Received: October 10, 2024. Revised: October 31, 2024.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Enhanced Deep Learning Model to Detect Violence and Gore in Child-Friendly Online Game

Jasson Prestiliano[1] **Azhari Azhari**[1]* **Arif Nurwidyantoro**[1]

[1]*Department of Computer Sciences and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia*
* Corresponding author's Email: arisn@ugm.ac.id

Abstract: Child-friendly online games retain significant potential for violence due to the user-generated content element inherent in the game, and parents often overlook this. Games frequently portray hyperbolic violence, including the depiction of blood and gore. Researchers have conducted several deep-learning studies to detect violence in games. However, the research is still limited to common types of violence, such as fights and firearms, and has not been successful in detecting blood, gore, and other types of violence. An enhanced deep-learning algorithm is necessary to identify those kinds of violence in child-friendly online games. This research will propose a model architecture integrating 3D Convolutional Neural Network (3DCNN) and Bidirectional Long Short-Term Memory (BiLSTM) to extract the spatiotemporal characteristics of frame sequences or videos. The proposed model also adds attention mechanics to categorize videos according to violent content. The 3DCNN-BiLSTM-Att model achieves an average accuracy of 99.14% when evaluated on three separate datasets: the hockey fights dataset, the movies dataset, and the proposed new dataset of online games (consisting of video with non-violent and violent scenes such as blood, gore, fire, and physical confrontation known as melee). The accuracy is higher compared to several other existing deep-learning methods. Using the suggested approach, a system was implemented to notify parents of their children's online gaming activities.

Keywords: Violence detection, Child-friendly online game, Deep learning, Attention mechanism.

1. Introduction

Children love playing online games, particularly on mobile devices, to have fun and socialize. This is especially true after the new average period when it's challenging for them to engage in direct social interactions [1]. Currently, parents are limiting their activities outside the home due to the monkeypox outbreak [2]. Unfortunately, not all parents can supervise children intensively and continuously when they play online games. Some parents choose not to allow their children to play online games, even though it can harm children's social skills [3, 4]. Some other parents allow their children to play online games that are considered safe for them because they have followed game rating guidelines such as Entertainment System Rating Board (ESRB) or Pan European Game Information (PEGI) [5].

In online games, there is currently a feature that can create other game content, also known as user-generated content [6]. Some games with child-friendly ratings already have this feature. Roblox and Minecraft are the most popular online games with these features among children. In ESRB, the rating system is symbolized by a letter: E as Everyone above six years old, T as Teenagers above thirteen years old, and M as Mature above seventeen years old. Meanwhile, PEGI uses numbers (3, 7, 12, 16, and 18) to indicate the minimum age at which someone can access a game. Roblox and Minecraft have gotten the E rating from the ESRB and 7 (seven) from PEGI [7, 8]. However, several online game players use this functionality for their benefit, including perpetrating acts of violence or participating in sexual harassment [9]. Content generated by negligent gamers might have detrimental and violent aspects when accessed or encountered by other minors engaged in online

gaming [10]. The risk lies in the potential for violence in video games to influence children's conduct since they often emulate what they see in these games [11].

Research on violence detection has produced notable findings, yet the violence shown in games varies somewhat from that seen in real life. Games often portray hyperbolic violence, including blood and gore [12]. This research will provide an enhanced deep-learning algorithm for identifying violence in child-friendly online games. The proposed model incorporates sophisticated feature extraction methods to identify severe manifestations of violence, such as blood and gore, found in user-generated content of online games. This guarantees superior detection accuracy compared to conventional models developed for real-world violent situations. This model combines an enhanced 3D Convolutional Neural Network (3DCNN) and Bidirectional Long Short-Term Memory (BiLSTM) with an attention mechanism that adjusts to the dynamic, user-generated material in child-friendly online games such as Roblox and Minecraft, markedly decreasing false positives while maintaining the gaming experience.

This paper is organized as follows: Section 2 introduces related works. Section 3 describes the research methods. Section 4 presents the results and discussion, and Section 5 gives the conclusion and future work directions.

2. Related works

2.1. Deep learning for real-life violence detection

Previous research on the visual side of violence detection using deep learning has been extensive. Usually, the research will use all or some of these video datasets: The Hockey Fight, Violent Crowd, and Violence in Movies. These three popular datasets for violence detection training have yielded very high accuracy when tested using various deep-learning models.

A study conducted by Khan et al. [13] used a Mobile Neural Network (MobileNet) and a Convolutional Neural Network (CNN) to detect violence in film scenes, achieving an average accuracy of 93.25%. After that, Vijjeikis et al. [14] Combined Mobilenet V2 with UNet; they achieved 82% accuracy and 81% precision in light computer conditions. However, the accuracy reached for the Hockey Dataset is only 96.1%, and for the Violent Movie Dataset, it is 99.5%.

On the other hand, Ullah et al. [15, 16] Achieved 97.97% average accuracy using spatiotemporal features and 3D CNN for movies and then developed the research to detect the surveillance video. Furthermore, using CNN combined with long short-term memory (LSTM) can reach a 98% average accuracy trained with the three datasets mentioned [17]. Meanwhile, CNN combined with BiLSTM achieved an accuracy of 99.03% on the three datasets [18]. Those studies were then improved by Mumtaz et al. [19]. They conducted research using CNN's Deep Multinet (DMN), achieving results comparable to GoogleNet and AlexNet but with a learning speed 1.33 times faster than AlexNet and 2.28 times faster than GoogleNet. This study reaches 99.82% for the Hockey dataset and 100% for the violent movie datasets. However, when faced with a video containing blood and gore, the model's accuracy dropped significantly to only 90%.

Another direction for this kind of research is to reduce the level of supervision on survey cameras, as in a study conducted by Choqueluque-Roman and Camara-Chavez [20]. This study used two 2D and 3D CNN streams simultaneously to detect violence. The results had the best accuracy of 88%. Nonetheless, the approach has some deficiencies when the action tubes are improperly removed because of undetected individuals and occlusions.

Another research uses the Full Temporal Cross Fusion (FTCF) network to detect violence visually. This proposed approach performs superior in processing violence feature maps, and the proposed FTCF performs outstanding violence detection on four existing datasets. The accuracy obtained was more than 95% [21]. This method is excellent when tested with fighting scenes but not with fighting with blood scattered.

In addition, the Multi-Scale Spatio-Temporal (MSTN) network is used. MSTN can be thought of as a single-stream framework. The Long Time Building (LTB) branch has a slow frame rate, and the Short Time Building (STB) branch has a fast frame rate. These two branches capture movement features at different semantic levels in time and space. The experimental results show the superiority of the proposed MSTN with an accuracy of 90.25% on the RWF-2000 dataset and 99% on the Hockey dataset [22]. The current violence datasets often exhibit significant time discrepancies between violent and non-violent instances. However, owing to the intricate nature of severe violence that frequently happens in online games, many non-violent activities exhibit significant variability in their temporal dimensions.

Other research uses a combination of approaches or algorithms. A study using a decision tree combined with support vector machines (SVM) to detect school violence got 97.6% accuracy [23], while other

combinations of SVM and feature extraction using deep transfer learning got 82% accuracy and 92% precision [24]. The ability to classify the two classes needs to be balanced and happen to those two combinations. Another group combines Gated Recurrent Units (GRU) and discrete wavelet transforms (DWT) to achieve 96% accuracy [25]. This combination performs well but still needs help detecting some live-streaming videos.

Some recent studies have improved the robustness or efficiency. One implements a Violence Detection Network (VD-Net) to facilitate a final judgment and efficiently notify necessary parties during violent occurrences [26]. However, the accuracy for the popular video datasets is slightly lower than that of other methods. Next, a study called Residual Neural Network (Resnet-34) combines 3D convolutional-based violence detection using optical flow and RGB. It shows robust performance in videos with diverse backgrounds [27]. However, the model's efficacy diminishes somewhat in situations including films with a large concentration of individuals, and it has not been well evaluated in instances of extreme violence often shown in online games.

A study to improve the identification of severe acts of violence is necessary. This paper proposes a model that integrates the optimal characteristics of the 3DCNN spatiotemporal feature extractor with BiLSTM. This model enables the capture of long-term temporal relationships from both directions, along with attention mechanisms, to classify videos to detect severe violence, typically seen in user-generated content within certain child-friendly online games.

2.2. Violence detection in game

Researchers have conducted extensive research on detecting game violence, particularly to forecast the ratings game developers assign. This is necessary to ensure that game players do not receive inappropriate content for their age. Ji et al. [28] detected violence to predict game ratings by comparing videos on game content in a structured manner using two streams from CNN, with an accuracy of 60.33%. Other methods involve classifying dialogue transcripts from the game using the Semantics Matrix, calculated with the Correlated Occurrence Analogue to Lexical Semantics (COALS) algorithm [5]. Machine learning using the Random Forest (RF) method can be a model for detecting violent content in video games with an accuracy of 84.9%, with the most detected content being blood [10]. However, gore and fire haven’t been included in this detection.

In mobile games, violence detection is carried out on Android games that appear on the Google Play Store by analyzing APK/source code on several games that have been published using the Outline language structure tree (OLST), control-stream chart (CSC), and program reliance diagram (PRD) [29] and other study uses the narrative story as violence detection [30]. Meanwhile, a study has utilized deep learning with Convolutional Long Short-Term Memory (ConvLSTM) to detect violence in video game visual content, achieving an accuracy rate of approximately 73.39% [31]. These studies can identify violent in-game content created by game developers. However, they still need to identify user-generated content in online games. Therefore, additional research is required to identify violent elements in user-generated content, particularly in games rated as child-friendly.

2.3. Violence in child-friendly online game

Roblox and Minecraft have been classified as child-friendly online games because game developers removed potentially hazardous elements. Both games have considerable popularity, attracting millions of active users. A significant component contributing to their prominence is the user-generated content function, which allows any member to produce material at their convenience. It may also serve as a tool for teaching and simulation. However, misuse of this capability can produce inappropriate or violent material [6].

Researchers conducted a study to examine participants' reckless behavior in Roblox games. After their seven-year-old daughter's Roblox avatar experienced sexual molestation by multiple individuals, some parents banned Roblox from their homes [32]. A game avatar represents a player via the gaming character they control, and it usually becomes the object of violence in online games. This kind of violence is often beyond parental supervision.

Game developers sometimes fail to notice the violent content in child-friendly online games because they don't design the content for such themes. Regrettably, the game's ability to generate fresh content, combined with a large number of online players, enables the occurrence of violence to be visible to all participants, as age restrictions only apply to the primary game [33]. The Brookhaven content is a prime example of how Roblox games incorporate violence [34], and Piggy's violent content won the Game of the Year award at the eighth Roblox Bloxy Awards [35]. Based on those occasions, a violence detection system is required when a parent feels unsafe with their children’s online game session

without blocking all the fun and educational parts of gaming.

3. Research methods

3.1. Proposed method

Detecting violence in a video recorded during an online gaming session requires multiple processes. The architecture of the proposed model is shown in Fig. 1.

The recorded video must be split into a series of frames. After the first stage, preprocessing for each frame is executed via many methods, namely resizing and contrast enhancement, followed by grayscaling, and concluding with edge and saliency detection.

The first step after video preprocessing is extracting the present features. A 3D convolutional neural network (3DCNN) is used to extract the spatiotemporal features of the preprocessed video frames. After extracting spatiotemporal features, they are progressively processed by Bidirectional Long-Short Term Memory (BiLSTM) in both forward and reverse directions to capture long-term temporal relationships. BiLSTM will provide many hidden states that will later serve as inputs for the attention mechanism. The attention mechanism assigns increased significance to the crucial elements of the feature sequence, enabling the model to concentrate on the most relevant information for a specific job [36]. This process enhances the ability to forecast the presence of violence in the video accurately. By combining 3D CNN for spatiotemporal feature extraction, BiLSTM for temporal dependency modeling, and the attention mechanism for focusing on essential features, this architecture is well-suited for tasks like video classification, where understanding spatial and temporal relationships is crucial.

3.2. Online game dataset

This research proposes a novel dataset of 2364 films evenly divided between violent and nonviolent content typically seen in child-friendly online games such as Roblox and Minecraft. Each video has a runtime of 4 seconds and will be divided into 12 frames per second, resulting in 48 frames for processing. The total number of frames to be processed is 113,742. The dataset categorizes violence into four distinct classifications [37]. They are:

1. Blood. Blood is seen emitted from the game avatar due to impacts from bare hands, melee weapons, or firearms. It also refers to a pool of blood or blood dispersed from gaming avatars.
2. Gore. In a game, the term "gore" refers to the dismemberment of an avatar's bodily parts caused by impactful actions. In a child-friendly online game, gore can exist with or without blood; therefore, we classify severed body parts from a game avatar as gore.
3. Fire. The usage of firearms, explosives, rocket launchers, flamethrowers, or any other incendiary weapon to inflict damage on other game avatars. This category also includes explosive or incendiary gaming avatars.
4. Melee. Melee refers to the physical confrontation or battle between game avatars, including unarmed combat or using melee weapons such as swords, clubs, axes, etc.

The game's four predominant types of violence are the most prevalent. Violence remains evident despite the typical depiction of gaming avatars as humorous or whimsical figures. This dataset is used since the model must identify the blocky gaming avatars not found in other datasets.

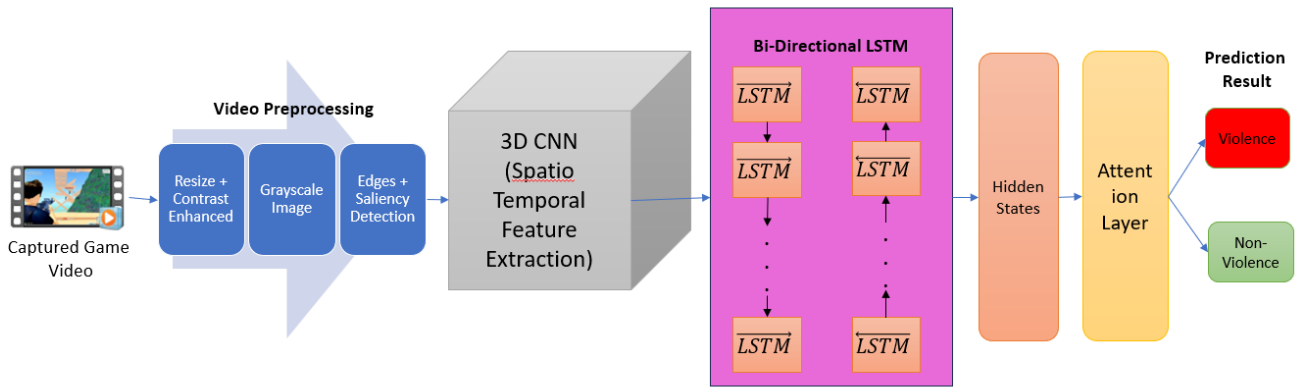

Figure. 1 The architecture of the proposed model

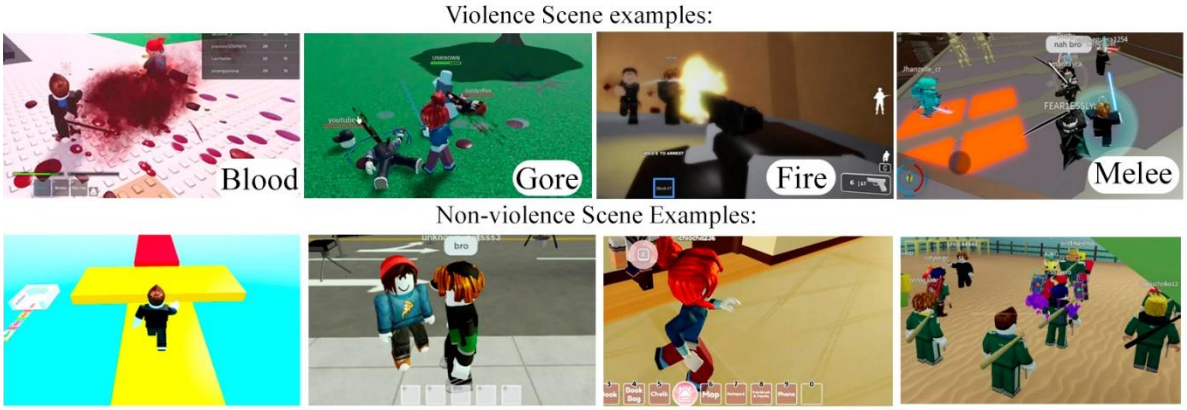

Figure. 2 In-Game Violence Examples and Non-Violence Examples

Meanwhile, the non-violence scenes include running, dancing, jumping, swimming, chopping wood, and other non-violent activities. This dataset is called the Child-friendly Online Game Violence Dataset. Fig. 2 shows examples of violence categories that have been explained before and some non-violent examples.

3.3. Preprocessing

Each video will be divided into sequences of frames to minimize duplication. Each second of the video will be divided into 12 frames despite the video's framerate of 30 frames per second.

Subsequently, image preprocessing was conducted. Resizing guarantees uniformity in input dimensions and minimizes computational demands. It guarantees that the input conforms to the model's design, particularly in convolutional neural networks (CNNs), which need fixed input dimensions. Contrast enhancement improves the visibility of essential details, particularly in low-light video scenarios. In poor visibility conditions (e.g., fog or shadows), contrast enhancement may improve the extraction of spatio-temporal characteristics.

Grayscaling is applied because sometimes, color is insignificant. It can reduce complexity and preserve processing resources. Edge detection highlights structural information, such as object borders, which helps with feature extraction. It improves the ability to recognize shapes and objects, which benefits action recognition and object localization tasks. Saliency detection emphasizes the most essential elements of frames, enhancing model efficiency. This strategy may eliminate extraneous background information and concentrate on high-information regions, enhancing training efficacy.

Fig. 3 shows an original frame from a video in the dataset and the result of each preprocessing step.

3.4. 3D convolutional neural network

The 3D CNN was selected because it is efficacious in feature extraction from video and simultaneously captures spatial and temporal information. In video, a significant attribute exists not only inside each frame (spatial dimension) but also in the transitions between frames (temporal dimension) [38].

A 3D CNN processes video by applying 3D convolution over the spatial dimensions (height, width) and the temporal dimension (depth, i.e., time). The 3D CNN applies 3D convolutions to extract spatio-temporal features from video data. The operation for 3D convolution can be expressed as:

$$y_{(t,h,w)} = \sum_{c=1}^{C} \sum_{i=1}^{T} \sum_{j=1}^{H} \sum_{k=1}^{W} x_c\,(t+i, h+j, w+k).\, W_c(i,j,k) + b \quad (1)$$

Here's the notation list of Eq.1:

- $y_{(t,h,w)}$: the convolution output for the specific position *(t,h,w)* where *(t,h,w)* denote time, height, and width, respectively.
- x_c : the input tensor for channels C with dimension T x H x W x C where:
  - T: the time dimension (or number of frames in the video)
  - H: height dimension
  - W: width dimension
- W_c: the 3D convolutional filters,
- b: the bias term,
- i, j, k: Indices for iterating over the convolutional filter's time, height, and width.

If given an input video represented as a tensor $X \in \mathbb{R}^{T\times H\times W\times C}$. In this representation, C denotes the number of channels (e.g., RGB). The 3D convolution can be expressed as:

$$F_t = \sigma(W_{3D} * X_t + b) \quad (2)$$

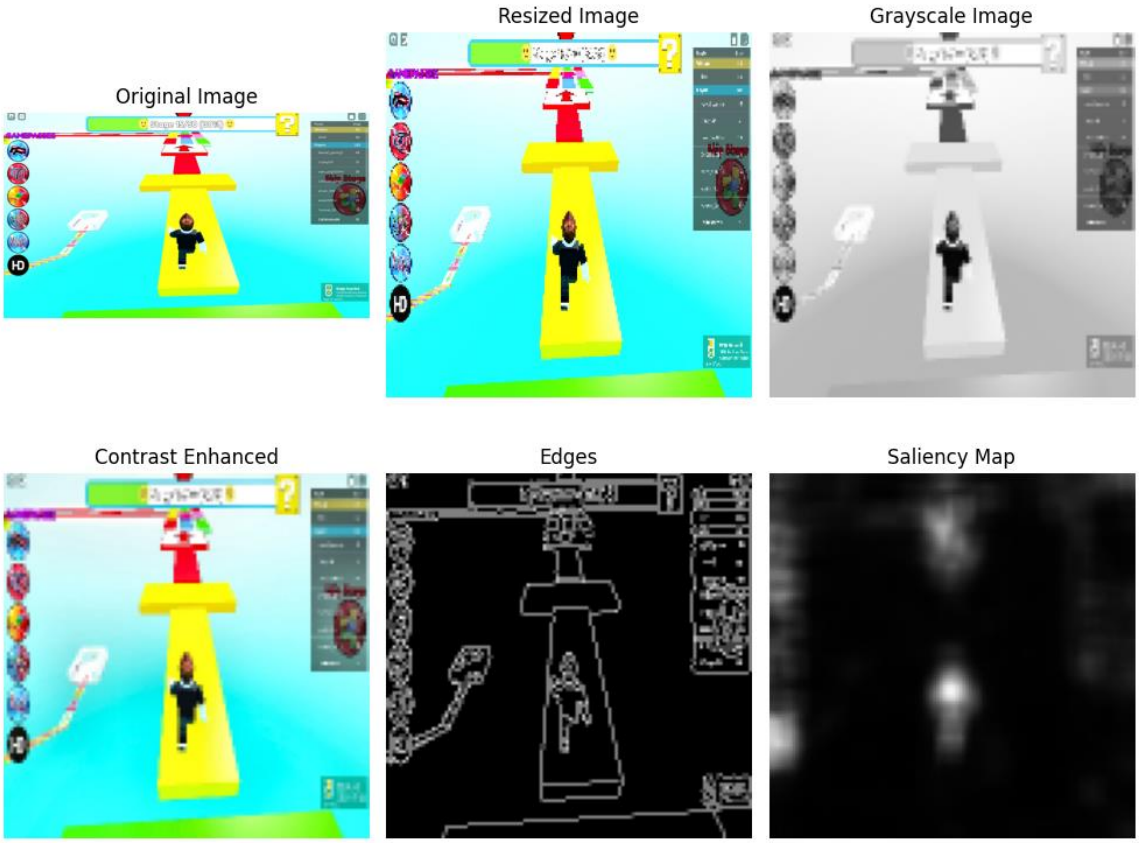

Figure. 3 Example of results for each preprocessing step

Where W_{3D} denotes 3D convolutional kernel (filter), ∗ represents the 3D convolution operation, b is the bias term, σ is a non-linear activation function (such as ReLU), $F_t \in \mathbb{R}^{T\times H\times W\times C}$ represents the spatiotemporal features extracted from the video.

Suppose a video with a 4s duration, T is 48, and the neural network gathers characteristics from the 48 frames; hence, the input dimensions are 3x48xHxW, with H and W denoting the height and width of the frames and the first element indicating the number of channels. The 3D convolution operation consists of several layers of convolution arranged, as seen in Fig. 4.

This step produces a sequence of spatiotemporal features $\{F_1, F_2, \ldots, F_T\}$ for the entire video, where T', H' and W' are reduced dimensions after convolution.

3.5. Bidirectional Long Short-Term Memory (BiLSTM)

BiLSTM was combined in the model architecture because the need for temporal (frame sequence) and spatial (content inside each frame) information for comprehending video material, including action identification and event detection, is excellent. BiLSTM may be used to understand the temporal interactions between frames in greater depth. This enables the model to examine the temporal variations between frames [18].

Based on Figure 1, the extracted spatio-temporal features F_t are fed into a BiLSTM to capture dependencies between different time steps, both forward and backward [39]. The BiLSTM operates as follows for the forward LSTM:

$$\vec{H}_t = LSTM(\vec{H}_{t-1}, F_t) \tag{3}$$

While for the backward LSTM, the BiLSTM operates as follows:

$$\overleftarrow{H}_t = LSTM(\overleftarrow{H}_{t+1}, F_t \tag{4}$$

Where $\vec{H}_t$ and $\overleftarrow{H}_t$ are the hidden states from the forward and backward LSTM, respectively. The final hidden state at each time step H_t is obtained by concatenating the forward and backward hidden states as follows:

$$H_t = \left[\vec{H}_t, \overleftarrow{H}_t\right];\ H_t \in \mathbb{R}^d \tag{5}$$

Where d is the combined dimensionality of the hidden states.

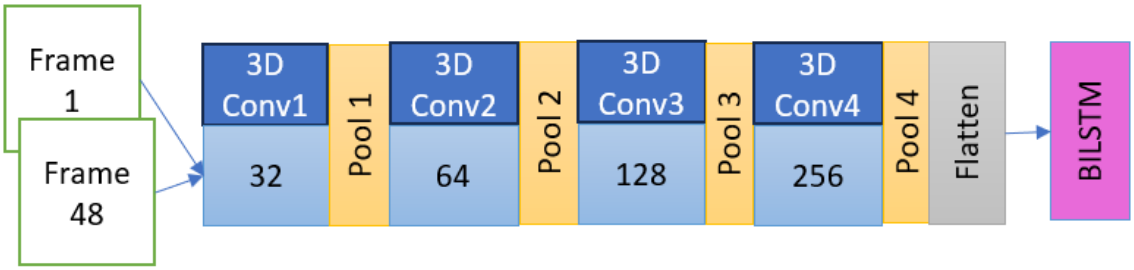

Figure. 4 3D Convolution Layers

3.6. Attention mechanism

The Attention Mechanism computes a weighted sum of the hidden states from the BiLSTM, assigning higher weights to more relevant sequence parts. The mechanism is typically composed of four steps [40].

First, the score calculation that is usually called the alignment score, the equation is as follows:

$$e_t = v^T tanh\,(W_a H_t + b_a) \quad (6)$$

Where H_t denotes the hidden state at time step t while W_a denotes a learnable weight matrix. b_a is a bias term, *tanh* is the hyperbolic tangent activation function, and v is a learnable vector. The result $e_t$is a scalar that represents the importance (or relevance) of the hidden state H_t.

The next step is to calculate attention weights via softmax. The alignment scores e_t is normalized using the softmax function to obtain the attention weights α_t. The equation is represented as follows:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \quad (7)$$

Where $\alpha_t \in [0,1]$ represents the attention weight for the hidden state H_t, and the sum of all attention weights equals to 1. The third step is context vector calculation. The context vector C, which is the weighted sum of the hidden states, is computed as:

$$C = \sum_{t=1}^{T} \alpha_t H_t \,; C \in \mathbb{R}^d \quad (8)$$

Where $C \in \mathbb{R}^d$ is the final context vector that aggregates the most relevant information from the sequence based on the attention weights α_t.

Finally, the final prediction calculation. The context vector C is then passed through a fully connected (FC) layer followed by a softmax or sigmoid function (depending on the task, e.g., classification or multilabel classification):

$$\hat{Y} = Softmax\,(W_c C + b_c) \quad (9)$$

Where W_c denotes the weight matrix for the output layer, b_c denotes the bias term and $\hat{Y}$ denotes the predicted output; in this research, the classes are violence and non-violence.

4. Result and discussion

4.1. Experimental preparation and settings

Data from the dataset explained in section 3.2 is used to evaluate the proposed model's performance. Other datasets used are the Hockey Dataset and the Violent Movie Dataset. The experiments were performed with an NVIDIA RTX 3050 GPU, a Ryzen 7 6800H, and 32 GB of RAM. The video descriptors are computed using CUDA core (version 12.5) and cuDNN acceleration (version 9.1).

The proposed model will be trained using K-fold cross-validation; k = 5 is used for this training. For each fold, 80% of the dataset will be used for training, and the rest will be used for the test. The prediction score from the proposed model will be tested with the tutor's rate in the dataset to find the accuracy, precision, and F1 score for each fold [14].

4.2. Result and discussion

The first phase involves training the proposed model using established violence detection datasets, including the Hockey Fights and Violent Movies datasets, to evaluate its performance. Next, the Online Game Dataset is used to train the model. Multiple training sessions with Hyperparameter tuning adjustments include modifying batch sizes (8 or 16) and the number of epochs (about 5-20). Our proposed model's optimal accuracy measure was 97.84% upon training completion. Besides the accuracy, the precision, recall, and F1 Score of each training was also calculated. The optimal result of the suggested training yielded comparable values for accuracy, precision, and F1 score, all at 97.84%.

Table 1. Accuracy, Precision, and F1 Score for each Tested Datasets

Dataset	Accuracy	Precision	Recall	F1 Score
Hockey Dataset	99.58	99.59	99.57	99.58
Violent Movies Dataset	100.00	100.00	100.00	100.00
Online Game Dataset	97.84	97.84	97.84	97.84

Accuracy is the proportion of adequately classified samples to the total number of samples. Accuracy is calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

Where TP denotes true positive, TN denotes true negative, FP denotes false positive, and FN denotes false negative.

Precision is the proportion of accurately classified "Violent videos" to the total count of labeled "Violent videos." Precision is calculated as:

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

The F1 score is the harmonic mean of precision and recall. To calculate the F1 score, the recall must be calculated first. Recall is the proportion of accurately identified "Violent Videos" to the total number of really violent videos, calculated as follows:

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

After a recall is calculated, the F1 score is calculated as follows:

$$F1\ Score = \frac{2 \times precision \times recall}{precision + recall} \quad (13)$$

Fig. 5 shows the accuracy measure corresponding to Fold 4 of the trials, using hyper-tuning with a batch size of 8 and 8 epochs. Table 1 shows accuracy, precision, and F1 score when the proposed model is trained using the Hockey Dataset, Violent Movies Dataset, and Online Game Dataset.

Fig. 6 shows the confusion matrix for the accuracy measure. The confusion matrix reveals that the model mostly misclassified non-violent films as violent, with 19 instances of misprediction. Upon closer examination, mispredictions often occur when a game avatar is running while carrying an object that might be interpreted as a weapon. Other things misinterpreted as violence are when the avatar jumps over another avatar or passes by other gaming avatars that may be seen as engaging in physical combat. On the other hand, some shooting scenes are considered not violent. This is possible because the bullets or fire are not visible when shot.

The experiments are also done in the same manner as some machine learning models to compare our proposed model with the existing ones. The proposed model has been compared with some existing deep-learning models.

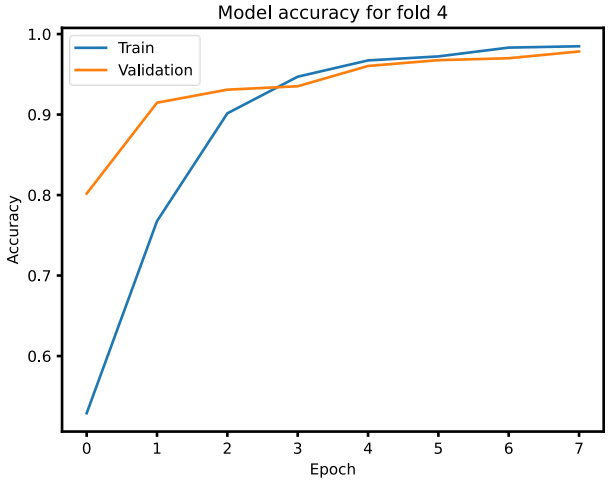

Figure. 5 Accuracy Metric for the proposed model, trained with the child-friendly online game violence dataset

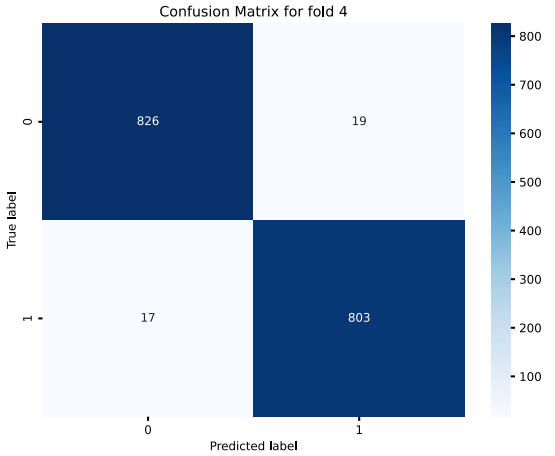

Figure. 6 Confusion Matrix for the proposed model, trained with the child-friendly online game violence dataset

Figure. 7 The implementation of the proposed model to check other video

Table 2. Accuracy outcomes of each model in comparison to the proposed model

Model	Hockey Fights Dataset	Violent Movie Dataset	Online Game Dataset	Average Accuracy
CNN-BiLSTM [18]	99.27	100.00	86.70	95.32
3DCNN [15, 16]	96.00	100.00	93.35	96.45
MSTN [22]	99.00	-	93.95	96.48
VD-Net [26]	98.50	99.00	92.25	96.58
CNN Deep Multinet [19]	**99.82**	100.00	90.00	96.61
MobileNet V2-Unet [20]	96.10	99.50	94.56	96.72
FCTF Net [21]	99.50	100.00	92.55	97.35
Resnet-34 [27]	98.10	100.00	94.10	97.40
3DCNN-BiLSTM-Att (The Proposed Model)	99.58	**100.00**	**97.84**	**99.14**

Some models were mentioned as proven in video surveillance violence detection techniques [41]. The other models are the latest existing models that perform more than 90% accuracy for datasets such as Hockey Dataset and Violent Movie Dataset. A novel dataset included in the comparison is the Online Game Dataset, where severe violent scenes such as blood, gore, fire, and melee happened. These deep-learning models are CNN-BiLSTM [18], 3DCNN [15, 16], MSTN [22], VD-NET [26], CNN Deep Multinet [19], MobileNet V2-Unet [20], FCTF [21], and Resnet-34 [27]. Table 2 shows that the proposed model (3DCNN-BiLSTM-Att) surpassed the overall comparison models, with an average accuracy of 99.14% across all datasets.

4.3. Implementation and limitations

The suggested concept entails deploying an application that operates in the background and automatically records video when a mobile user engages with a mobile game. The program will capture footage till the game is terminated. The video will be uploaded to a server equipped with a violence detection system. The system will segment the video, preprocess it, and then use the pre-trained model to ascertain violent content in each scene. When a violent incident is identified, the system will send a message with an image to the parent associated with the application, allowing them to evaluate the context of their children's exposure to visual violence in the online game. Fig. 7 displays the implementation of the proposed model when it predicts other videos for violent content.

The application's limitations are that parents should install it, and it will affect the children’s device’s memory. Other things that should be considered are the internet connection and the ability to upload the video to the server. On the server side, processing the large size of the video will take more time, so the result will not reach the parent immediately.

5. Conclusion and future works

The proposed model 3DCNN-BiLSTM-Att attains an average accuracy of 99.14% across three distinct datasets. The hockey battles dataset yielded

an accuracy of 99.58%, while the movies dataset achieved an accuracy of 100%. The most notable aspect is its ability to achieve an accuracy of 97.84% when trained on the online game dataset, which includes children's online gaming avatars. This research may assist parents in supervising their children while online gaming, regardless of the game's child-friendly designation. It can identify violence, including blood, gore, fire, and melee confrontations.

Future studies may investigate multilabel categorization because a video may depict multiple forms of violence at the same time. The suggested model may be integrated with specific machine learning techniques capable of simultaneously predicting many classes. Moreover, the study may be further advanced by using a multimodal strategy. For instance, it can be integrated with machine learning to identify terror or intense acoustic stimuli. An additional multimodal example is seen in an online game where several participants communicate, possibly including violent or inappropriate language that is unsuitable for children. The online game can filter some profane terms but not determine if a phrase conveys an objectionable connotation. The suggested model may be integrated with various Natural Language Processing (NLP) techniques to address multimodal visual and verbal aggression.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contribution

The contributions of the authors are as follows: conceptualization, Jasson Prestiliano, Azhari Azhari, and Arif Nurwidyantoro; methodology, Jasson Prestiliano; formal analysis: Azhari Azhari and Arif Nurwidyantoro; Software: Jasson Prestiliano; Validation, Azhari Azhari, and Arif Nurwidyantoro; Verification: Azhari Azhari and Arif Nurwidyantoro; Supervision, Azhari Azhari, and Arif Nurwidyantoro; Writing – original draft preparation, Jasson Prestiliano, Writing – Review and Editing, Jasson Prestiliano, Azhari Azhari, and Arif Nurwidyantoro.

Acknowledgments

The author would like to express his heartfelt gratitude to the supervisors for their guidance and unwavering support during this research.

References

[1] J. J. Hew, V. H. Lee, S. T. T'ng, G. W. H. Tan, K. B. Ooi, and Y. K. Dwivedi, "Are Online Mobile Gamers Really Happy? On the Suppressor Role of Online Game Addiction", *Information Systems Frontiers*, Vol. 26, No. 1, pp. 217–249, 2024, doi: 10.1007/ s10796-023-10377-7.

[2] S. Anil, B. Joseph, M. Thomas, V. K. Sweety, N. Suresh, and T. Waltimo, "Monkeypox_ A Viral Zoonotic Disease of Rising Global Concern", *Infectious Diseases & Immunity*, Vol. 4, No. 3, pp. 121–131, 2024.

[3] D. Gür and Y. K. Türel, "Parenting in the digital age: Attitudes, controls, and limitations regarding children's use of ICT", *Comput Educ*, Vol. 183, 2022, doi: 10.1016/ j.compedu.2022.104504.

[4] J. Prestiliano, "Strengthening Multicultural Community for Teenagers Using Role Playing Game Development", In: *Dynamics of Dialogue, Cultural Development, and Peace in the Metaverse*, IGI Global, Ch. 14, pp. 160–174, 2022, doi: 10.4018/978-1-6684-5907-2.ch014.

[5] A. Maisto, G. Martorelli, A. Paone, and S. Pelosi, "Automatic Classification and Rating of Videogames Based on Dialogues Transcript Files", In: *Lecture Notes on Data Engineering and Communications Technologies*, Vol. 65, 2021, doi: 10.1007/978-3-030-70639-5_28.

[6] H. Duan, Y. Huang, Y. Zhao, Z. Huang, and W. Cai, "User-Generated Content and Editors in Video Games: Survey and Vision", In: *IEEE Conf. on Computational Intelligence and Games, CIG*, pp. 536–543, 2022, doi: 10.1109/CoG5 1982.2022.9893717.

[7] P. 'asher' Rospigliosi, "Metaverse or Simulacra? Roblox, Minecraft, Meta and the turn to virtual reality for education, socialization, and work", 2022, doi: 10.1080/10494820.2022. 2022899

[8] L. Xiao, "Beneath the label unsatisfactory compliance with ESRB, PEGI, and IARC industry self-regulation requiring loot box presence warning labels by videogame companies", *R Soc Open Sci*, Vol. 10, No. 3, pp. 1–33, 2023.

[9] R. C. M. Callou, S. J. B. Bezerra, F. T. L. dos S. Moreira, J. M. Belém, and G. A. Albuquerque, "Cyberbullying e violência de gênero em jogos online", *Saúde e Pesquisa*, Vol. 14, No. 3, pp. 1–15, 2021, doi: 10.17765/2176-9206. 2021v14n3e7920.

[10] F. Zhipeng and H. Gani, "Interpretable Models for the Potentially Harmful Content in Video Games Based on Game Rating Predictions", *Applied Artificial Intelligence*, Vol. 36, No. 1, 2022, doi: 10.1080/08839514.2021. 2008148.

[11] M. Stojanovic, "The effects of playing violent video games on children and youth", *Specijalna Edukacija i Rehabilitacija*, Vol. 18, No. 2, pp. 199–220, 2019, doi: 10.5937/ SPECEDREH18-20876.

[12] J. Denham, S. Hirschler, and M. Spokes, "The reification of structural violence in video games", *Crime Media Cult*, Vol. 17, No. 1, pp. 85–103, 2021, doi: 10.1177/ 1741659019881040.

[13] 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 (Switzerland)*, Vol. 9, No. 22, 2019, doi: 10.3390/APP9224963.

[14] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient Violence Detection in Surveillance", *Sensors*, Vol. 22, No. 6, 2022, doi: 10.3390/s22062216.

[15] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence detection using spatiotemporal features with 3D convolutional neural network", *Sensors*, Vol. 19, No. 11, 2019, doi: 10.3390/s19112472.

[16] F. U. M. Ullah, M. S. Obaidat, A. Ullah, K. Muhammad, M. Hijji, and S. W. Baik, "A Comprehensive Review on Vision-based Violence Detection in Surveillance Videos," *ACM Comput Surv*, 2022, doi: 10.1145/3561971.

[17] A. M. R. Abdali and R. F. Al-Tuma, "Robust Real-Time Violence Detection in Video Using CNN and LSTM", In: *SCCS 2019 - 2019 2nd Scientific Conf. of Computer Sciences*, 2019, doi: 10.1109/SCCS.2019.8852616.

[18] R. Halder and R. Chatterjee, "CNN-BiLSTM Model for Violence Detection in Smart Surveillance", *SN Comput Sci*, Vol. 1, No. 4, 2020, doi: 10.1007/s42979-020-00207-x.

[19] A. Mumtaz, A. Bux Sargano, and Z. Habib, "Fast Learning Through Deep Multi-Net CNN Model For Violence Recognition In Video Surveillance", *Computer Journal*, Vol. 65, No. 3, pp. 457–472, 2022, doi: 10.1093/ comjnl/bxaa061.

[20] D. Choqueluque-Roman and G. Camara-Chavez, "Weakly Supervised Violence Detection in Surveillance Video", *Sensors*, Vol. 22, No. 12, 2022, doi: 10.3390/ s22124502.

[21] T. Zhenhua, X. Zhenche, W. Pengfei, D. Chang, and Z. Weichao, "FTCF: Full temporal cross fusion network for violence detection in videos", *Applied Intelligence*, 2022, doi: 10.1007/s10489-022-03708-9.

[22] W. Zhou, X. Min, Y. Zhao, Y. Pang, and J. Yi, "A Multi-Scale Spatio-Temporal Network for Violence Behavior Detection", *IEEE Trans Biom Behav Identity Sci*, pp. 1–1, 2023, doi: 10.1109/tbiom.2022.3233399.

[23] L. Ye, L. Wang, H. Ferdinando, T. Seppänen, and E. Alasaarela, "A video-based dt–svm school violence detecting algorithm", *Sensors (Switzerland)*, Vol. 20, No. 7, 2020, doi: 10.3390/s20072018.

[24] Karisma, E. M. Imah, and A. Wintarti, "Violence Classification Using Support Vector Machine and Deep Transfer Learning Feature Extraction", In: *Proc. of 2021 International Seminar on Intelligent Technology and Its Application: Intelligent Systems for the New Normal Era, ISITIA 2021*, pp. 337–342, 2021, doi: 10.1109/ISITIA52817.2021. 9502253.

[25] E. M. Imah, I. K. Laksono, K. Karisma, and A. Wintarti, "Detecting violent scenes in movies using Gated Recurrent Units and Discrete Wavelet Transform", *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, Vol. 8, No. 2, pp. 94–103, 2022, doi: 10.26594/ register.v8i2.2541.

[26] M. Khan, A. El 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, doi: 10.1109/ACCESS.2024.3380192.

[27] J. H. Park, M. Mahmoud, and H. S. Kang, "Conv3D-Based Video Violence Detection Network Using Optical Flow and RGB Data", *Sensors*, Vol. 24, No. 2, 2024, doi: 10.3390/s24020317.

[28] Y. Ji, Y. Wang, J. Kato, and K. Mori, "Predicting violence rating based on pairwise comparison", *IEICE Trans Inf Syst*, Vol. E103D, No. 12, pp. 2578–2589, 2020, doi: 10.1587/ transinf.2020EDP7056.

[29] D. Saravanan, J. Feroskhan, R. Parthiban, and S. Usharani, "Secure violent detection in Android application with trust analysis in Google Play", In: *Journal of Physics: Conf. Series*, IOP Publishing Ltd, 2021, doi: 10.1088/1742-6596/1717/1/012055.

[30] G. C. Dobre, M. Gillies, and X. Pan, "Immersive machine learning for social attitude detection in virtual reality narrative games", *Virtual Real*,

Vol. 26, No. 4, pp. 1519–1538, 2022, doi: 10.1007/s10055-022-00644-4.

[31] H. A. Correia and J. H. Brito, "Violence detection in video game metadata using ConvLSTM", In *Proc. of SeGAH 2021 - 2021 IEEE 9th International Conf. on Serious Games and Applications for Health*, Institute of Electrical and Electronics Engineers Inc., 2021, doi: 10.1109/SEGAH52098.2021.9551853.

[32] A. Ghosh, "Analyzing Toxicity in Online Gaming Communities", *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* Vol. 12, No. 10, pp. 4448-4455. 2021.

[33] U. S. Yahsy and M. Syas, "Komodifikasi Users pada Platform Game Online Roblox", *JURNAL INTERACT*, Vol. 11, No. 2, 2022. [Online] Available: https://ojs.atmajaya.ac.id/index.php/fiabikom/index

[34] P. Saputro, "Apa Itu Room Brookhaven di Roblox yang Bisa Beradegan Dewasa ", Detik. Accessed: Oct. 10, 2024. [Online]. Available: https://inet.detik.com/games-news/d-5896856/apa-itu-room-brookhaven-di-Roblox-yang-bisa-beradegan-dewasa

[35] O. B. Mondo, "8th Annual Bloxy Awards: Complete Winners List", Roblox Official Site. Accessed: Oct. 10, 2024. [Online]. Available: https://blog.roblox.com/2021/03/8th-annual-bloxy-awards-complete-winners-list/

[36] 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 (Switzerland)*, Vol. 10, No. 13, 2021, doi: 10.3390/electronics10131601.

[37] F. Alvarez Igarzábal, M. S. Debus, C. L. Maughan, and G. Y. A. W. Clash of Realities (9th : 2018 : Cologne, *Violence, perception, video games : new directions in game research : young academics at the clash of realities 2017-2018*, EBook. Transcript Publishing, 2019.

[38] S. Accattoli, P. Sernani, N. Falcionelli, D. N. Mekuria, and A. F. Dragoni, "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, doi: 10.1080/08839514.2020.1723876.

[39] S. R. Dinesh Jackson *et al.*, "Real-time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM", *Computer Networks*, Vol. 151, pp. 191–200, 2019, doi: 10.1016/j.comnet.2019. 01.028.

[40] H. Mohammadi and E. Nazerfard, "Video violence recognition and localization using a semi-supervised hard attention model", *Expert Syst Appl*, Vol. 212, 2023, doi: 10. 1016/j.eswa.2022.118791.

[41] B. Omarov, S. Narynov, Z. Zhumanov, A. Gumar, and M. Khassanova, "State-of-the-art violence detection techniques in video surveillance security systems: A systematic review", *PeerJ Comput Sci*, Vol. 8, 2022, doi: 10.7717/PEERJ-CS.920.