible to a vanishing gradient. The model has been trained on the AVD dataset [16], and additionally tested on the hockey-fights (HF) [6], movie-fights (MF) [6] and RWF2000 [17] datasets to investigate its ability to accurately classify violent and non-violent videos for various real-time applications.

This paper provides a detailed account of the procedures used, implementation strategies, and performance analysis for the development of an AVDC system. In Sec. 2, we outline the procedures and datasets used in other related works, followed by the architecture and implementation details of the proposed model, BrutNet in Sec. 3. The experiment settings including dataset pre-processing and model evaluation approaches are discussed in Sec. 4, while Sec. 5 presents the results and analysis demonstrating the violence detection and classification performance of BrutNet. Finally, Sec. 6 highlights our concluding remarks, summarizing the outcomes and significance of our work, and potential avenues for future research.

2 | RELATED WORK

AVDC models are primarily developed for different scenarios such as video surveillance, movie content detection, and traffic monitoring [18,10,19]. Irrespective of the scenario, DL based approaches have recently shown great potential in violence characterization using multiprocessing layer models, resulting in significant improvements in AVDC through human activity identification, image or video-based object or pattern recognition, anomaly detection, and emotion detection [20,21,22]. There are several approaches for addressing these problems, and we will analyse these models and datasets in this section that are primarily developed for violence characterization.

2.1 | AVDC Models

Developing DL models requires several general considerations, such as selecting the appropriate dataset for the target application, selecting the appropriate neural network architecture, and optimizing the model's hyperparameters. The chronological development of DL models for AVDC can be examined based on different factors, such as spatial features, temporal features, or hybrid approaches. Spatial feature based models typically handle image recognition tasks, whereas temporal feature-based models handle tasks involving sequential data. To address complex tasks like activity recognition, hybrid approaches have emerged to combine spatial and temporal features. To develop our new BrutNet model, we studied these approaches based on the features considered and identified gaps in the state-of-the-art approaches for AVDC to address them in this paper.

2.1.1 | Spatial Feature-based Models

Most of the approaches that deal with the spatial features of the dataset are DCNN based. DCNNs have significantly impacted various domains of pattern recognition [23,24]. For instance, Das *et al.* [25] presented a technique for recognizing violence that involves selecting several frames from each video segment using image removal and averaging and extracting lower-level features using HOG. Subsequently, they utilized SVM, LDA, Naive Bayes, and K-Nearest Neighbors (KNN) for classification. Besides, Wang *et al.* [21] discussed the different CNNs employed for video violence detection, their advantages, and their drawbacks. The technique suggested by Honarjoo *et al.* [26] involves utilizing pre-trained deep neural networks, specifically ResNet-50 and VGG16, to identify violent actions using extracted features from pre-trained models, providing a technique with a minimal level of complexity for identifying instances of violence.

Guedes *et al.* [27] proposed a CNN and SVM classifier-based strategy for identifying instances of aggressive behaviour in video streams containing violent altercations. To identify and locate violent activities in video surveillance, Roman *et al.* [28] used dynamic pictures to categorize a video as violent or non-violent and then used CNNs and weakly supervised localization algorithms to locate violent regions in HF, Violent Flows and UCFCrime2Local datasets.

Similarly, Gruosso *et al.* [29] developed a content grading system based on CNN for evaluating materials for children, teens, and adults. They also created an algorithm to categorize and restrict violent situations automatically. To train and verify the Inception v3 architectural model, they utilized a large hand-labelled dataset containing visual components useful for categorization. For model evaluation, they created an algorithm to enhance the network performance for video input.

2.1.2 | Temporal Feature-based Models

A few approaches have considered the temporal property of the video datasets. Wang *et al.*[30] demonstrated that DL approaches based on RNNs, such as LSTM and GRU, outperform other sequential models when used to predict traffic flow. Cheng *et al.*[17] introduced the RWF-2000 database, which contains 2,000 films recorded by security cameras in real-world scenarios and utilizes 3D-CNN and optical flow. The model employed self-learned pooling to adapt to both appearance and temporal features.

According to Chatterjee *et al.*[31], the purpose of their study was to improve the categorization of violent and non-violent actions in public settings. They employed a convolutional bidirectional LSTM to identify violent activities, and the results were compared to other techniques, demonstrating a higher classification accuracy for the popular HF dataset.

2.1.3 | Spatio-temporal Feature-based Models

To address the issue of activity detection considering both spatial and temporal features, Ullah *et al.*[15] proposed a DL model. They employed a CNN network trained on two surveillance datasets to detect a person in a surveillance stream initially. They used an Efficient LiteFlowNet CNN and Deep Skip Connection GRU (DS-GRU) based approach to learn the spatio-temporal variations in a frame sequence for activity detection.

Besides, Peixoto *et al.*[32] proposed using two deep neural networks (DNN) frameworks, C3D and CNN-LSTM, to detect violence in movies. The frameworks were applied to learn spatio-temporal information from video segments under subjective and conceptual scenarios. The fusion of ideas was analysed as a whole to determine the higher-level notion of violence. Besides, Ditsanthia *et al.*[33] also suggested a DL-based video-based AVD. They analysed the findings of numerous AVD approaches and found ResNet50+LSTM to be the most accurate on the common datasets like HF, MF and real-violent.

Recently, Vijeikis *et al.*[34] time-distributed MobileNetV2 and LSTM-based model for addressing an efficient violence detection problem. For higher classification accuracy, Ehasan *et al.*[35] proposed an UNet + PatchGAN-based unsupervised action translation network utilizing spatio-temporal features to identify violent behaviours and overcome the problem related to the insufficiency of relevant data. Similarly, Mohtavipour *et al.*[36] proposed a multi-stream CNN-based AVDC approach. Despite the promising classification performance of this model, its computational efficiency in terms of total parameters remains suboptimal.

While studying all these approaches based on spatial and/or temporal features, several other models like Efficient 3D CNN[37], Xception + BiLSTM + Attentions[38], C3D[39], AlexNet + LSTM[40], Hough Forests + 2D CNN[41], Three Streams + LSTM[42], MoSIFT[43], motion intensities + AdaBoost[44], ResNet50 + ConvLSTM[45], Fine-tuned MobileNet[46] and Motion Blobs + Random Forest[47] and Double-AE[48] have also attempted to address the violence detection problems and demonstrated notable results. Hence, to better understand the progress of AVDC so far, the prominent approaches discussed above have been summarized in the Table. 1.

2.2 | Datasets

Most of the datasets used for violence identification are made up of several video segments, have poor resolution, and are often constructed on close-context examples, resulting in high false positives from inadequate learning. So, it is crucial to use high-resolution datasets to assess the resilience of violence detection systems against false positives[16]. Hence, we considered several datasets that can overcome such limitations and help train a more generalized model for AVDC. Some of these datasets have been discussed in this section.

AVD Dataset. The AVD dataset is a recent and valuable resource for studying violence detection with high-resolution videos. The dataset consists of 350 video segments, each captured at a resolution of 1920×1080 and a frame rate of 30fps. These video segments were categorized and labelled as either *violent* or *non-violent*, forming the foundation of the AVD dataset. The segment durations varied, ranging from 75 to 435 frames. To ensure unbiased evaluation, 80% (222 segments) were assigned to the training dataset, while the remaining 20% (56 segments) were allocated to the validation dataset. Notably, the dataset exhibited an inherent class imbalance between *violent* and *non-violent* samples. The number of samples in the violent class was approximately twice that of the non-violent class. Consequently, addressing this variation during training requires tailoring the weighted loss function within the proposed BrutNet architecture.

HF Dataset. It is a popular and widely used dataset for violence detection. The videos mainly comprise different scenes in several hockey matches. The dataset comprises a total of 1000 video segments. Each video segment of the dataset had between 41 and 50 frames. These video segments were of two types, 50% were with fighting scenes and 50% did not contain any fighting scenes. The video segments with fight scenes were considered to be violent scenes and the rest were considered non-violent, hence, were labelled 1 and 0 accordingly. There were 500 video segments for each type of sample.

MF Dataset. It is another popular dataset for violent content detection. It contains a total of 200 video segments. These

TABLE 1 Summary of prominent approaches.

Model	Target Feature	Limitation
ResNet50 + LSTM [33]	spatio-temporal	Limitted to aggressive behaviour detection
Efficient 3D [37]	spatio-temporal	Time-consuming on low-end devices
Multi-stream CNN [36]	spatio-temporal	Large number of model parameters
Hough Forests +2D CNN [41]	spatial	Computationally expensive
Xception + BiLSTM + Attentions [38]	spatio-temporal	Limitted to fighting scenario detection
Motion Blobs + Random Forest [47]	spatio-temporal	Low accuracy Poor detection of continuous movement
C3D [39]	spatio-temporal	Low accuracy & computationally expensive
AlexNet + LSTM [40]	spatio-temporal	Computationally expensive
Motion intensities + AdaBoost [44]	spatio-temporal	Poor detection of aggressive behaviours
ResNet50 + ConvLSTM [45]	spatio-temporal	Low accuracy
UNet + PatchGAN [35]	spatio-temporal	Low accuracy
MobileNetV2 + LSTM [34]	temporal	Lack of diversity in used datasets
VGG16 [26]	spatial	Disregarded the sequential nature of activities in video frames
DWT-CNN-BiLSTM [31]	temporal	Lack of diversity in used datasets

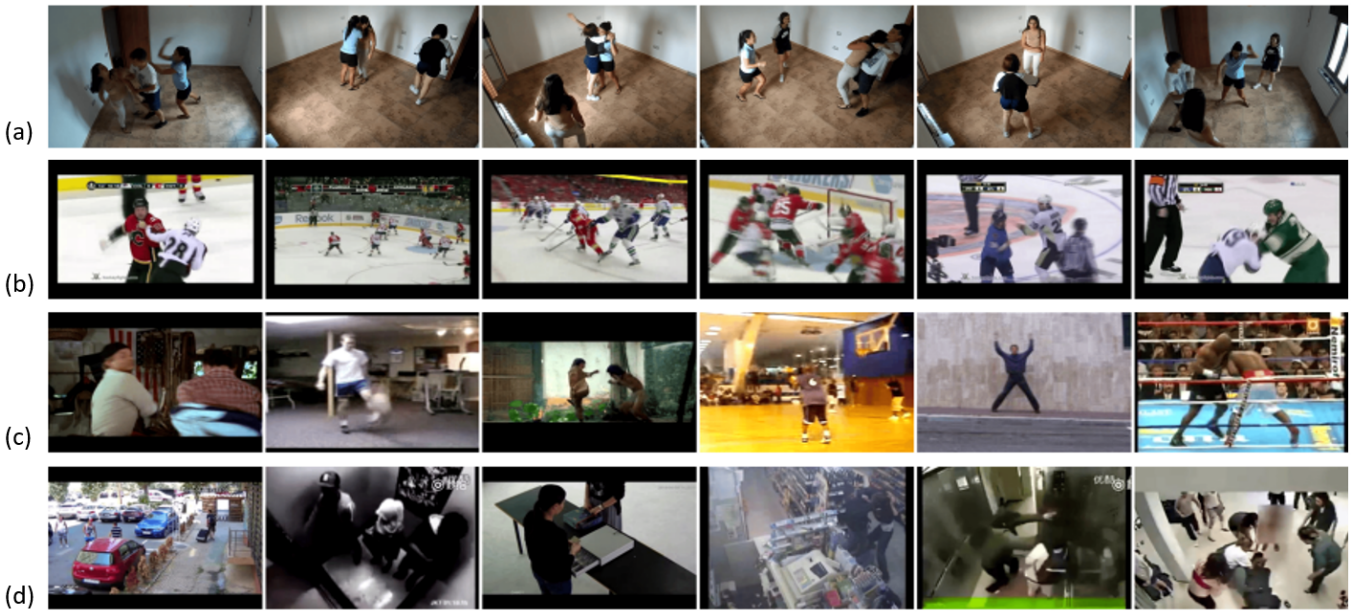

FIGURE 1 Samples of different datasets: (a) AVD Dataset, (b) HF Dataset, (c) MF Dataset and (d) RWF-2000 Dataset

segments are divided into two classes, *fights* and *non-fights*, each class containing 100 segments. The fight scenes were considered to be violent scenes and the non-fights scenes were considered non-violent, hence, were labelled 1 and 0 accordingly.

RWF-2000 Dataset. It is another dataset similar to HF and MF datasets with 2000 samples, each containing 151 frames with highly varying resolutions and aspect ratios. The total dataset is labelled into two classes, *fight* and *non-fight* each with 1000 samples. The *fight* classes are considered violent classes and *non-fight* classes are considered non-violent classes.

2.3 | Scope of Development

The preceding discussion highlights an ample scope for enhancing the current AVDC approaches in terms of accuracy and computational efficiency. To this end, we initially attempted to develop BrutNet by combining a time-distributed DCNN

and GRU-based architecture, and our early results presented in[49] confirmed that it can reliably recognize and categorize the visual content depending on the presence of violent actions. Being trained on the AVD video dataset, that earlier approach demonstrated a fairly high test accuracy of 90%.

However, a significant challenge in developing an efficient AVDC system lies in striking a balance between performance and the number of parameters in the model. Existing models that focus solely on either spatial or temporal features of video datasets have fewer parameters but tend to exhibit lower accuracy. Conversely, models that incorporate hybrid (*i.e.*, spatio-temporal) features tend to achieve higher accuracy but at the expense of increased parameters, resulting in a computationally heavy architecture.

Addressing the said trade-off necessitates redefining the earlier architecture of BrutNet[49] by reducing the number of convolution layers while increasing the number of hidden layers, and incorporating additional data pre-processing steps. This approach aims to decrease the overall number of model parameters without rendering it susceptible to the vanishing gradient problem. As a result, the proposed new architecture of BrutNet is expected to enhance both computational efficiency and classification performance.

Furnishing with essential technical details of the new model (Sec. 3), diverse training and testing using the widely used AVD[16], HF[6], MF[6] and RWF2000[17] datasets, and thorough validation to demonstrate its anticipated improvement against the other prominent models (Sec. 5), this paper aims to address the aforementioned challenges and contribute to the advancement of AVDC systems.

3 | A NEW BRUTNET MODEL

Development of an efficient AVDC system requires considering both the spatial and temporal features of each video segment of the considered datasets, while ensuring the number of parameters is as low as possible to make our model computationally efficient. We thus have designed BrutNet to incorporate the spatio-temporal features of the wide-ranging violent scenes. The primary processing phases of this system are depicted in Fig. 2. In what follows, we present the architecture of the proposed model (Fig. 3) that combines a custom TD-DCNN, GRU and dense layers for AVDC. We also provide an essential mathematical foundation for this development and specified the parameters to highlight its computational requirements.

3.1 | Architecture

As an AVDC model, BrutNet analyses pre-processed video segments as a set of spatio-temporal data. Given that videos consist of a sequence of images, it becomes crucial to capture the temporal characteristics of specific features present in these images to effectively detect activities within the video. In particular, after undergoing the pre-processing steps outlined in section 4.1, each individual sample from the dataset, namely video segments with dimensions (24, 90, 160, 3), is fed into the BrutNet network. This fresh approach allows the model for comprehensive analysis and classification of the data.

Moreover, the proposed model explicitly adopts a set of sequential layers, enabling the recognition of patterns across multiple frames. While DCNN-based AVDC models have already demonstrated their potential for image and video-based detection and classification, they fail to address the temporal nature of any action. Hence, for the spatio-temporal approach, DCNN cannot be used directly for the surveillance and identification of violent content because it operates on static images to learn its properties. Determining the intrinsic characteristics of a violent incident in a video requires analysis of the sequential patterns from successive image sets or frames. This indicates that sequential models, such as GRU, are more capable of recognizing the pattern between multiple frames of a video segment to recognize violent activities. Our model consequently combines custom TD-DCNN and GRU-based RNN to optimize the learning of violent video characteristics, followed by dense layers that further refines the learned features. In the subsequent sections, we provide a detailed technical explanation of these components and their integration within our model.

3.1.1 | TD-DCNN Layers

As illustrated in Fig. 3, BrutNet is ideally designed to initialize with a set of convolutional layers defined in Eq. (1) for each frame of the time-distributed layer. Considering $\mathbf{W}$, X and $\mathbf{C}$ to be the kernel parameters, the input image, and the output of each convolutional layer ($C_0 = X$), respectively with m and n being the number of rows and columns of the input image. Here, $\mathbf{R}$ is the ReLU activation function such that $R(x) = \max(0, x)$.

$$\mathbf{Y}_L = \sum_{i=0}^{m} \sum_{j=0}^{n} \mathbf{C}_{L-1}(m-i, n-j) \cdot \mathbf{W}_L(i,j) \tag{1a}$$

$$\mathbf{C}_L(m,n) = \mathbf{R}(Y_L) \tag{1b}$$

After every two convolutional layers, batch normalization and MaxPooling layers were used. Batch normalization was used to normalize the activations of each layer, leading to improved training speed and generalization using equation (2a) such that μ is the mean, σ^2 is the variance and ϵ is the constant used for numerical stability. During training, the running means and variance is updated using a momentum (ρ) value

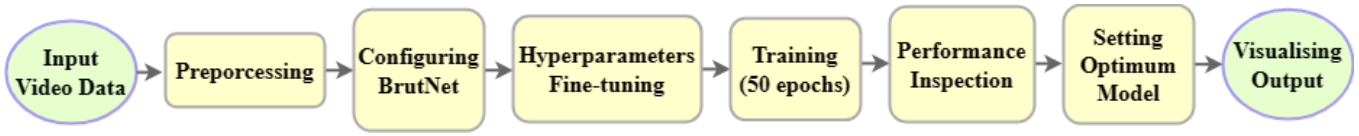

FIGURE 2 Processes of the proposed AVDC system.

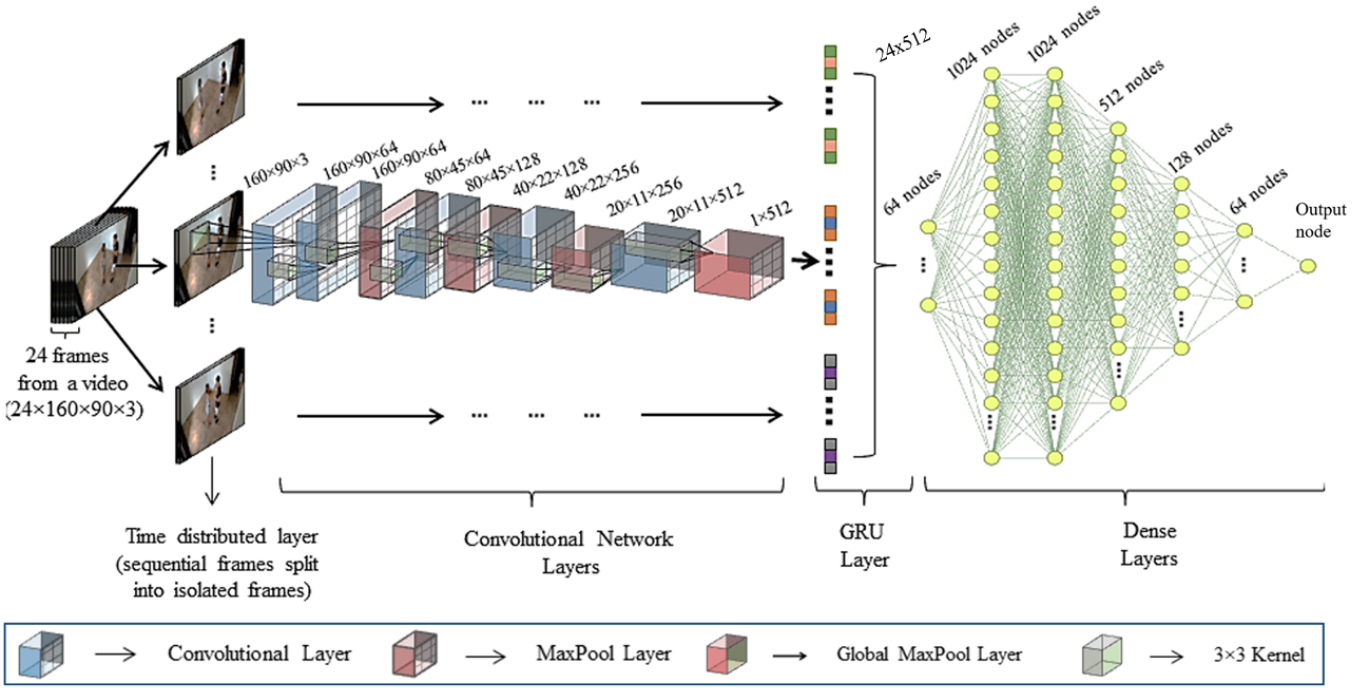

FIGURE 3 Architecture of BrutNet

using Eq. (2b) and (2c).

$$\hat{\mathbf{B}} = \frac{\mathbf{C}_L - \mu_{batch}}{\sqrt{\sigma^2_{batch} + \epsilon)}} \tag{2a}$$

$$\mu_{running} = (1 - \rho) \cdot \mu_{running} + \rho \cdot \mu_{batch} \tag{2b}$$

$$\sigma^2_{running} = (1 - \rho) \cdot \sigma^2_{running} + \rho \cdot \sigma^2_{batch} \tag{2c}$$

On the other hand, to downsample feature maps and extract the most salient information in the local regions of the input matrix, the MaxPooling2D operation is carried out with a specified pool size (p_H, p_W) and stride (s_H, s_W) as defined in Eq. (3). Here, $\mathbf{P}$ is the output matrix, and n, c, h, and w denote the batch index, channel index, height index, and width index, respectively. This process helps determine the maximum value within each pool region, considering the specified stride to shift the pooling window.

$$\mathbf{P} = \max_{i=0}^{p_H-1} \max_{j=0}^{p_W-1} \left(\hat{\mathbf{B}}_{n,c,(h \cdot s_H+i),(w \cdot s_W+j)} \right) \tag{3}$$

Passing through the series of convolution layers, as defined in Eq. (4), the model encodes each frame into a vector of 512 spatial features, $\mathbf{G}$ using the Global Max Pooling operation for capturing the most salient features across all spatial locations in the input matrix.

$$\mathbf{G}_{n,c} = \max_{h=1}^{H} \max_{w=1}^{W} \left(\mathbf{X}_{n,c,h,w} \right) \tag{4}$$

Combining the $\mathbf{G}_{n,c}$ values of the 24 frames of the input video segment, the $\mathbf{C_t}$ matrix of shape (512, 24) is obtained that captures the inter-frames correlations of the video segment.

3.1.2 | GRU Layer

To explore the temporal features or patterns across the given 24 frames, we introduced an RNN-based GRU layer in the BrutNet architecture. Thus, the matrix of spatial features, C_t, obtained from the preceding TD-DCNN layers, is utilized by the GRU layer designed in this stage. The GRU layer identifies the time-dependent sequential pattern within these spatial feature sets from the TD-CNN using the set of equations in (5), where z and r are the update and reset gate activations, $\tilde{h}$ and h are the current and updated hidden states, respectively, with W and U being the weights of each stage, resulting in an output of a 1D vector that represents the extracted spatio-temporal features. The *the sigmoid* activation function, σ in Eq. (6) is used to normalize the feature sets.

$$\mathbf{z}(t) = \mathbf{S}(C_t(t) \cdot W_z + h(t-1) \cdot U_z) \tag{5a}$$

$$\mathbf{r}(t) = \mathbf{S}(C_t(t) \cdot W_r + h(t-1) \cdot U_r) \tag{5b}$$

$$\tilde{\mathbf{h}}(t) = \tanh\left(C_t(t) \cdot W_h + r \odot h(t-1) \cdot U_h\right) \tag{5c}$$

$$\mathbf{h}(t) = z \odot h(t-1) + (1-z) \odot \tilde{h}(t) \tag{5d}$$

$$\mathbf{S}(x) = \frac{1}{1+e^{-x}} \tag{6}$$

3.1.3 | Dense Layers

For enabling the BrutNet model to determine whether a video segment is violent or not, the relation between the spatio-temporal features obtained from the output stage of the GRU layer, was passed through several dense layers defined in Eq. (7). Here, W, X_L and D_L are the weights, input and output of the dense layers ($X_0 = h$), respectively. Thus, the series of dense layers in the network operates on the 1D vector, $\mathbf{h}$ to execute a binary classification task. The resulting output corresponds to either 1 or 0, representing the classification of the video segment into the violent or non-violent category, respectively.

$$\mathbf{D}_L = \begin{cases} R(\mathbf{X}_L \cdot W_L), & \textit{hidden layers} \\ S(\mathbf{X}_L \cdot W_L), & \textit{output layer} \end{cases} \tag{7}$$

Further to reduce the risk of overfitting, we have also included dropout layers within the model's hidden layers, with a dropout rate of 0.5. These deliberately positioned layers serve as protections, reducing over-reliance on specific nodes and improving generalization abilities. The hidden layers were also enhanced with ReLU activation functions to increase nonlinearity and allow the model to capture complicated relationships in the data. Finally, the output layer was outfitted with a sigmoid activation function to provide binary outputs for accurate classification.

3.2 | Hyperparameters

The careful selection and optimization of hyperparameters are essential for the development of effective and reliable ML models, as they have a significant impact on the model's performance, generalization ability, efficiency, stability, and interoperability. To determine the value of the hyperparameters in our model, we consider factors including the number of layers, their dimensions, connectivity, filter size, presence of pooling layers, recurrent connections, regularization techniques, activation functions, and other architectural choices.

Specifically, the BrutNet training, initialized with a learning rate of 10^{-5}, continues with the necessary optimization for minimized loss values. For optimization, weighted binary cross-entropy, $\mathcal{L}_{BCE}$ in Eq. (8) was used as the loss function and class-wise weights were assigned in configuring the training. (Here, N is the number of samples in the dataset and y is the label for a given input.) This consideration also helps tackle the imbalanced nature of existing datasets. Besides, the batch size was determined to be 24, considering the size of each sample and the available hardware resources. Moreover, the Adam optimizer has been used to optimize the minimization process.

$$\mathcal{L}_{BCE}(y, D_{\text{output}}) = -\frac{1}{N}\sum_{m=1}^{N}\Big[y \log(D_{\text{output}}) + (1-y)\log(1-D_{\text{output}})\Big] \tag{8}$$

The complete list of trainable parameters for our model can be found in Table 2. Based on our calculations, the total number of parameters in the model is determined to be 3,415,745 (approximately 3.416 million).

TABLE 2 Parameters of BrutNet

Layer (type)	Output Shape	Number of Parameters
TD-DCNN	(None, 24, 512)	1,591,744
GRU	(None, 64)	110,976
Dense	(None, 1024)	66,560
Dense	(None, 1024)	1,049,600
Dropout	(None, 1024)	0
Dense	(None, 512)	524,800
Dropout	(None, 512)	0
Dense	(None, 128)	65,664
Dropout	(None, 128)	0
Dense	(None, 64)	8,256
Dense	(None, 1)	65
Total Parameters		**3,415,745**

4 | EXPERIMENTAL SETTINGS

This section presents the implementation of our BrutNet model and other experiment settings for its training and performance evaluation. The model was implemented using a Google Colaboratory environment having a 2.30GHz Intel(R) Xeon(R) CPU, 12.63 GB RAM, and 12 GB NVIDIA Tesla K80 GPU.

4.1 | Dataset Pre-processing

As outlined in Sec. 2, the primary challenge in AVDC research is that the available datasets usually contain low-resolution and limited samples (*i.e.*, video segments). The widely used datasets also have less diversity of violent scenes, raising the concern of limited learning capability. Hence, our model was trained using the AVD[16] dataset and additionally, evaluated using the HF[6], the RWF-2000[17], and the MF[6] datasets.

Additionally, to address the limitations of the hardware resources utilized in training, the resolution of each video segment in the AVD dataset was resized to 160×90 from its initial resolution for training, with each frame of its RGB frames having the shape of (90, 160, 3). However, in other datasets used for testing, the video segments were of varying resolutions and aspect ratios. To address this limitation, we developed a method to scale up/down the resolution and add the necessary padding to the images to ensure their fixed resolution and aspect ratio. We consider h_0 and w_0, to be the number of pixels along the height and width of the frames of the raw video segment, where the aspect ratio is a_0 in Eq. (9a). In contrast, h_1 and w_1 are considered to be the desired number of pixels along the height and width of the frames, where the aspect ratio is a_1 in Eq. (9b).

$$a_0 = \frac{w_0}{h_0} \tag{9a}$$

$$a_1 = \frac{w_1}{h_1} \tag{9b}$$

Now, to fit the raw image with aspect ratio a_0 into the desired shape with aspect ratio a_1, the raw frames are to be resized with width and height of w_0' and h_0', respectively. Padding is also added to the top and bottom of the resized image with $a_1 < a_0$, and left and right of the resized image for $a_1 > a_0$, as illustrated in Fig. 4a and Fig. 4b, respectively. Thus, an image of the desired shape without distortion due to a change of aspect ratio using Eq. (10a), (10b) and (10c) is obtained.

$$h_0' = \begin{cases} \frac{h_0 w_1}{w_0}, & a_1 < a_0 \\ h_1, & a_1 \geq a_0 \end{cases} \tag{10a}$$

$$w_0' = \begin{cases} w_1, & a_1 \leq a_0 \\ \frac{w_0 h_1}{h_0}, & a_1 > a_0 \end{cases} \tag{10b}$$

$$p = \begin{cases} \frac{h_1 - h_0}{2}, & a_1 < a_0 \\ 0, & a_1 = a_0 \\ \frac{w_1 - w_0}{2}, & a_1 > a_0 \end{cases} \tag{10c}$$

Moreover, the range of pixel values for each RGB video frame is between 0 and 255. These pixel values have been normalized to a scale of 0 to 1 to prevent biasing of our model during its training testing. To avoid the inconsistency of the datasets and to ensure a fixed number of frames in the input video segment, we have further considered selecting 24 random and equally spaced frames such that these frames could represent the overall content of the whole video. As a result, each sample of the dataset has a 4D data shape of (24, 90, 160, 3). For assessment purposes, the violent and non-violent labels have been binarised. The processed samples were compressed and saved locally, which were later loaded using a data generator on TensorFlow (TF) 2.5 framework.

4.2 | Model Training and Evaluation

Once the data was pre-processed, the model underwent training for 50 epochs to monitor the minimization of the loss function and the improvement in accuracy. To ensure the optimal performance of the model, we validated it after each epoch using a separate validation dataset that was created beforehand. This process allowed for readjustment and fine-tuning of the weights of the model.

To optimize the model for maximum accuracy, we stored the training and validation accuracies and losses of each epoch during training. Additionally, we saved the model after every epoch to retain the best accuracy achieved throughout the training process. Once the model is trained, we evaluated the performance using a dedicated test dataset to assess the model's ability to accurately classify violent instances. For this classification task with the optimized model, we developed a Python script employing OpenCV 4.5 and TensorFlow, incorporating a user-friendly graphical user interface (GUI).

For the effectiveness of the selected threshold in achieving a high recall/true positive rate (TPR) (Eq. (11a)) and low false positive rate (FPR) (Eq. (11b)), we plotted the Receiver Operating Characteristic (ROC) curve and evaluated the performance of the BrutNet classifier. The ROC curve illustrates the trade-off between TPR and FPR. The TPR and FPR were calculated based on the True Positive (TP), True Negative (TN), False Positive, and False Negative (FN) values for various thresholds.

$$TPR = \frac{TP}{TP + FN} \tag{11a}$$

$$FPR = \frac{FP}{FP + TN} \tag{11b}$$

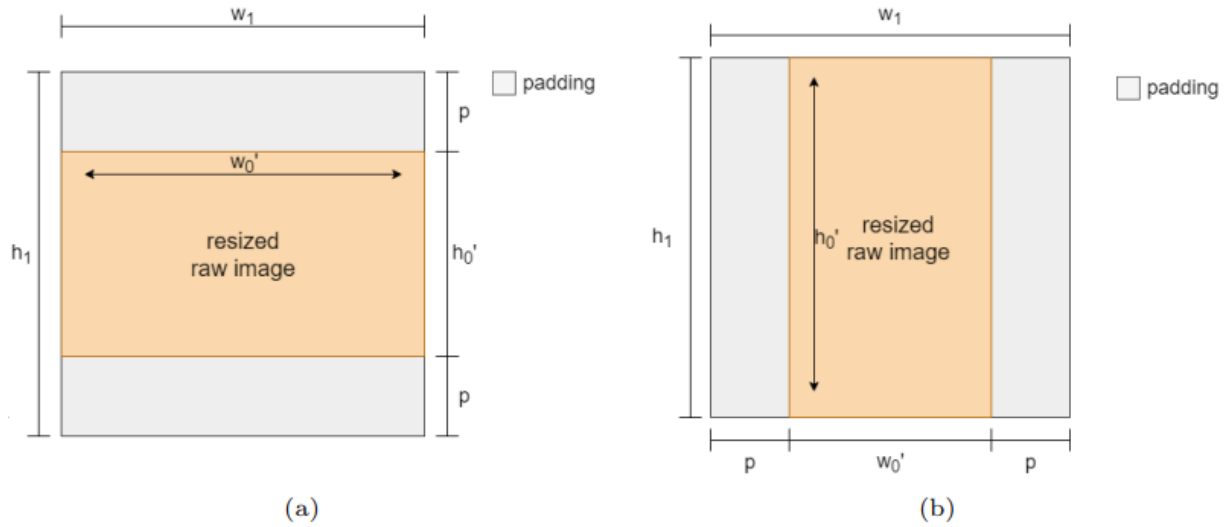

FIGURE 4 Resizing and padding of the raw image, keeping the aspect ratio intact: (a) $a_0 > a_1$ and (b) $a_0 < a_1$

.

5 | RESULTS AND ANALYSIS

The results of our study demonstrate that the proposed BrutNet model achieves state-of-the-art performance in violence detection. The training and validation accuracies, as well as the loss values per epoch, were visualized and compared in Fig. 5 throughout the 50 epochs of model training. The figure suggests that the model offers the most optimal accuracy and loss value at 24 epochs, and hence the model obtained at 24 epochs of training, was considered to be the optimized model.

To demonstrate the attainment of a critical requirement for an AVD classifier, we investigated the proposed model's ability to detect the maximum number of violent scenes. This performance is illustrated with the scatter plot of the model on the test data of the AVD Dataset in Fig. 6a. We observed that the trade-off between recall and FPR, as illustrated in Fig. 7, was optimal at an FPR of 0.056, yielding a recall of 97.14%. This favourable outcome was achieved at a threshold value of approximately 0.2. Thus, from Fig. 6a, the threshold is set at 0.2 such that the maximum number of violent scenes are correctly detected without increasing the number of false detections significantly. Furthermore, the proposed model achieves an impressive area under the ROC curve (AUC) of 0.983, approaching the maximum achievable area of 1 for an ideal classifier according to theoretical expectations. The confusion matrix has also been shown in Fig. 6b to demonstrate the ability of the model to classify violent scenes. We illustrate some classified images by the model in Fig. 8 that reflect the above performance of our proposed model. Despite some erroneous outputs illustrated in Fig. 9, the performance of the proposed model outperforms the other models. These erroneous output frames have been observed for the HF, AVD, and RWF2000 datasets. It is to be noted, there were no erroneous outputs for the MF dataset as it recorded 100% classification accuracy.

To validate the improved performance of the model, its accuracy and number of parameters have also been compared against the other prominent models for violence detection based on their performance on AVD, HF, MF and RWF2000 datasets. Specifically, we have considered the performance of several AVDC models like UNet + PatchGAN[35], ResNet50+LSTM[33], Multi-stream CNN[36], Efficient 3D CNN[37], Hough Forests + 2D CNN[41], Xception + BiLSTM + Attentions[38], Motion Blobs + Random Forest[47], C3D[39], AlexNet + LSTM[40], motion intensities + AdaBoost[44], ResNet50 + ConvLSTM[45], and the state-of-the-art MobileNetV2 + LSTM[34] models. We have also compared the improvement of the new BrutNet over its earlier results[49].

The performance of the proposed BrutNet model is compared across various datasets and related AVDC models, which is illustrated in the Table. 3. Here, we observe that the Multi-stream CNN[36], Efficient 3D CNN[37], Xception + BiLSTM + Attentions[38] and AlexNet + LSTM[40] demonstrate the state-of-the-art classification performance for the MF dataset. For the same dataset, while our model attains the same classification accuracy, its total number of parameters is significantly reduced by 77.82%, 53.84%, 68.95% and 64.42% than the said models, respectively. A similar trend in the reduction of the total number of parameters is also observed in the case of the RWF2000 dataset. Our model outperforms in terms of both the accuracy (4.43% increase) and total parameters (16.15% reduction) over the MobileNetV2 + LSTM[34] model.

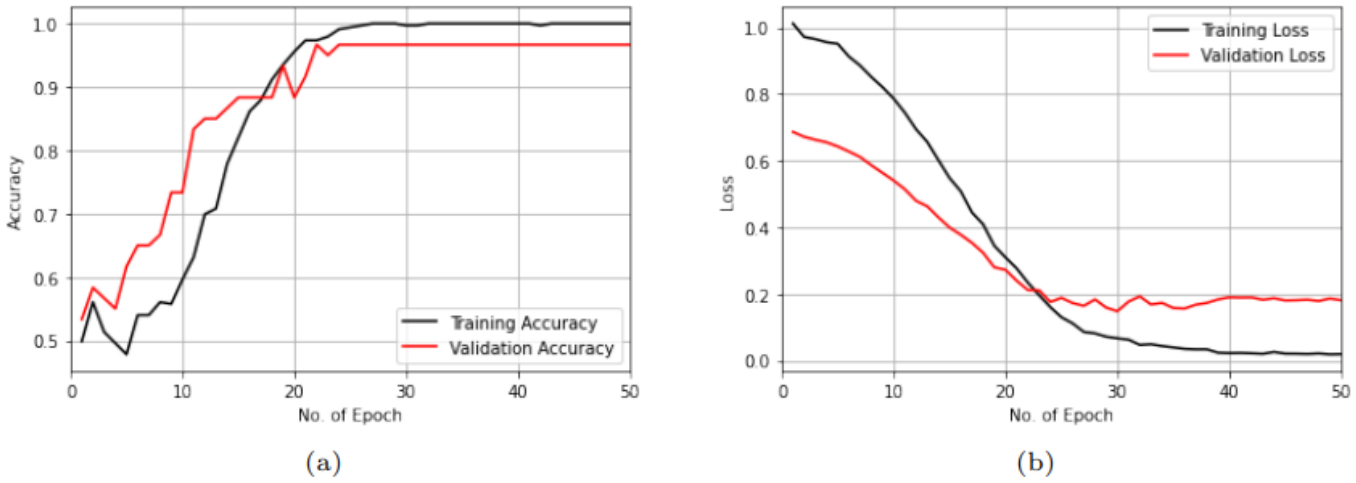

FIGURE 5 Training and validation characteristics per epoch of BrutNet: (a) accuracy and (b) loss.

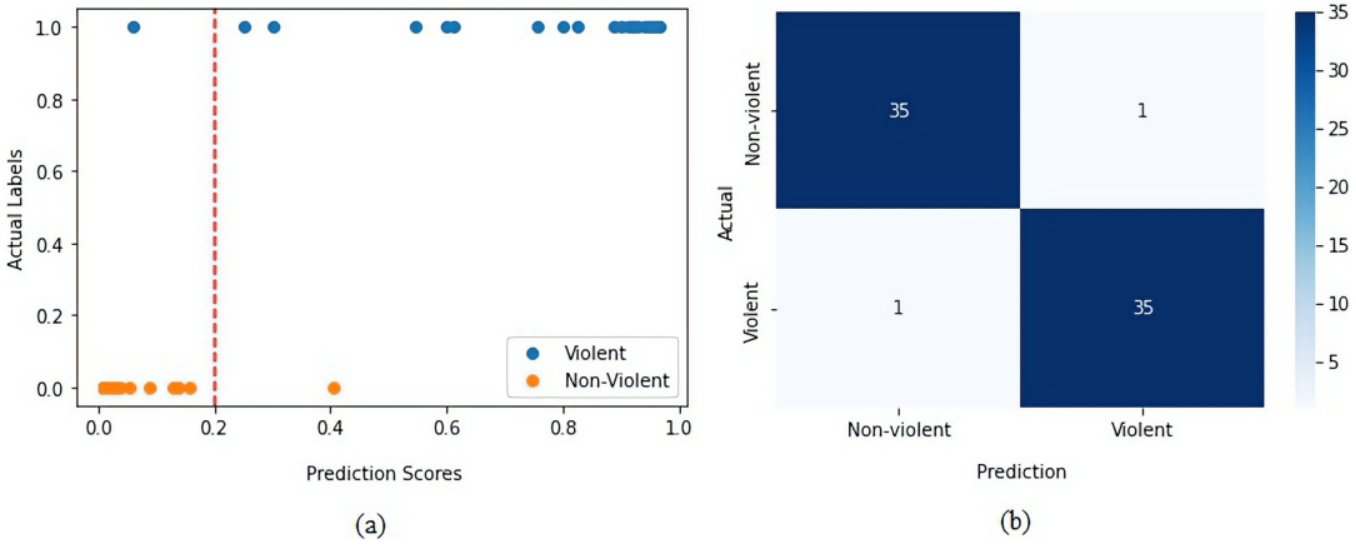

FIGURE 6 Violence classification by BrutNet on AVD Dataset: (a) scatter plot and (b) confusion matrix.

Additionally, for the HF dataset, our model demonstrates a comparable classification accuracy to the prominent models. For example, the high accuracies of 100%, 98.3%, 97.5%, 97.1%, and 96.1% were offered by the Multi-stream CNN[36], Efficient 3D CNN[37], Xception + BiLSTM + Attentions[38], AlexNet + LSTM[40] and MobileNetV2 + LSTM[34] models, respectively. In contrast, our model offers 97.62% classification accuracy for this dataset, which is 2.38% and 0.68% lower than the highest accuracies offered by Multi-stream CNN[36] and Efficient 3D CNN[37] respectively. This fairly diminished classification accuracy of our model is effectively offset by the notable improvement in total parameters reduced by 77.82% and 53.84% compared to the respective models, resulting in a favourable trade-off.

Table. 3 also demonstrates a significant improvement of BrutNet over its initial development presented in[49]. While BrutNet's accuracy has increased by 7.22% (on AVD dataset), its number of parameters has decreased by 2.05M, making the new model computationally 40% more efficient. Furthermore, the model achieved a lower loss value of 0.185, demonstrating over 50% improved optimization over its previous version with a test loss value of 0.386.

Several key factors have contributed to the demonstrated improvements of the proposed model discussed above. For

TABLE 3 Performance of BrutNet across HF, MF, AVD and RWF2000 datasets.

Model	Accuracy on Different Datasets				Total parameters
	HF	MF	AVD	RWF2000	
UNet + PatchGAN [35]	92.61%	91.5%	79.42%	-	-
ResNet50+LSTM [33]	83.19%	88.74%	-	-	-
Multi-stream CNN [36]	**100%**	**100%**	-	-	15.4 M
Efficient 3D CNN [37]	98.3%	**100%**	-	-	7.4 M
Hough Forests + 2D CNN [41]	94.6%	99%	-	-	-
Xception + BiLSTM + Attentions [38]	97.5%	**100%**	-	-	11 M
Motion Blobs + Random Forest [47]	82.4%	97.8%	-	-	-
C3D [39]	87.4%	93.6%	-	-	17.5 M
AlexNet + LSTM [40]	97.1%	**100%**	-	-	9.6 M
Motion intensities + AdaBoost [44]	90.1%	98.9%	-	-	-
ResNet50 + ConvLSTM [45]	87.5%	92%	-	-	-
MobileNetV2 + LSTM [34]	96.1%	99.5%	-	82.0%	4.074 M
BrutNet (previous) [49]	-	-	90%	-	5.466 M
BrutNet (new)	97.62%	**100%**	**97.22%**	**86.43%**	**3.416 M**

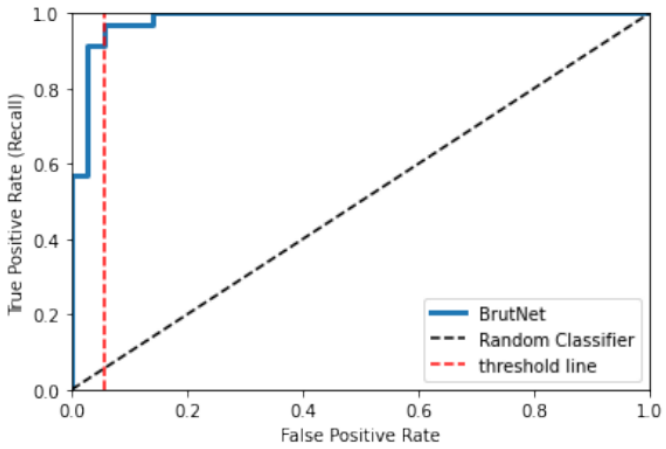

FIGURE 7 ROC curve for BrutNet Model.

example, (*i*) optimization of network architecture and hyper-parameters by employing new hidden layers, while reducing convolutional layers, (*ii*) tackling the skewed datasets with data-preprocessing steps, and (*iii*) an increased frame-feeding rate, *i.e.*, number of frames (*i.e.*, 24 frames, instead of 12 frames) per video segment, to allow for a more thorough analysis of video data, which primarily resulted in the reported performance improvement. Thus, the proposed BrutNet model not only delivers state-of-the-art performance with significantly higher accuracy than most of the prominent models but also accomplishes this with a minimal number of parameters, signifying its computational efficiency.

6 | CONCLUSIONS

Addressing the growing need for public safety and security, there is a pressing demand for an intelligent and efficient AVDC system capable of seamlessly integrating with real-time application scenarios. To meet this demand, this paper introduces a new AVDC model called BrutNet. Our approach combines a custom DCNN with GRU stands out in achieving impressive test accuracies of 97.62%, 100%, 97.22%, and 86.43% on the HF, MF, AVD, and RWF2000 datasets. Furthermore, the model demonstrates efficient parameter utilization, requiring only 3.416 million parameters, thus enhancing its practicality and computational efficiency. With an optimized architecture, fine-tuned hyperparameters, improved handling of skewed datasets, and increased frame-feeding rate, our model offers state-of-the-art performance for real-time application scenarios. This advancement paves the way for a more efficient and effective AVDC system, enabling enhanced public safety, content moderation, censorship, computer-aided investigations, and law enforcement.

BrutNet's improved performance significantly contributes to the advancement of AVDC research. Moreover, it creates opportunities for future investigations, including: (*i*) the development of more suitable loss functions tailored to handle complex and diverse datasets, (*ii*) the establishment of spatio-temporal feature-based evaluation parameters to assess the model's ability to interpret features, and (*iii*) the extension of the model to incorporate audiovisual data for an AVDC system. Despite the progress made in ML-based AVDC approaches, there are still critical areas that require attention, such as the

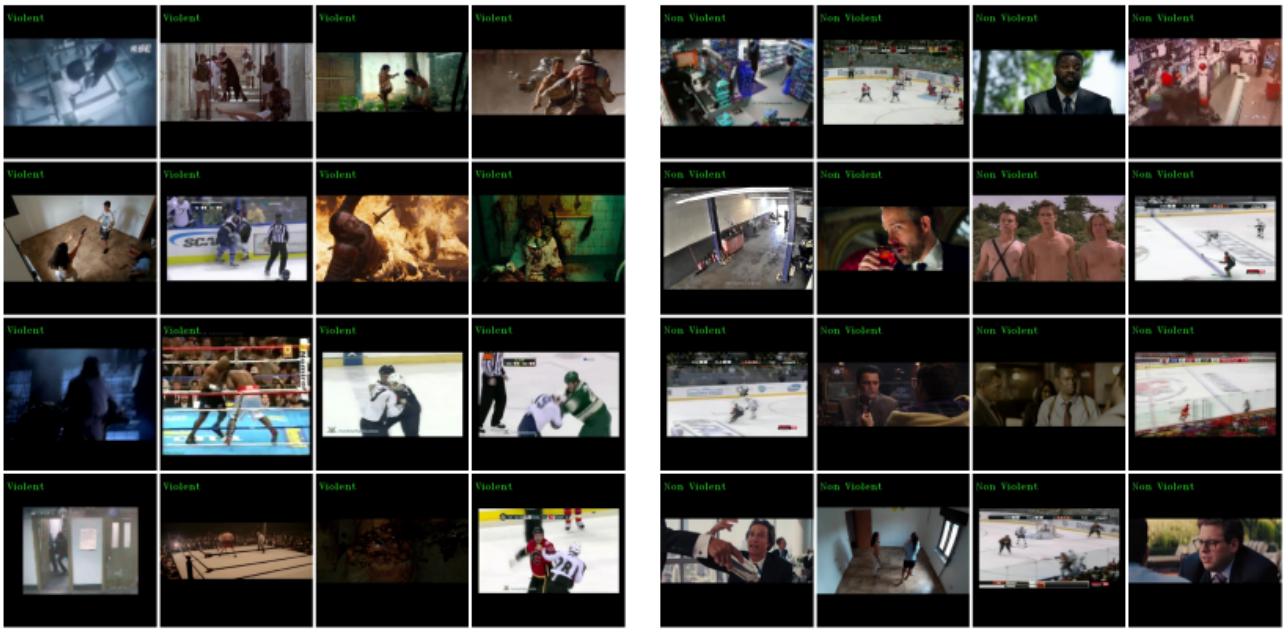

FIGURE 8 Example of classified frames by the BrutNet: (a) violent and (b) non-violent.

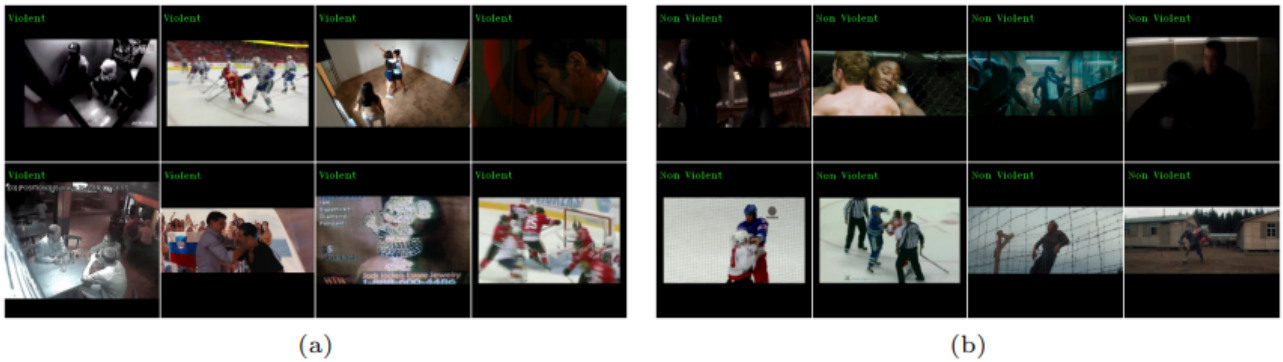

FIGURE 9 Some erroneous output frames from the HF, AVD and RWF2000 datasets: (a) false positive and (b) false negative.

exploration of advanced techniques for video data augmentation and the creation of a comprehensive dataset that captures more complex and diverse scenarios of violence.

DECLARATION

Conflict of interest

The authors have no competing interests to declare that are relevant to the content of this article.

Data availability

Data sets generated during the current study are available from the corresponding author upon reasonable request.

Credit authors statement

Mahmudul Haque: Conceptualization, Methodology, Software, Validation, Investigation, Resources, Writing - Original Draft, Visualization.
Hussain Nyeem: Conceptualization, Methodology, Visualization, Supervision, Administration, Project Management, Resources, Writing - Review & Editing.

Syma Afsha: Conceptualization, Methodology, Investigation, Resources, Visualization.

References

1. Ramzan M, Abid A, Khan HU, et al. A review on state-of-the-art violence detection techniques. *IEEE Access.* 2019;7:107560–107575.
2. Mumtaz N, Ejaz N, Habib S, et al. An overview of violence detection techniques: Current challenges and Future Directions. *Artificial Intelligence Review.* 2022;56(5):4641–4666. doi: 10.1007/s10462-022-10285-3
3. Alkinani MH, Khan WZ, Arshad Q. Detecting human driver inattentive and aggressive driving behavior using deep learning: Recent advances, requirements and open challenges. *IEEE Access.* 2020;8:105008–105030.
4. Pujol FA, Mora H, Pertegal ML. A soft computing approach to violence detection in social media for smart cities. *Soft Computing.* 2020;24:11007–11017.
5. Yang Y, Angelini F, Naqvi SM. Pose-driven human activity anomaly detection in a CCTV-like environment. *IET Image Processing.* 2023;17(3):674–686.
6. Bermejo Nievas E, Deniz Suarez O, Bueno García G, Sukthankar R. Violence detection in video using computer vision techniques. In: Springer. 2011:332–339.
7. Pauwels L, Schils N. Differential online exposure to extremist content and political violence: Testing the relative strength of social learning and competing perspectives. *Terrorism and Political Violence.* 2016;28(1):1–29.
8. Sangwan SR, Bhatia M. D-BullyRumbler: a safety rumble strip to resolve online denigration bullying using a hybrid filter-wrapper approach. *Multimedia Systems.* 2022;28(6):1987–2003.
9. Constantin MG, Ştefan LD, Ionescu B, et al. Affect in multimedia: Benchmarking violent scenes detection. *IEEE Transactions on Affective Computing.* 2020;13(1):347–366.
10. Afsha S, Haque M, Nyeem H. Machine Learning Models for Content Classification in Film Censorship and Rating. In: IEEE. 2022:396–401.
11. Mahmoodi J, Nezamabadi-pour H, Abbasi-Moghadam D. Violence detection in videos using interest frame extraction and 3D convolutional neural network. *Multimedia Tools and Applications.* 2022:1–17.
12. Ullah W, Hussain T, Khan ZA, Haroon U, Baik SW. Intelligent dual stream CNN and echo state network for anomaly detection. *Knowledge-Based Systems.* 2022;253:109456.
13. Pawar K, Attar V. Deep learning model based on cascaded autoencoders and one-class learning for detection and localization of anomalies from surveillance videos. *IET Biometrics.* 2022;11(4):289–303.
14. Haque M, Afsha S, Ovi TB, Nyeem H. Improving Automatic Sign Language Translation with Image Binarisation and Deep Learning. In: IEEE. 2021:1–5.
15. Ullah A, Muhammad K, Ding W, Palade V, Haq IU, Baik SW. Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications. *Applied Soft Computing.* 2021;103:107102.
16. Bianculli M, Falcionelli N, Sernani P, et al. A dataset for automatic violence detection in videos. *Data in brief.* 2020;33:106587.
17. Cheng M, Cai K, Li M. RWF-2000: an open large scale video database for violence detection. In: IEEE. 2021:4183–4190.
18. Do P, Pham P, Phan T. Some Research Issues of Harmful and Violent Content Filtering for Social Networks in the Context of Large-Scale and Streaming Data with Apache Spark. *Recent Advances in Security, Privacy, and Trust for Internet of Things (IoT) and Cyber-Physical Systems (CPS).* 2020:249–272.
19. Khaksar Pour A, Chaw Seng W, Palaiahnakote S, Tahaei H, Anuar NB. A survey on video content rating: taxonomy, challenges and open issues. *Multimedia Tools and Applications.* 2021;80(16):24121–24145.
20. LeCun Y, Bengio Y, Hinton G. Deep learning. *nature.* 2015;521(7553):436–444.
21. Wang P, Wang P, Fan E. Violence detection and face recognition based on deep learning. *Pattern Recognition Letters.* 2021;142:20–24.
22. Wan B, Jiang W, Fang Y, Luo Z, Ding G. Anomaly detection in video sequences: A benchmark and computational model. *IET Image Processing.* 2021;15(14):3454–3465.
23. Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In: IEEE. 2017:1–6.
24. Yang X, Wang Z, Wu K, Xie Z, Hou J. Deep social force network for anomaly event detection. *IET Image Processing.* 2021;15(14):3441–3453.
25. Das S, Sarker A, Mahmud T. Violence detection from videos using hog features. In: IEEE. 2019:1–5.
26. Honarjoo N, Abdari A, Mansouri A. Violence detection using pretrained models. In: IEEE. 2021:1–4.
27. Guedes ARM, Chávez GC. Real-time violence detection in videos using dynamic images. In: IEEE. 2020:503–511.
28. Roman DGC, Chávez GC. Violence detection and localization in surveillance video. In: IEEE. 2020:248–255.
29. Gruosso M, Capece N, Erra U, Lopardo N. A deep learning approach for the motion picture content rating. In: IEEE. 2019:137–142.
30. Wang S, Zhao J, Shao C, Dong C, Yin C. Truck traffic flow prediction based on LSTM and GRU methods with sampled GPS data. *IEEE Access.* 2020;8:208158–208169.
31. Chatterjee R, Halder R. Discrete wavelet transform for cnn-bilstm-based violence detection. In: Springer. 2021:41–52.
32. Peixoto BM, 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.* 2021;78:103174.
33. Ditsanthia E, Pipanmaekaporn L, Kamonsantiroj S. Video representation learning for cctv-based violence detection. In: IEEE. 2018:1–5.
34. Vijeikis R, Raudonis V, Dervinis G. Efficient violence detection in surveillance. *Sensors.* 2022;22(6):2216.
35. Ehsan TZ, Nahvi M, Mohtavipour SM. An accurate violence detection framework using unsupervised spatial–temporal action translation network. *The Visual Computer.* 2023:1–21.

36. Mohtavipour SM, Saeidi M, Arabsorkhi A. A multi-stream CNN for deep violence detection in video sequences using handcrafted features. *The Visual Computer.* 2022:1–16.

37. Li J, Jiang X, Sun T, Xu K. Efficient violence detection using 3d convolutional neural networks. In: IEEE. 2019:1–8.

38. Aktı Ş, Tataroğlu GA, Ekenel HK. Vision-based fight detection from surveillance cameras. In: IEEE. 2019:1–6.

39. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. In: IEEE. 2015:4489–4497.

40. Sudhakaran S, Lanz O. Learning to detect violent videos using convolutional long short-term memory. In: IEEE. 2017:1–6.

41. Serrano I, Deniz O, Espinosa-Aranda JL, Bueno G. Fight recognition in video using hough forests and 2D convolutional neural network. *IEEE Transactions on Image Processing.* 2018;27(10):4787–4797.

42. Dong Z, Qin J, Wang Y. Multi-stream deep networks for person to person violence detection in videos. In: Springer. 2016:517–531.

43. Xu L, Gong C, Yang J, Wu Q, Yao L. Violent video detection based on MoSIFT feature and sparse coding. In: IEEE. 2014:3538–3542.

44. Deniz O, Serrano I, Bueno G, Kim TK. Fast violence detection in video. In: . 2. IEEE. 2014:478–485.

45. Sharma M, Baghel R. Video surveillance for violence detection using deep learning. In: Springer. 2020:411–420.

46. Khan SU, Haq IU, Rho S, Baik SW, Lee MY. Cover the violence: A novel Deep-Learning-Based approach towards violence-detection in movies. *Applied Sciences.* 2019;9(22):4963.

47. Serrano Gracia I, Deniz Suarez O, Bueno Garcia G, Kim TK. Fast fight detection. *PloS one.* 2015;10(4):e0120448.

48. Ehsan TZ, Nahvi M, Mohtavipour SM. Learning deep latent space for unsupervised violence detection. *Multimedia Tools and Applications.* 2023;82(8):12493–12512.

49. Haque M, Afsha S, Nyeem H. Developing BrutNet: A New Deep CNN Model with GRU for Realtime Violence Detection. In: IEEE. 2022:390–395.

AUTHOR BIOGRAPHY

Mahmudul Haque is currently an Adjunct Faculty and Research Assistant at the Department of CSE, IUB, Dhaka. He holds a B.Sc. degree in Electrical, Electronic and Communication Engineering from MIST, Dhaka, which he completed in 2022. His research focuses on the development of robust and intelligent systems that can understand and operate seamlessly in dynamic and complex environments using deep learning in the fields of computer vision, robotics and human-computer interaction. Over the past five years, he has published seven peer-reviewed articles.

Dr. Hussain Nyeem is currently serving as an Associate Professor (Instructor Class 'A') at MIST, Mirpur. He received a B.Sc. degree in electronics and communication engineering (ECE) from the Khulna University of Engineering & Technology (KUET), Bangladesh in 2007 and a Ph.D. degree in electrical engineering and computer science (EECS) from the Queensland University of Technology (QUT), Australia, in 2014 with distinction and nomination for the QUT best Ph.D. thesis award. He was the Assistant Professor and Lecturer at KUET, Bangladesh; Sessional Academic and Doctoral Research Fellow at QUT, Australia; Guest Lecturer at Khulna University and Bangladesh University of Professionals (BUP); and Exchange Research Scholar at the University of Fukui, Japan in his eighteen years of profession. Dr. Nyeem has supervised four masters and more than twenty-five honors projects and theses. He has published a total of fifty peer-reviewed articles, and received five best-papers, one best-presentation, and the Elsevier outstanding reviewer (Optik) awards in the past five years.

Syma Afsha completed her B.Sc. degree in Electrical, Electronic and Communication Engineering from MIST. Dhaka, in 2022. Her research interests revolve around machine learning. robotics, computer vision, and natural language processing. Currently, she is pursuing her Masters in Intelligent Field Robotics System (IFRoS). Now, her research focused on real-time applications of deep learning, robotics particularly in inappropriate content detection, mobile robotics and autonomous vehicles. In the past five years, she has published four peer-reviewed articles.