International Journal of Serious Games

ISSN: 2384-8766
https://journal.seriousgamessociety.org/

Article

Multimodal Deep Learning Violence Detector for Child-Friendly Online Game

Jasson Prestiliano[1], Azhari Azhari[2] and Arif Nurwidyantoro[2]

[1]Visual Communication Design Study Program, Faculty of Information Technology, Universitas Kristen Satya Wacana, Salatiga, Indonesia; [2]Department of Computer Science and Electronics, Faculty of Math and Science, Universitas Gadjah Mada, Yogyakarta, Indonesia
jasson.prestiliano@uksw.edu ; {arisn, arifn}@ugm.ac.id

Keywords:

Multimodal
Deep Learning
Child-Friendly Rated
Online Games
Visual Violence
Verbal Violence

Received: March 2025
Accepted: December 2025
Published: January 2026
DOI: 10.17083/9k96e890

Abstract

The violence present in child-friendly internet games includes both visual and verbal aggression. Visual violence occurs when players perform actions that harm themselves or another player's avatar. On the other hand, verbal aggression often happens during player interactions, even if no physical action takes place. This study explores whether a multimodal deep-learning framework can more effectively detect violence by simultaneously analyzing visual and verbal signals, and whether a hybrid late fusion approach provides better results than traditional fusion methods. Methodologically, the visual modality integrates 3DCNN, BiLSTM, and attention mechanisms, while the verbal modality incorporates BERT and BiLSTM. Each modality is handled independently. The hybrid late fusion employs rule-based and softmax probability to integrate the outcomes of each modality. The proposed multimodal model achieves an average accuracy of 96.72%, with 99.14% for the visual modality and 94.30% for the verbal modality. This performance clearly surpasses existing state-of-the-art fusion methods. The novelty of this study lies in the combination of each modality model and its integration of a hybrid late fusion multimodal approach. Additionally, the study outlines the process and stages for incorporating the model into a system suitable for any child-friendly online game, creating an early warning system for parents.

1. Introduction

Violence in an online game should be indicated by the game ratings appropriate for the player's age [1]. Therefore, if the online game's rating is child-friendly, violence should not occur. However, the user-generated content system accompanying the child-friendly rated online game makes violence possible for players. [2]. Because of its child-friendly rating, parents usually don't pay much attention to what happens when their children play these online games, even though the dangers they could face can be more harmful than online games with a more mature rating [3].

Online games like Roblox and Minecraft are considered child-friendly because the developers do not include elements harmful to children. Both games are very popular, with millions of active users, and are even used in some elementary schools as teaching tools for beginner-level computer programming. The feature that makes these two games very popular is the user-generated content feature offered by both games, which allows each player to create content as they wish. It can even be used for education and simulation. [4]. A study was conducted to observe the behavior of irresponsible players in the Roblox game. Some parents banned Roblox from their homes because of an incident where many players sexually assaulted their seven-year-old daughter's Roblox Avatar [5].

An avatar represents a game player, symbolized by the character they use. Although Roblox has a maximum security feature that filters chats in English by game developers, irresponsible users can still display inappropriate or violent content that can significantly affect the psychology of children who play [6].

The violence that frequently occurs in child-friendly online games includes both visual and verbal violence. Visual violence happens when players perform or witness actions that harm their own avatars or those of others. Verbal violence often happens during chat interactions between players, regardless of whether there is an active in-game situation or not [7]. Some examples of verbal violence are shown in Figure 1.

Figure 1. Some examples of verbal violence that often happen in a child-friendly online game.

Playing games also provides benefits for children. For example, it can introduce players to multicultural communities, [8] and help them develop social skills [9]. However, exposure to violence in online games can increase the risk of psychological harm and mental distress for young players [10]. Detecting violence in online games has its difficulties. For example, parents cannot constantly supervise their children while they are playing online games, or when their children are playing, parents do not realize that their children are encountering visual or verbal violence. [11]. To this end, automated approaches like artificial intelligence technology can assist parents in detecting violence.

Previous works of artificial intelligence models have proposed several approaches to detect visual violence [12], [13], [14], [15], [16] and verbal violence [17], [18], [19], [20],[21], [22], in video or social media, which can also be applied to games. This paper proposes using BERT and BiLSTM to detect verbal violence in Bahasa Indonesia. Then, the result is combined using the late fusion approach with visual violence detection using 3DCNN, BiLSTM, and Attention Mechanics combination proposed by Prestiliano et al. [23] to develop multimodal violence detection. Thus, the contributions of this paper are as follows:

1. A curated dataset of verbal violence in online games, consisting of 10.702 chats written in Indonesian, including common swear word variations used by online game players.
2. A verbal violence detection model using the combination of BERT and BiLSTM.

3. A multimodal violence detection model that integrates the verbal detection model (2) with the visual detection model proposed by Prestiliano et al. [23] Using the hybrid late fusion strategy.
4. A proof of concept implementation demonstrating the multimodal model's possible use in a real-world game system scenario.

2. Related Works

In serious games, machine learning can be used for many things, such as assessing depression and anxiety [24] or detecting violence. Many studies have been conducted to detect violence in visually moving objects. Video violence detection has been conducted with some machine learning techniques, such as Decision Tree – Support Vector Machine (DT-SVM) [12], Full Temporal Cross Fusion (FCTF) network [25], or spatiotemporal modeling methods [26]. Some other studies use deep learning techniques to detect violence in the video, such as a 3D Convolutional Neural Network (3DCNN) [27], [28], there is also a combination of 3DCNN and SVM [13], a combination of Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) [14], [29], or a combination of CNN and Bidirectional LSTM (BiLSTM) [16].

A recent study used Vision Transformer to form a framework called CrimeNet. [30]. Other studies used a multi-scale spatiotemporal network to detect violent behavior. [31] or edge vision-based systems formed as a Violent Detector Network (VD-Net) [32]. Those studies are conducted for surveillance videos. On the other hand, a study focused on movies used optical flow and RGB based on Convolutional 3D to detect violent action [33]. Those studies focused on surveillance videos or films, which usually detect human forms. When the models were tested in child-friendly online game avatars, the models didn't give satisfying results.

Studies about verbal violence detection are usually conducted on social media. Deep learning and machine learning methods, such as CNN and GloVe840, are used to detect cyberbullying tweets [17]. Another study suggests using the CNN and Gate Recurrent Unit (GRU) combination [18]. Meanwhile, the Salp Swarm Algorithm (SSA) and the Deep Belief Network (DBN) can also be used to detect and classify cyberbullying on Social Media [19]. Another study used online game chat logs and text-mining techniques to analyze sexism or toxic language [34]. The latest studies are focused on cyberbullying detection on social media using a transformer-based approach to the Bangla language [21] or using Stacked bidirectional GRU attention with the BERT model [22]. In the field of games, recently, research on games using Bidirectional Long-Short-Term Memory (BiLSTM) to recognize cyberbullying chats was conducted [20]. This model still needs some development in a language other than English. The study in this paper will use Bahasa Indonesia to develop the verbal violence model and add semantic tokenizers to improve its performance in detecting violence in chats.

Multimodal approaches to violence detection are usually conducted in a recorded video or movie. Peixoto et al. [35], [36] use deep convolutional neural networks (dCNN) to detect violence by combining visual and audio detection. This study then focused on video and audio fusion techniques [37]. One study uses semantic and multimodal cross-fusion networks to detect violence in a video [38], and another one uses semantic Violence Correspondence Detection (VCD) to detect violent activities in videos [39]. There is also a C3D fusion to detect violence in a surveillance video [37]. On the other hand, verbal text is included in multimodal cyberbullying detection using the residual BiLSTM model [40] or some congruent reinforced perceptron to mimic the human perceptual system for hateful meme detection [41]. Unlike other studies, this paper's multimodal approach will use visual and verbal modalities. Each modality will be processed individually and combined with late fusion using the hybrid method.

In games, violence detection has also become a concern for many researchers. A study uses ConvLSTM to detect violence from the game metadata [42]. Other studies use Random Forest

Models to detect potentially harmful content based on game rating prediction [43], social attitude detection from the game narrative [44], or by tracking the violence from the code and manifest made by the game developer [45]. Those studies focused on what game developers do in the game development or testing phases. The issue is that using the user-generated system in the online game they develop can cause the player to commit some violence, although the game is rated as child-friendly.

3. Methods and Materials

3.1 Visual Violence Detection

The visual violence detection that becomes the visual modality of this study combines a 3D Convolutional Neural Network (3DCNN), BiLSTM, and an Attention mechanism in the model architecture. The architecture model is shown in Figure 2.

First, the preprocessing methods, such as resize, grayscale, add contrast, edge detection, and saliency detection, are conducted. After that, 3DCNN is used to extract the spatiotemporal characteristics of the preprocessed video frames. Following the extraction of spatiotemporal traits, they are systematically processed by BiLSTM in both forward and backward orientations to capture long-term temporal dependencies. BiLSTM will generate several hidden states that will subsequently function as inputs for the attention mechanism. The attention mechanism enhances the importance of key components within the feature sequence, allowing the model to focus on the most relevant information for a particular task [15].

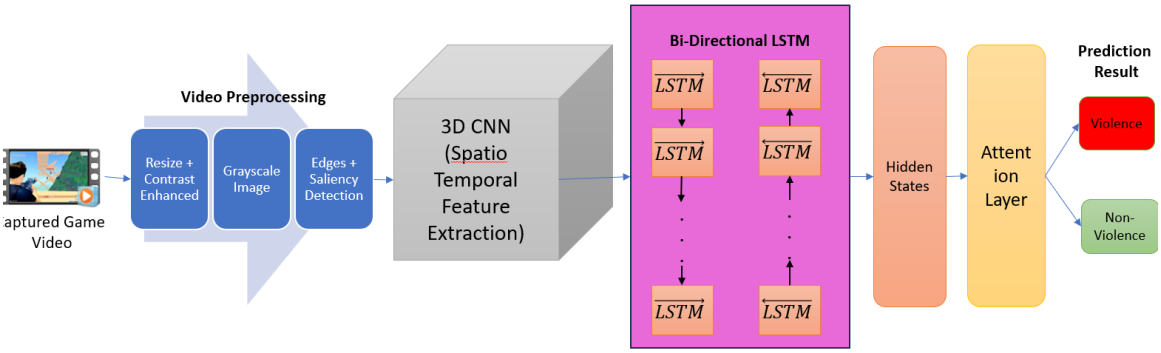

Figure 2. The model architecture for visual violence detection combines 3DCNN, BiLSTM, and the Attention Mechanism. It is applied after several preprocessing methods, such as resizing, contrast enhancement, grayscaling, edge detection, and saliency detection.

This approach improves the accuracy of predicting the occurrence of violence in the video. This architecture effectively integrates 3DCNN for spatiotemporal feature extraction, BiLSTM for modeling temporal dependencies, and an attention mechanism to emphasize critical features, making it highly suitable for video classification tasks that require comprehension of spatial and temporal interactions.

Those combinations have been experimented with and tested with three datasets. Two datasets are used as video violence detection benchmarks: the Hockey Dataset and the Violent Movie Dataset. The model achieved an accuracy of 99.58% for the Hockey Dataset and 100% for the Violent Movie Dataset. The last dataset used is the Online Game Dataset, where the blocky figures of Roblox and Minecraft avatars appeared. The model achieves the best accuracy of 99.14% when it is tested with the Online Game Dataset. The visual violence provided in the test is the captured game videos that consist of blood, gore, fire (or gunshots), and melee (or fights) [23].

3.2 Indonesian Chat Dataset

The Indonesian language has a unique structure compared to the English language. It has no chronological-based tense but has some subject-predicate-object-adjective forms of sentence. A study that uses BiLSTM to detect hate speech in the Indonesian language on social media [46]. This study inspired the use of the Indonesian Chat Dataset. This dataset uses Bahasa Indonesia. In Roblox and Minecraft, many English words are blocked, mainly if they contain swearing or direct bullying.

The interesting thing that can be found is that in Bahasa Indonesia, there are so many association words that Indonesian people recognize as violent words. For example, the word "anjing" means a dog, and in Bahasa Indonesia, it is initially used for an animal called a dog. On the other hand, many people in Indonesia associate this word with a swear word for others, especially when mad. This word has many writing variations, such as “anjir, anjay, njing, anjrot, ajg, 4nj1n9, etc.” These variations may develop even more over time. Therefore, word-by-word detectors cannot detect violent chats in their entirety.

The Indonesian Chat Dataset used in this study collects about 14.000 Indonesian chats from Roblox and Minecraft players. The chats were collected from approximately 42 students who are also Roblox and Minecraft players who regularly play these online games. They then engaged other players in game sessions or rooms that could potentially involve violent, racist, or harassing content for about 2 months. Each captured chat was captured using screenshots, which were then extracted using OCR and parsing to separate the chat from other elements in the captured images. Only publicly available sources were used, and all messages were anonymized to safeguard privacy and prevent the identification of individuals. Care was taken to avoid reinforcing biases or disproportionately representing specific groups. Nevertheless, societal implications must be acknowledged: misclassification could unjustly censor benign interactions or fail to capture harmful content, and large-scale monitoring of conversations raises concerns regarding privacy and freedom of expression.

The parsed chats are then checked manually for validity, and chats containing too many emoticons or meaningless words are deleted. So, the curated chats that can be used as the dataset are about 10.702 chats. The authors manually conducted the annotation by thoroughly examining each chat message. Given the explicit nature of violent, racist, and harassing expressions in the collected data, each instance could be clearly identified when read carefully in context. This careful review ensured that the labeling process was consistent and reliable. Those chats are labeled into four classes: neutral, violence, racist, and harassment. The explanation and data distribution of the Indonesian Chat Dataset are shown in Table 1.

As seen in Table 1, although the dataset does not contain identical samples in each class, the distribution can be considered relatively balanced, with proportions ranging from 23.21% to 27.87% across the four categories. The racism and harassment chat is more challenging, so the frequency of those classes is lower. Nonetheless, the dataset avoids extreme imbalance, ensuring that each class is adequately represented for training and evaluation purposes. The full Indonesian Chat dataset is publicly available. The Dataset can be found here: https://www.kaggle.com/datasets/jprestiliano/indonesian-chat-dataset.

Table 1. Each class's content explanation and data distribution in the Indonesian Chat Dataset: Neutral, Violence, Racist, and Harassment.

Class	Content	Amount of Data	Percentage
Neutral	no violent sentences, casual chat without any means to harm someone	2729	25.49
Violence	swearing sentences, threats, incitement to harm others, or associating people with some animal or creature	2983	27.87
Racist	discriminate against or demean people based on race, religion, ethnicity, or nationality, including slurs, hate speech, or promoting racial superiority	2506	23.42
Harassment	porn, sexual abuse words, body shaming, derogation	2484	23.21
	Total Data	**10.702**	

3.3 Verbal Violence Detection

This study provides the verbal violence detection modality before implementing the multimodal approach. The source is similar to Figure 1, a captured game video. An optical character recognition (OCR) system is implemented to capture the chats, and then it is parsed and separated; for example, it will get rid of the user name that is usually written before a colon symbol and after the colon symbol, which is the start of a chat. After the chats are parsed, the next preprocessing is to lowercase and normalize sentences. For example, if a word has more than two similar letters, like 'heeeey,' the system will let one letter and get rid of the other similar letter so that the word will read as 'hey.' This preprocessing is conducted to prepare the word to be tokenized and become weighted word vectors by the following process.

The proposed violence chat detection model uses BERT combined with BiLSTM. Transformers are the foundation of BERT, a deep learning model that dynamically calculates the weightings between each input and output element based on their connections. BERT is particularly adept at several functions that enable this, including sequence-to-sequence-based language generation tasks, such as abstract summarization, sentence prediction, question answering, and conversational response generation [47]. BERT embeddings are contextualized to capture the meaning of a word in its context within a sentence. Consequently, BERT is an exceptional option for integration into automated essay assessment systems. One of the most effective NLP methods for enabling the machine to understand the context of a sentence is BiLSTM, which can generate more meaningful outputs by integrating LSTM layers from both directions [48]. The proposed model architecture to detect violent chats is shown in Figure 3.

The model proposed will detect and classify the chats into four classes. They are (1) neutral chats that use some potentially violence associated words in the usual way, (2) violent chats that use the associated words as violence to others, (3) racist chats, which use other types of sentences that promote racism and religiosity derogation, and (4) harassment chats that contains some obscene, impolite words that are offensive to human body parts.

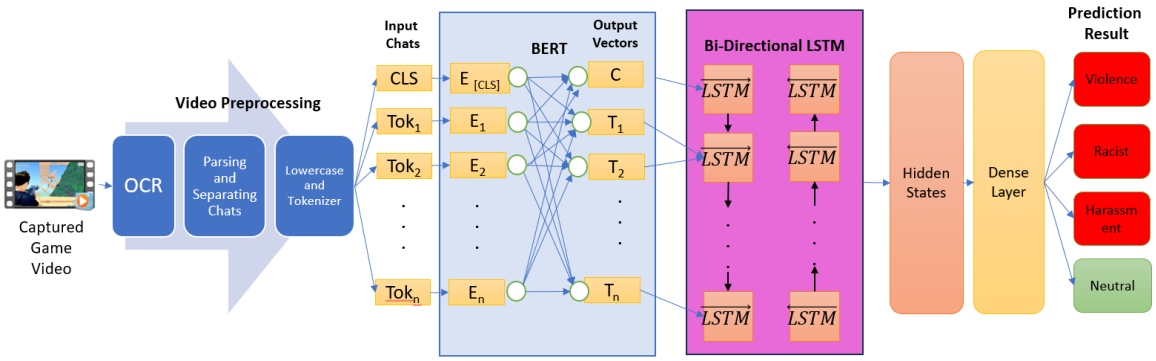

Figure 3. The model architecture for verbal violence detection involves capturing chat text using OCR and applying several preprocessing methods, such as parsing, separating chats, lowercasing, and tokenizing. The combination of BERT and BiLSTM is then implemented to detect four classes: neutral, violence, racist, and harassment.

The feature extraction that uses BERT has several steps that can be represented as a mathematical equation. First, in Equation 1, the input chat sequences are defined as follows:

$$Tok = [CLS, Tok_1, Tok_2, \ldots, Tok_n] \quad (1)$$

CLS is a unique classification token, and Tok_i represents each token in the sequence. Each token is then mapped into an embedding space using a trainable embedding matrix, as shown in Equation 2:

$$E_i = W_E.Tok_i + b_E \quad (2)$$

where W_E is the embedding matrix, b_E is the bias term, and E_i is the embedding representation for the token Tok_i. The embeddings are processed through L transformer layers, where each layer consists of multi-head self-attention and feed-forward networks. The contextualized word representation is computed as shown in Equation 3.

$$H^l = TransformerLayer^l(H^{l-1}) \; where \; H^0 = [E_{CLS}, E_1, E_2, \ldots, E_n] \quad (3)$$

H^0 represents the input embedding matrix, and H^l represents the hidden representation at the l-th transformer layer. Equation 4 is the final contextualized embedding for token i, which is given by:

$$T_i = H_i^L \quad (4)$$

Where H_i^L represents the final presentation after L BERT layers. The CLS token representation, denoted as $C = T_{CLS}$, is the feature representation for the entire chat message.

The contextualized embeddings derived by BERT are input into a BiLSTM network to capture sequential dependencies within the text. The BiLSTM comprises two LSTM networks: a forward LSTM that analyzes the sequence from left to right and a backward LSTM that examines the sequence from right to left [49]. For the forward LSTM, the hidden state at time step i is calculated as shown in Equation 5.

$$\vec{H}_i = LSTM_f(T_i, \vec{H}_{i-1}) \quad (5)$$

Where $\vec{H}_i$ is the hidden state of the forward LSTM or $LSTM_f$ at step i. While T_i is the contextualized embedding from BERT and $\vec{H}_{i-1}$ is the hidden state of the previous time step.

Likewise, with the backward LSTM, the hidden state at step i is determined, as shown in Equation 6.

$$\overleftarrow{H_i} = LSTM_b(T_i, \overleftarrow{H}_{i+1}) \quad (6)$$

Where $\overleftarrow{H_i}$ represents the hidden state of the backward LSTM or $LSTM_b$, and $\overleftarrow{H}_{i+1}$is the hidden state of the subsequent time step in reverse order. H_i represents the bidirectional hidden representation that captures past and future dependencies in the chat sequence. Equation 7 shows how the final BiLSTM representation for token i or H_i is derived by concatenating the forward and backward hidden states.

$$H_i = \left[\vec{H}_i, \overleftarrow{H_i}\right] \quad (7)$$

After the sequential feature learning is conducted using BiLSTM, the classification uses a dense layer and softmax activation. First, the final representation H obtained from BiLSTM is passed through a fully connected dense layer to map the learned features into a lower-dimensional space. The transformation is shown in Equation 8.

$$Z = W_H.H + b_H \quad (8)$$

W_H is the weight matrix, b_H is the bias term, and Z is the intermediate representation. To classify the chat message into one of the four categories (Neutral, Violence, Racist, Harassment), the model applies a softmax activation function, which converts the dense layer outputs into probability scores shown in Equation 9.

$$P(y|X) = softmax\,(Z) \quad (9)$$

P(y|X) is the probability score from the softmax activation function of the intermediate representation *Z*. If there are four classes, the probability for each class *i* is calculated as shown in Equation 10.

$$P(y_i) = \frac{e^{Z_i}}{\sum_{j=1}^{4} e^{Z_i}} \quad (10)$$

Equation 10 represents the softmax function used in classification models to convert logits (raw model outputs) into probabilities. The exponential function e^{Z_i} (where e is an Euler number) ensures that all outputs are positive (since exponentiation never results in negative values). The largest logit gets the highest probability, and the sum of probabilities across all classes equals 1, making it a proper probability distribution. The final classification decision is made by selecting the class with the highest probability. It is shown in Equation 11.

$$\hat{y} = \arg\max P(y_i) \quad (11)$$

Where $\hat{y}$ represents the predicted class label. If *P(Violence)* is the highest, the chat will be classified as violent; if *P(Racist)* is the highest, it will be classified as racist; if *P(Harassment)* is the highest, it will be classified as harassment, and if *P(Neutral)* is the highest, it will be classified as neutral chat.

3.4 Multimodal Approach with Hybrid Late Fusion Algorithm

The multimodal approach is required because many modalities can determine violence within a video capture. This study uses video and chat as the modalities. The video will show the violence visually, while the chat could show violence verbally. Audio is not included because the sound in child-friendly online games usually uses funny or ordinary sounds to disguise the violence. The video and the chat will be processed using an unimodal approach, and then the result of each modality will be combined using a late fusion technique. Late fusion is the most straightforward and often used fusion technique. It integrates data after distinct comprehensive processing in several unimodal streams. The different modalities may be addressed using robust, targeted methods customized to the unique characteristics of each modality. Following a thorough sequence of unimodal processing, usually after label prediction in a recognition task, the outcomes are consolidated, often by summing or averaging. Late fusion has a significant limitation due to its restricted capacity to leverage cross-correlations across various unimodal data [50].

This study proposed a model to detect visual violence using 3DCNN combined with BiLSTM and Attention Mechanics, which results in either violence or non-violence. The verbal violence results in this study are divided into four classes: neutral (non-violent), violence, racist, and harassment, with the last three classes considered violence. The results of each modality have different class numbers, so they should be combined using the rule-based fusion.

The rule-based late fusion could be the logical decision for the final result. It is based on explicit if-else rules applied to visual and verbal predictions [51]. The final result class will consist of four classes: non-violence, visual violence, verbal violence, and visual-verbal violence. The rules for visual and verbal detection and the final results are shown in Table 2.

Table 2. The rules for the visual detection, verbal detection, and final results

Visual Detection	Verbal Detection	Final Result
Non-violence	Neutral	Non-Violence
Violence	Neutral	Visual Violence
Non-Violence	Violence, Racist or Harassment	Verbal Violence
Violence	Violence, Racist or Harassment	Visual-Verbal Violence

The rule-based table works well for clear-cut cases. Still, probability weighting is needed in borderline scenarios where visual and verbal probabilities are close to the threshold (e.g., 51% violence vs. 49% non-violence) or multiple verbal categories have non-negligible probabilities (e.g., 24% Harassment, 24% Neutral, 24% Racist, 28% Violence).

The case that usually happens in an online game that requires probability is needed; there are game scenes with mild violence and aggressive chat. This case has not shown violence, like pushing another character; however, with aggressive chat, mild violence should be detected as violence. In another case, some sentences can be classified as violent or neutral at once. For example, when someone says "Aku melihat si anjing" in the chat, that means "I saw the dog," it can be determined as neutral if there's a dog in the scene. However, it can be defined as violent if only humans are on the scene.

So, if visual or verbal probabilities are close to the threshold, use the weighted formula shown in Equation 12.

$$P_{final} = \alpha P_{visual} + \beta P_{verbal} \tag{12}$$

P_{visual} is the probability of violence or non-violence class in visual detection, *and* P_{verbal} is the probability of neutral, violence, racist, and harassment class in verbal detection. α and β are the weights for each probability. They could be set as $= 0.6$ and $= 0.4$ if the model prioritizes visual violence detection and vice versa. The weight is adjustable. After adjusting the probabilities, Table 2 can be modified into Table 3.

Table 3. Modified Table with Probability Weights

Visual Detection	Verbal Detection	Final Result
Non-violence (p > 0.75)	Neutral (p > 0.75)	Non-violence (rule-based)
Violence (p > 0.75)	Neutral (p > 0.75)	Visual violence (rule-based)
Non-violence (p > 0.75)	Violence, Racist or Harassment (p>0.75)	Verbal violence (rule-based)
Violence (p > 0.75)	Violence, Racist or Harassment (p>0.75)	Visual-verbal violence (rule-based)
Non-Violence (p < 0.75)	Neutral (p < 0.75)	Weighted fusion to decide between Non-Violence, Visual Violence, Verbal Violence & Visual Verbal Violence
Violence (p < 0.75)	Violence, Racist or Harassment (p < 0.75)	Weighted fusion to decide between Non-Violence, Visual Violence, Verbal Violence & Visual Verbal Violence

To summarize the hybrid late fusion proposed in this study, Figure 4 shows the architecture of the multimodal approach for visual and verbal violence detection with hybrid late fusion.

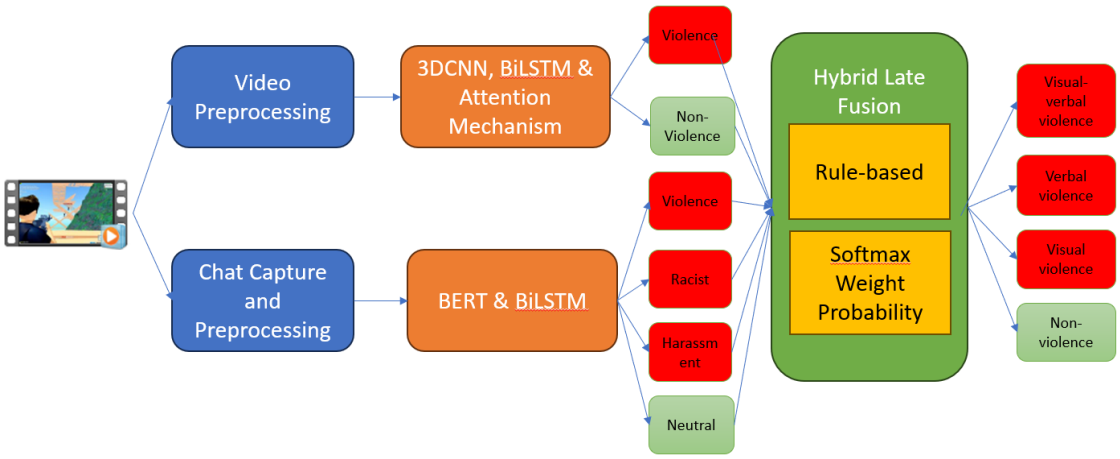

Figure 4. The architecture of a multimodal approach with hybrid late fusion, with Rule-based and softmax weight probability that combines the visual and verbal modalities in violence detection.

3.5 In-Game Early Warning System as A proof of Concept Implementation

The multimodal violence detection model is packed into a system that can be embedded in any game and works as an early warning system for the parents. The system flow of the in-game early warning system is shown in Figure 5.

The system will have two essential parts. The first part is the in-game system that runs in the background when the game is played on a device (marked with the numbers 1 to 4). This part will capture video displays as long as the game is played. Video captures are separated into 5-minute durations and then saved. Then, the system will record again until the game is finished. The duration of 5 minutes for each video was chosen so as not to overload the network when uploading videos.

The other part of the system is run on the server (marked with the numbers 5 to 8). This part will receive the captured videos, which are sent individually, and then each video is processed using the proposed model. Suppose violence occurs in one or both modalities. In that case, the system will send an early warning message to parents about what their children experience when playing online games via the typical messaging system. If there is no violence, an early warning is not sent. After all these processes are complete, the videos checked will be deleted from the device. The system will continue the detection process until all videos on the device are checked.

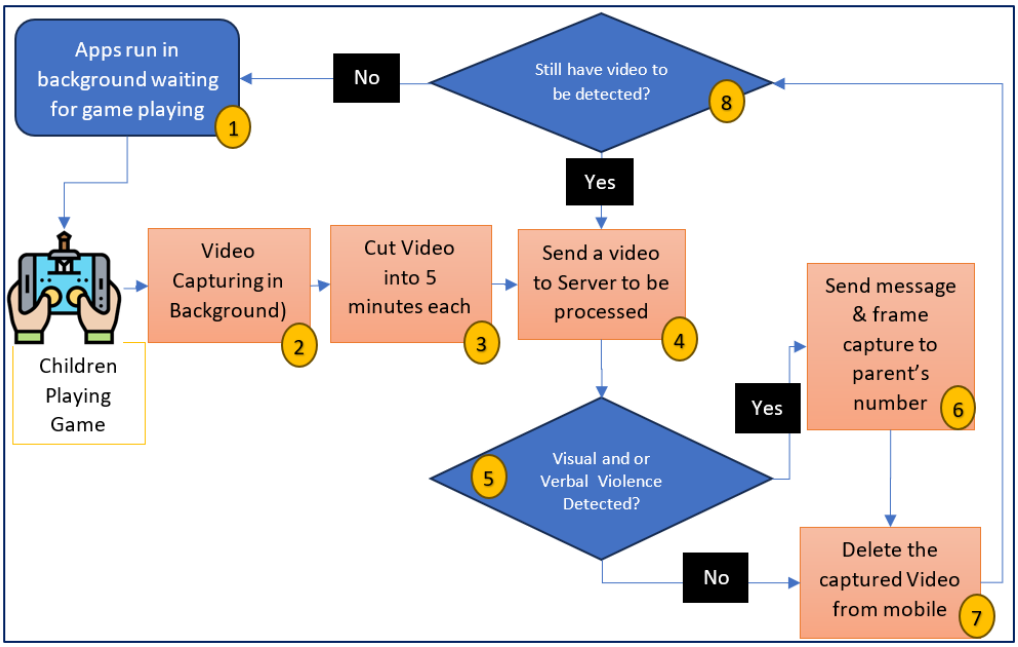

Figure 5. The flow diagram of the in-game early warning system, where the application runs in the background and captures the game session. Each video is cut into 5 minutes. Then, the videos are sent to the servers to check whether they contain visual or verbal violence or both. If found, the system will send the message to the registered parent's number, after all check is conducted, delete all the videos from the device.

To handle the multisession and scalability of the system, each video is indexed with session identifiers and timestamps for multiple sessions, enabling parallel uploads and efficient retrieval. In practice, only flagged frames or short excerpts are retained for caregiver notification, while non-relevant clips are discarded, reducing storage demands. Regarding scalability, the architecture can be extended to cloud-based infrastructures where hundreds or thousands of players can be supported simultaneously. This can be achieved through load balancing, distributed storage, and container-based deployment (e.g., Docker/Kubernetes). As video clips are as short as 5 minutes and processed independently, the system can allocate resources dynamically, ensuring timely detection and alerting without bottlenecks.

4. Results

For the visual modality, the model that combines 3DCNN, BiLSTM, and Attention Mechanism for visual violence detection is tested with the Hockey Fight dataset, Violent Movie Dataset, and Game Online dataset. It got an average accuracy of 99.14%. The best result from the Game Online dataset is 97.84% in accuracy, precision, recall, and F1 score. [23].

The visual modality achieves the best results from training with the Game Online dataset using the K-Fold method, with k=5. The experiments were done multiple times using epochs

10 (early stopping is used when there is no significant change in accuracy) and batch sizes 8-16. The proposed model achieves performance accuracy of 97.84%, precision of 97.84%, recall of 97.84%, and F1 score of 97.84%. The Metric evaluation is shown in Figure 6, and the confusion matrix of this result is shown in Figure 7.

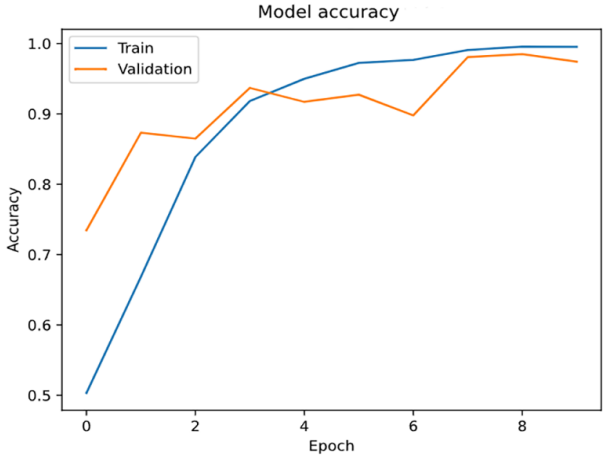

Figure 6. The Average Accuracy Metrics of the Proposed Model for Visual Violence Detection show that the proposed model achieves 97.84% trained with the Online Game Dataset.

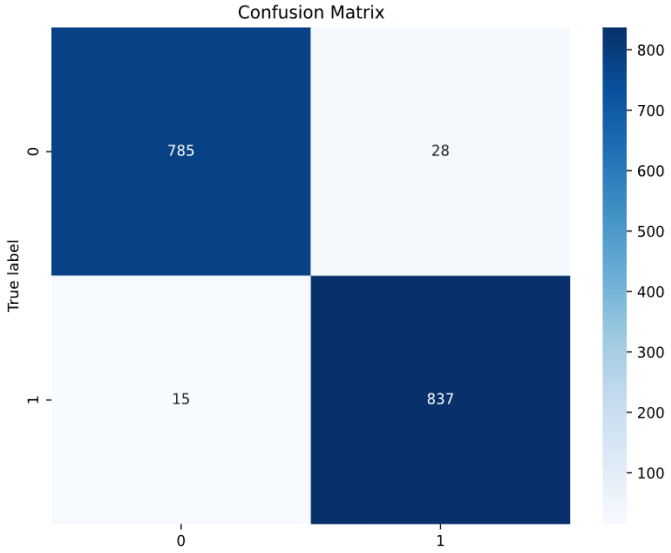

Figure 7. The Confusion Matrix of the proposed model for Visual Violence Detection for two classes: 0 is non-violent and 1 is violent.

The verbal modality result was trained with the Indonesian Chat dataset using the stratified K-Fold method, with k=5, and it balances the distribution of each class for each fold. The experiments were done multiple times using epochs 5-10 (early stopping is used when there is no significant change in accuracy) and batch sizes of 8 and 16. The accuracy, precision, recall, and F1 score are shown in Table 4.

Table 4. Accuracy, Precision, Recall, and F1 Score (in percentage) for the Indonesian Chat dataset with various batch parameters

Batch	Fold	Accuracy	Precision	Recall	F1 Score
8	1	81.83	81.82	81.83	81.63
8	2	94.16	94.22	94.16	94.17
8	3	98.08	98.10	98.08	98.09
8	4	98.64	98.65	98.64	98.65
8	5	98.79	98.79	98.79	98.79
AVG		**94.30**	**94.32**	**94.30**	**94.26**
16	1	81.64	81.84	81.64	81.69
16	2	90.71	91.06	90.71	90.71
16	3	94.58	94.69	94.58	94.59
16	4	96.96	96.97	96.96	96.96
16	5	97.38	97.39	97.38	97.39
AVG		92.25	92.39	92.25	92.27

Table 3 shows that the best result is achieved using a batch size of 8. The accuracy is 94.30%, the precision is 94.32%, the recall is 94.30%, and the F1 score is 94.26%. Figure 8 shows the accuracy metrics of all of the folds.

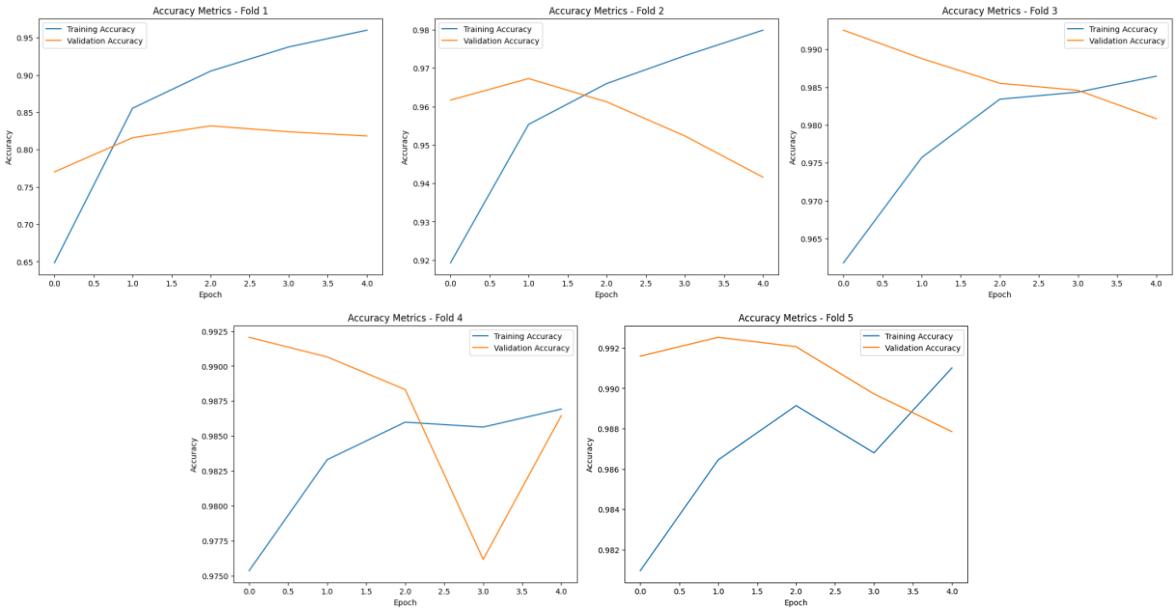

Figure 8. The proposed model's Average Accuracy Metrics for verbal violence detection has an average 94.30% accuracy when trained with the Indonesian Chat Dataset. All five folds results are presented.

Figure 9 shows the confusion matrix from the Fold 5 training result for four classes resulting from verbal violence detection: neutral, violent, racist, and harassment. Some misclassified words are ambiguous sentences; for example, "Edan, golnya Fabiano," which means "Fabiano's goal, crazy." This sentence is often recognized as ambiguous. Some say this sentence is neutral;

it's just a euphoric comment indicating the goal is amazingly crazy. However, others perceive this sentence as referring to Fabiano's craziness, which is considered violent.

The hybrid late fusion used in the multimodal approach can reduce this misclassification. It decreases false positives in ambiguous verbal cases like this euphoric word, which was classified as violent, while also improving recall for visually subtle acts of violence that rules alone would miss. These findings indicate that the combined strengths of deterministic rules and softmax weighting lead to more balanced and reliable multimodal detection.

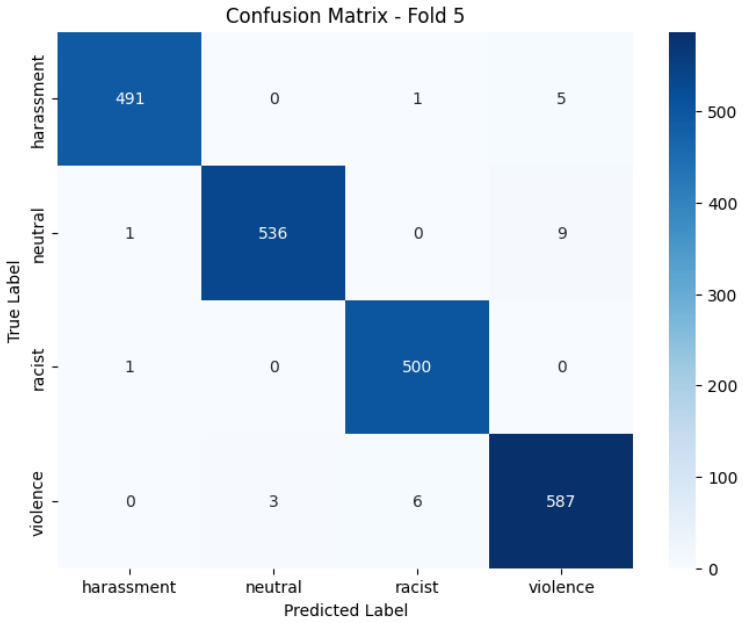

Figure 9. The Confusion Matrix of the proposed model for Verbal Violence Detection for four classes: Neutral, Violence, Racist, and Harassment.

Some video tests were conducted, including 15-minute videos containing visual and verbal violence content. One of the video frames is shown in Figure 10. In that frame, verbal violence is detected: "Dasar cupu lo semua," which means "you all are stupid." This chat should be identified as harassment because it derogates another player as stupid or a noob. In another frame, a fire burns the character, which can be visually recognized as violence since the model detects fire and weapon shooting as violent actions. These frames occur within five minutes of the captured video, so the final multimodal result should be classified as visual-verbal violence.

Figure 10. Example of a 5-minute video frame containing visual and verbal violence in some frames.

For a system that can be embedded in-game, the output is a text message that determines whether the five minutes of video captured from the game session contain violence visually or verbally. After explaining the result of each modality, the message also shows the final result of the video. An example message of the system's result is shown in Figure 11.

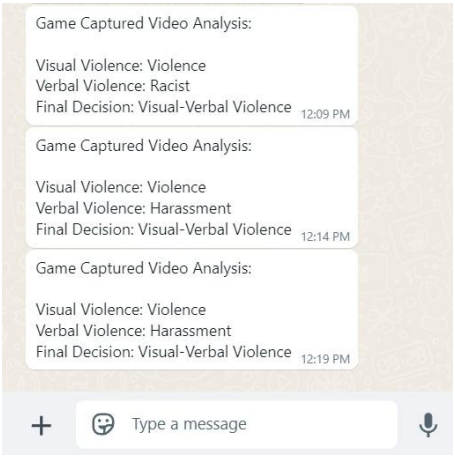

Figure 11. The Example Message was sent due to Multimodal Violence Detection conducted on the server. This message can be an early warning system for parents, showing what happened visually and verbally.

5. Discussion

This study analyzed the process and results of the verbal modality. The tests show that training a model with a combination of BERT and BiLSTM, using batch sizes of 16 over 10 epochs, produces the best results, with an accuracy of 98.27%. This indicates that the model effectively detects violent sentences in Bahasa Indonesia, accurately classifying them as neutral (non-violent), violent, racist, or harassing. In contrast, the combination of 3DCNN, BiLSTM, and attention mechanisms achieves an accuracy of 99.14% for the visual modality. The high accuracy demonstrates that the model can identify most instances of violence in videos [23].

The proposed visual violence detection model is also compared to several state-of-the-art models, such as the Full Temporal Cross-Fusion Network (FCTF Net) [25], 3D Convolutional Neural Network (3DCNN) [27]-[28], VD-Net [32], and the combination of CNN and Bidirectional LSTM (BiLSTM)[16]. Each model is trained with three datasets: the Hockey Fight Dataset, the Violent Movies Dataset, and the Online Game Dataset. The full results and discussion are published in [23]. Table 5 shows the performance comparison.

Table 5. The performance comparison of the proposed model's accuracy (in percentage) with several other models was performed when trained with three datasets (Hockey Fight Dataset, Violent Movies Dataset, and Online Game Dataset).

Model	Hockey Fight Dataset	Violent Movies Dataset	Online Game Dataset	Average
FCTF Net [25]	99.50	100.00	92.55	97.35
3DCNN [27]-[28]	96.00	100.00	93.35	96.45
VD-Net [32]	98.50	99.00	92.25	96.58
CNN-BiLSTM [16]	99.27	100.00	86.70	95.32
Proposed Model (3DCNN-BiLSTM-Att)	**99.58**	**100.00**	**97.84**	**99.14**

A comparison of the proposed model (BERT-BiLSTM) with some other latest models is shown in Table 6; each model is trained with the Indonesian Data Set used in this study, with a batch size of 16 and a maximum epoch of 10. The compared models are the CNN-GRU [17], SSA-DBN [18], BiLSTM [20], Transformer XLM-R [21], and Bi-Gru attention with the BERT model [22]. According to Table 5, the proposed model BERT-BiLSTM surpasses other models in accuracy, precision, recall, and F1 Score.

Table 6. The comparison of Accuracy, Precision, Recall, and F1 Score (in percentage) of the proposed model with several other models when trained with the Indonesian Chat dataset.

Model	Accuracy	Precision	Recall	F1 Score
CNN-GRU [17]	80.20	80.35	80.20	80.25
SSA-DBN [18]	93.80	91.20	93.20	92.18
BiLSTM [20]	78.84	79.60	78.84	79.05
Transformer-XLM-R [21]	82.61	82.00	84.00	83.00
Bi-Gru Att-BERT [22]	92.12	92.73	92.45	92.07
Proposed Model (BERT-BiLSTM)	**94.30**	**94.32**	**94.30**	**94.26**

When trained on other datasets or languages, the proposed BERT-BiLSTM model would likely sustain its strong performance trend due to BERT's contextualized embeddings, which are transferable across domains. However, the level of effectiveness would depend on the linguistic and cultural traits of the new dataset. For example, languages with richer morphology, different slang patterns, or code-switching phenomena may pose challenges that require additional pretraining or domain adaptation. Therefore, while the architecture is expected to generalize well, retraining or fine-tuning with representative samples from the target domain would be necessary to maintain robustness and ensure similar accuracy, precision, recall, and F1 scores.

Implementing hybrid late fusion in the multimodal approach combines rule-based and softmax probability fusion. It can detect visual and chat-based violence, racism, and harassment. Based on a 5-minute video input, the test can determine whether the final results are non-violent, visual, verbal, or both visual and verbal. Consequently, the average accuracy of this multimodal violence detection technique is 96.72%.

This study also demonstrates how the multimodal violence detection system is structured and how using this system can help parents understand what their children encounter while playing child-friendly online games. The proposed hybrid late fusion combines rule-based decision logic with probabilistic weighting, addressing the limitations of traditional fusion methods. While purely rule-based approaches are understandable, they can be too rigid in ambiguous situations. Conversely, strictly probabilistic fusion, such as softmax-based weighting, offers flexibility but may lead to misclassification if one modality dominates. Combining both, the strategy uses the clarity and reliability of deterministic rules for straightforward cases and softmax probabilities to resolve uncertain or conflicting predictions between visual and verbal modalities. This approach enhances robustness in multimodal violence detection by maintaining a balance between interpretability and flexibility, providing a practical improvement over existing late fusion techniques. It can help promote a more child-friendly online gaming environment, not just in game ratings but also in actual gameplay experiences.

The limitation is that the visual modality still cannot detect multiple types of violence. However, if it can identify violence such as blood, gore, fire, and melee, it will improve the model's performance.

In verbal modality, the model was only trained and tested in Bahasa Indonesia and needs to be evaluated in other languages. Future work could expand the model to include other widely spoken languages like English, Spanish, and French. This multilingual extension would improve the system's usefulness across different cultural contexts and boost its potential to create safer online gaming environments worldwide. Additionally, the verbal violence detection model has not been tested to identify new types of violent language that may have emerged as language evolves among different generations. Therefore, a study to predict new kinds of sentences used by younger people would be very helpful in protecting more youngsters from toxic online gaming communities.

Multimodal fusion strategies can be explored further for future research. Early fusion, which concatenates modality-specific features before joint modeling, can enable the network to capture cross-modal dependencies, for example, aligning visual cues with verbal intent. In contrast, late fusion integrates modality-specific predictions and is often more robust to missing or noisy data. Employing a denser fusion network at the late stage may further capture higher-order interactions between modalities, though it introduces greater model complexity and a potential risk of overfitting. Beyond fusion design, future work could also investigate additional modalities, such as audio or avatar expressions, and expand the scope of classification to create better categories.

The messages sent to caregivers can also be improved for the early warning system, which serves as the proof of concept. Instead of just a simple alert, the system might include the type of aggression detected, whether visual or verbal, the exact timestamp in the session, a severity level (low, medium, high), and contextual suggestions such as "Encourage rest time" or "Monitor further interactions with the player with a certain name." These details provide actionable insights and enhance the system's practicality for caregivers.

6. Conclusions

In this study, the multimodal approach is divided into two modalities. The visual and verbal modalities are considered because violence in online games typically occurs visually and can be observed by the player. Additionally, verbal violence often takes place through game chat. The visual modality, which combines 3DCNN, BiLSTM, and attention mechanisms, achieves an accuracy of 99.14%, while the verbal modality, using BERT and BiLSTM, reaches 94.30%. These modalities are processed separately and then fused using hybrid late fusion. This hybrid approach integrates rule-based methods and softmax probabilities; the multimodal approach attains an overall accuracy of 96.72% by averaging the accuracies of the visual and verbal modalities.

This model can also be integrated into any child-friendly online game to detect violence from those two modalities using the proposed system, serving as an early warning tool for parents. This technology is expected to reduce parental worries about the online games their children play and help them take necessary actions if violence occurs during gameplay.

Despite these promising results, several limitations remain. The current framework is restricted to only two modalities and three types of violence, which may not capture the full spectrum of harmful behaviors in online games. Furthermore, the system has been evaluated under controlled conditions. It has not yet been tested for robustness in noisy or real-world scenarios, such as evolving slang in chat or low-quality video streams.

Future research should therefore focus on expanding the detectable classes, integrating additional modalities such as audio or physiological signals, and testing across more diverse datasets to ensure adaptability and robustness. Addressing these aspects could make the

multimodal-based approach more comprehensive, effective, and suitable for deployment in child-friendly online games. Ultimately, the proposed system holds the potential to serve as an early warning tool for parents, helping them monitor gameplay and take necessary actions when violence occurs.

Acknowledgments

We would like to express our heartfelt gratitude to the supervisors for their guidance and unwavering support during this research, and to all the students who helped collect data to form the Online Game Dataset and Indonesian Chat Dataset. We also extend our heartfelt thanks to the reviewers for their helpful insights and constructive comments that improved this paper.

Conflicts of interest

The authors declare no conflict of interest.

References

[1] Moreno-López, R., & Argüello-Gutiérrez, C. "Violence, Hate Speech, and Discrimination in Video Games: A Systematic Review," *Social Inclusion*, vol. 13, Article 9401, 2025. doi: 10.17645/si.9401.

[2] H. Duan, Y. Huang, Y. Zhao, Z. Huang, and W. Cai, "User-Generated Content and Editors in Video Games: Survey and Vision," in *IEEE Conference on Computational Intelligence and Games, CIG*, IEEE Computer Society, 2022, pp. 536–543. doi: 10.1109/CoG51982.2022.9893717.

[3] 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.

[4] P. 'asher' Rospigliosi, "Metaverse or Simulacra? Roblox, Minecraft, Meta, and the turn to virtual reality for education, socialisation, and work," *Interactive Learning Environments*, vol. 30, no. 1, 2022, doi: 10.1080/10494820.2022.2022899.

[5] 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, Jul. 2022, doi: 10.1016/j.compedu.2022.104504.

[6] Z. Yang, N. Grenon-Godbout, and R. Rabbany. "Game On, Hate Off: A Study of Toxicity in Online Multiplayer Environments," *ACM Games* vol. 2, no. 2, Article 14, pp. 1-13, 2024. doi: 10.1145/3675805.

[7] 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, Mar. 2021, doi: 10.1177/1741659019881040.

[8] 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, 2022, ch. 14, pp. 160–174. doi: 10.4018/978-1-6684-5907-2.ch014.

[9] T. G. Sharma, J. Hamari, A. Kesharwani, and P. Tak, "Understanding continuance intention to play online games: roles of self-expressiveness, self-congruity, self-efficacy, and perceived risk," *Behaviour and Information Technology*, vol. 41, no. 2, pp. 348–364, 2022, doi: 10.1080/0144929X.2020.1811770.

[10] 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, Feb. 2024, doi: 10.1007/s10796-023-10377-7.

[11] E. Lee, P. Schulz, and H. E. Lee. "The Impact of Violent Media Content and Knowledge of Viable Responses to Cyberviolence on Good Citizenship Behavior Among South Korean Adolescents," *Journal of Interpersonal Violence*, vol. 40, no. 19-20, pp. 4651-4685. doi: 10.1177/08862605241297377.

[12] 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, Apr. 2020, doi: 10.3390/s20072018.

[13] 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, Mar. 2020, doi: 10.1080/08839514.2020.1723876.

[14] S. Vosta and K.C. Yow, "KianNet: A Violence Detection Model Using an Attention-Based CNN-LSTM Structure," in *IEEE Access*, vol. 12, pp. 2198-2209, 2024, doi: 10.1109/ACCESS.2023.3339379.

[15] 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, Jul. 2021, doi: 10.3390/electronics10131601.

[16] R. Halder and R. Chatterjee, “CNN-BiLSTM Model for Violence Detection in Smart Surveillance,” *SN Computer Science*, vol. 1, no. 4, Jul. 2020, doi: 10.1007/s42979-020-00207-x.

[17] S. Bharti, A. K. Yadav, M. Kumar, and D. Yadav, “Cyberbullying detection from tweets using deep learning,” *Kybernetes*, vol. 51, no. 9, pp. 2695–2711, Sep. 2022, doi: 10.1108/K-01-2021-0061.

[18] R. Kumar and A. Bhat, “A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media,” *Int J Inf Secur*, vol. 21, no. 6, pp. 1409–1431, Dec. 2022, doi: 10.1007/s10207-022-00600-y.

[19] S. Neelakandan *et al.*, “Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media,” *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/2163458.

[20] M. S. Lekshmi, A. M. Shaji, S. K. Amrita. “Cyberbullying Detection Using BiLSTM Model.” In: *Gopi, E.S., Maheswaran, P. (eds) Proceedings of the International Conference on Machine Learning, Deep Learning and Computational Intelligence for Wireless Communication.* MDCWC 2023. Signals and Communication Technology. Springer, Cham. doi: 10.1007/978-3-031-47942-7_29.

[21] S. Sihab-Us-Sakib, Md. R. Rahman, Md. S. A. Forhad, and Md. A. Aziz, “Cyberbullying detection of resource-constrained language from social media using a transformer-based approach,” *Natural Language Processing Journal*, vol. 9, p. 100104, Dec. 2024, doi: 10.1016/j.nlp.2024.100104.

[22] M. K. Mali *et al.*, “Automatic detection of cyberbullying behaviour on social media using Stacked Bi-Gru attention with BERT model,” *Expert Syst Appl*, vol. 262, Mar. 2025, doi: 10.1016/j.eswa.2024.125641.

[23] J. Prestiliano, A. Azhari, and A. Nurwidyantoro, “Enhanced Deep Learning Model to Detect Violence and Gore in Child-Friendly Online Game,” *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 1, pp. 279–290, 2025, doi: 10.22266/ijies2025.0229.20.

[24] S. Elyasi, A. VarastehNezhad, and F. Taghiyareh, “From Play to Prediction: Assessing Depression and Anxiety in Players' Behavior with Machine Learning Models,” *International Journal of Serious Games*, vol. 12, no. 1, pp. 83–102, Feb. 2025, doi: 10.17083/ijsg.v12i1.897.

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

[26] M. S. Kang, R. H. Park, and H. M. Park, “Efficient Spatio-Temporal Modeling Methods for Real-Time Violence Recognition,” *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3083273.

[27] J. Mahmoodi, H. Nezamabadi-pour, and D. Abbasi-Moghadam, “Violence detection in videos using interest frame extraction and 3D convolutional neural network,” *Multimed Tools Appl*, vol. 81, no. 15, pp. 20945–20961, Jun. 2022, doi: 10.1007/s11042-022-12532-9.

[28] 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, Jun. 2019, doi: 10.3390/s19112472.

[29] 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 Conference of Computer Sciences*, 2019. doi: 10.1109/SCCS.2019.8852616.

[30] F. J. Rendón-Segador, J. A. Álvarez-García, J. L. Salazar-González, and T. Tommasi, "CrimeNet: Neural Structured Learning using Vision Transformer for violence detection," *Neural Networks*, vol. 161, pp. 318–329, Apr. 2023, doi: 10.1016/j.neunet.2023.01.048.

[31] 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, Jan. 2023, doi: 10.1109/tbiom.2022.3233399.

[32] 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.

[33] 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, Jan. 2024, doi: 10.3390/s24020317.

[34] U. Naseem, S. Shiwakoti, S. B. Shah, S. Thapa, and Q. Zhang. "GameTox: A Comprehensive Dataset and Analysis for Enhanced Toxicity Detection in Online Gaming Communities," In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 440–447, Albuquerque, New Mexico. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-short.37.

[35] B. M. Peixoto, B. Lavi, Z. Dias, and A. Rocha, "Harnessing high-level concepts, visual, and auditory features for violence detection in videos," *J Vis Commun Image Represent*, vol. 78, Jul. 2021, doi: 10.1016/j.jvcir.2021.103174.

[36] B. M. Peixoto, B. Lavi, P. Bestagini, Z. Dias, and A. Rocha, "Multimodal Violence Detection in Videos," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2020. doi: 10.1109/ICASSP40776.2020.9054018.

[37] P. Wu, J. Liu, Y. Sun, F. Shao, Z. Wu and Z. Yang. "Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision," in *Computer Vision – European Conference on Computer Vision (ECCV) 2020*, 2020. doi: 10.1007/978-3-030-58577-8_20.

[38] Y. Pu, X. Wu, S. Wang, Y. Huang, Z. Liu, and C. Gu, "Semantic multimodal violence detection based on local-to-global embedding," *Neurocomputing*, vol. 514, pp. 148–161, Dec. 2022, doi: 10.1016/j.neucom.2022.09.090.

[39] C. Gu, X. Wu, and S. Wang, "Violent Video Detection Based on Semantic Correspondence," *IEEE Access*, vol. 8, pp. 85958–85967, 2020, doi: 10.1109/ACCESS.2020.2992617.

[40] S. Paul, S. Saha, and M. Hasanuzzaman, "Identification of cyberbullying: A deep learning based multimodal approach," *Multimed Tools Appl*, vol. 81, no. 19, pp. 26989–27008, Aug. 2022, doi: 10.1007/s11042-020-09631-w.

[41] F. Wu, B. Gao, X. Pan, L. Li, Y. Ma, S. Liu, and Z. Liu. "Fuser: An enhanced multimodal fusion framework with congruent reinforced perceptron for hateful memes detection," *Information Process Management*, vol. 61, no. 4, Jul. 2024, doi: 10.1016/j.ipm.2024.103772.

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

[43] 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.

[44] G. C. Dobre, M. Gillies, and X. Pan, "Immersive machine learning for social attitude detection in virtual reality narrative games," *Virtual Reality*, vol. 26, no. 4, pp. 1519–1538, Dec. 2022, doi: 10.1007/s10055-022-00644-4.

[45] 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: Conference Series*, IOP Publishing Ltd, Jan. 2021. doi: https://doi.org/10.1088/1742-6596/1717/1/012055.

[46] A. P. J. Dwitama, D. H. Fudholi, and S. Hidayat, "Indonesian Hate Speech Detection Using Bidirectional Long Short-Term Memory (Bi-LSTM)," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 2, pp. 302–309, Mar. 2023, doi: 10.29207/resti.v7i2.4642.

[47] A. Azhari, A. Santoso, A. A. P. Ratna, and J. Prestiliano, "Optimization of AES using BERT and BiLSTM for Grading the Online Exams," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 5, pp. 395–411, 2024, doi: 10.22266/ijies2024.1031.31.

[48] J. Xie, B. Chen, X. Gu, F. Liang, and X. Xu, "Self-Attention-Based BiLSTM Model for Short Text Fine-Grained Sentiment Classification," *IEEE Access*, vol. 7, pp. 180558–180570, 2019, doi: 10.1109/ACCESS.2019.2957510.

[49] J. S. R. Dinesh *et al.*, "Real-time violence detection framework for football stadium comprising big data analysis and deep learning through bidirectional LSTM," *Computer Networks*, vol. 151, pp. 191–200, Mar. 2019, doi: 10.1016/j.comnet.2019.01.028.

[50] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs Late Fusion in Multimodal Convolutional Neural Networks," in *IEEE 23rd International Conference on Information Fusion (FUSION)*, 2020, pp. 1–6. doi: 10.23919/FUSION45008.2020.9190246.

[51] X. Chen, "MMRBN: Rule-Based Network for Multimodal Emotion Recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 8200–8204. doi: 10.1109/ICASSP48485.2024.10447930.