PHYSICAL VIOLENCE DETECTION SYSTEM TO PREVENT STUDENT MENTAL HEALTH DISORDERS BASED ON DEEP LEARNING

Sukmawati Anggraeni Putri[1*)] ; Achmad Rifai[2]; Imam Nawawi[3]

Sistem Informasi
Universitas Nusa Mandiri[1,2]
Jakarta, Indonesia
sukmawati@nusamandiri.ac.id, achmad.acf@nusamandiri.ac.id

Sistem Informasi
Universitas Bina Sarana Informatika[3]
Jakarta, Indonesia
imam.imw@bsi.ac.id

(*) Corresponding Author

Abstract— Physical violence in the educational environment by students often occurs and leads to criminal acts. Apart from that, repeated acts of physical violence can be considered non-verbal bullying. This bullying can hurt the victim, causing physical disorders, mental health, impaired social relationships, and decreased academic performance. However, monitoring activities against acts of violence currently being carried out have weaknesses, namely weak supervision by the school. A deep Learning-based physical violence detection system, namely LSTM Network, is the solution to this problem. In this research, we develop a Convolutional Neural Network to detect acts of violence. Convolutional Neural Network extracts features at the frame level from videos. At the frame level, the feature uses long short-term memory in the convolutional gate. Convolutional Neural Networks and convolutional short-term memory can capture local spatio-temporal features, enabling local video motion analysis. The performance of the proposed feature extraction pipeline is evaluated on standard benchmark datasets in terms of recognition accuracy. A comparison of the results obtained with state-of-the-art techniques reveals the promising capabilities of the proposed method for recognizing violent videos. The model that has been trained and tested will be integrated into a violence detection system, which can provide ease and speed in detecting acts of violence that occur in the school environment.

Keywords: Violence Detection, Deep Learning, Bullying, Long Short-Term Memory, Student

Abstrak— *Tindak kekerasan fisik dalam lingkungan pendidikan yang dilakukan oleh siswa sering terjadi dan berujung pada tindakan kriminal. Selain itu, tindak kerasan fisik yang dilakukan berulang dapat dikatakan sebagai tindak bullying verbal. Bullying ini dapat memberikan dampak negative bagi korban, seperti gangguan fisik, kesehatan mental, gangguan hubungan sosial dan performa akademik yang menurun. Namun kegiatan pengawasan terhadap tindak kekerasan yang dilakukan saat ini mempunyai kelemahan yaitu lemahnya pengawasan dari pihak sekolah. Sistem pendeteksi kekerasan fisik berbasis Deep Learning yaitu LSTM Network sebagai solusi permasalahan tersebut. Dalam penelitian ini, kami mengembangkan Convolutional Neural Network untuk mendeteksi tindakan kekerasan. Convolutional Neural Network digunakan untuk mengekstrak fitur pada level frame dari video. Pada level frame, fitur menggunakan long short-term memory di convolutional gate. Convolutional Neural Network dan convolutional short-term memory dapat menangkap fitur local spatio-temporal features, memungkinkan analisis gerakan video lokal. Performa pipeline ekstraksi fitur yang diusulkan dievaluasi pada kumpulan data benchmark standar dalam hal akurasi pengenalan. Perbandingan hasil yang diperoleh dengan teknik mutakhir mengungkapkan kemampuan yang menjanjikan dari metode yang diusulkan untuk mengenali video kekerasan. Model yang telah dilatih dan diuji akan diintegrasikan pada sistem deteksi kekerasan yang dapat memberikan kemudahan dan kecepatan medeteksi tindak kekerasan yang terjadi dilingkungan sekolah.*

Kata Kunci*: Deteksi Kekerasan, Deep Learning, Perundungan, Long Short-Term Memory, Siswa.*

INTRODUCTION

In recent years, there has been an increase in the frequency of acts of violence between students in educational environments, which harms victims physically and psychologically. According to

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 19, No. 2 September 2023 | DOI: 10.33480/pilar.v19i2.4600

Herwina, violence against children in the school environment includes all actions that result in physical, mental, sexual, or psychological harm, including torture (Bahar, 2013). Physical acts of violence carried out repeatedly can be said to be non-verbal acts of intimidation. As reported on the tempo.co news portal on August 4, 2023, the Federation of Indonesian Teachers' Unions (FSGI) reported 16 cases of bullying and 43 victims, consisting of 41 students and two teachers, during January – July 2023. There was physical violence, namely being kicked by an upperclassman. The increasing number of bullying cases has made the educational environment an emergency. Based on these conditions, this research analyzing videos from surveillance cameras using artificial intelligence technology (Roshan et al., 2020) can help detect acts of violence, reducing bullying cases that occur in the school environment.

In this research, in building artificial intelligence to detect violent acts, we propose using a deep learning methodology, namely convolutional neural networks, for binary classification, differentiating violent and non-violent content in the video analysis process (Ramzan et al., 2019)(Gkountakos et al., 2020). Additionally, *video* can be defined as a compilation of individual frames that transition quickly to create the illusion of movement. This methodology begins by extracting frames from the image and categorizing them as images depicting violence or non-violence (Gkountakos et al., 2020).

Convolutional Neural Networks (CNNs) have been empirically proven to be a very efficient model for analyzing image material (Honarjoo et al., 2021a). The main reasons for these results are the application of methods to expand the network to cover many parameters and the availability of extensive labeled datasets capable of facilitating the learning process (Bianculli et al., 2020). Convolutional Neural Networks (CNN) have demonstrated the ability to obtain robust and easy-to-understand visual characteristics in certain situations. Convolutional Neural Networks (CNN) are equipped with the ability to exploit the inherent visual features of individual stationary images and the complex temporal dynamics exhibited by these images (Dinesh Jackson et al., 2019).

Recently, several models have been created that utilize long short-term memory (LSTM) recurrent neural networks (RNNs) (Sudhakaran & Lanz, 2017) to address problems related to feature aggregation. Sharma et al. (Sharma et al., 2021) have presented a violence detection method that uses this model. The methodology involves leveraging convolutional neural networks to extract features from raw pixels, optical flow images, and acceleration flow maps. This is followed by coding using short-term memory models (LSTM) and final fusion techniques.

In a recent study by Majd et al. (Majd & Safabakhsh, 2020), the authors modified the short-term memory model (LSTM) by replacing the fully connected gate layer with a convolutional layer called convolutional LSTM. Convolutional LSTM (convLSTM) is a newer iteration of LSTM models. The convLSTM model encodes spatiotemporal information in its memory cells by replacing fully connected layers in the LSTM with convolutional layers. After the ConvLSTM model performs well, it is integrated with a violence detection system. This makes it easier for users to analyze videos to determine whether there are violent elements.

MATERIALS AND METHODS

This research aims to create a comprehensive deep neural network model that can be trained from start to finish to categorize videos as violent or non-violent accurately. The model that has been built will be integrated into the Violence Detection System. The training process of the proposed model is presented in Figure 1. In the proposed model, the network consists of convolutional layers, followed by a max pooling operation. This operation is used to extract discriminant features. Additionally, the network uses convolutional short-term memory (convLSTM) to encode changes within the video frame level. The convLSTM model for detecting acts of violence will be integrated with the system or application for acts of violence described in Figure 2.

Convolutional LSTM (convLSTM)

Video is a series of sequential images. To distinguish the occurrence of acts of physical violence between individuals depicted in videos, the system must be able to accurately detect and track the spatial position of the people involved and understand temporal variations in their movements. Convolutional neural networks (CNN) can produce high-quality representations of individual frames in video (Shi et al., 2015). Recurrent neural networks (RNNs) must be coded in the process of temporal changes. In general-purpose sequence modeling, LSTM as a special RNN structure has proven stable and robust for modeling long-term dependencies in previous studies (Pei et al., 2019). Considering our aim to examine changes in spatial and temporal dimensions, the convLSTM model emerged as an appropriate choice.

ConvLSTM is a development of the Convolutional Neural Network model, which performs well in detecting spatial (object shape) from each video frame of acts of violence (Wu et al.,

2020). However, it has shortcomings when analyzing videos' temporal (object movement) (Patel, 2021). So, LSTM is proven stable and robust for modeling long-term dependence on temporal change processes (Soeleman et al., 2022). In contrast to LSTM, convLSTM can encode spatial and temporal changes by utilizing convolutional gates inherent in its architecture. ConvLSTM will produce a more accurate picture of the video being analyzed. The convLSTM model equation can be seen in equations 1-3.

$$i_1=\sigma\ (w^i_x * I_t + w^i_h * h_{t-1} + b^i) \quad \text{.....(1)}$$
$$\int_1=\sigma\ (w^{\int}_x * I_t + w^{\int}_h * h_{t-1} + b^{\int}) \quad \text{.....(2)}$$
$$h_i=o_t\tanh(c_1) \quad \text{.....(3)}$$

In the preceding equations, '*' denotes the convolution operation, and' represents the Hadamard product. In the case of convLSTM, the hidden state ht, the memory cell ct, and the gate activations it, ft, and not are all 3D tensors.

To classify a video as violent or non-violent, the system needs to possess the ability to encode localized spatial elements and track their temporal variations. The utilization of handcrafted characteristics has the potential to accomplish this task but at the expense of heightened computational complexity. Convolutional neural networks (CNNs) can produce spatial features that effectively discriminate between objects or patterns. However, current approaches typically utilize the components generated from the fully connected layers of the CNN for temporal encoding, often employing long short-term memory (LSTM) networks for this purpose (Majd & Safabakhsh, 2020). The output generated by the fully connected layers serves as a comprehensive representation of the entire image, encapsulating its global characteristics. Therefore, the current approaches cannot effectively encode the localized spatial changes.

Consequently, they employ techniques incorporating additional data streams, such as optical flow images (Dong et al., 2016), which leads to heightened computational intricacy. The relevance of convLSTM arises since it can encode the convolutional characteristics inherent in the CNN. Additionally, the convolutional gates within the convLSTM model are trained to encode the temporal variations of local regions. Thus, the entire network exhibits the ability to encode localized spatiotemporal characteristics.

Design Model and Applications

This research aims to build a violence detection system by integrating the convLSTM model as a violence detection model for which training has been carried out. The system development process consists of several stages, which are explained in Figure 1 and Figure 2.

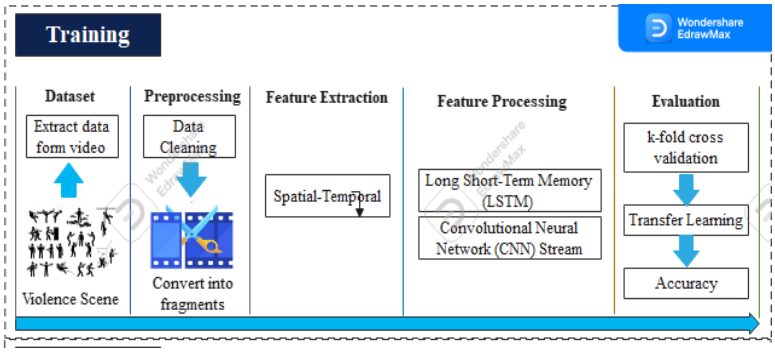

Figure 1. Training Process with Model

Figure 1 explains the training process for the designed model. The process begins by entering a dataset in the form of a video, then processing the video dataset, and then the feature extraction process spatially and temporally. Next is the training and test process with the convLSTM algorithm (CNN and LSTM). In the final approach, an evaluation is carried out to produce the percentage accuracy of the violence detection model from video analysis while training the dataset model used with public datasets. The following process is explained in Figure 2. After testing the model, integration of the physical violence detection system that has been developed is carried out. Through this system, videos tested on the system will be validated as containing elements of violence or non-violence.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 19, No. 2 September 2023 | DOI: 10.33480/pilar.v19i2.4600

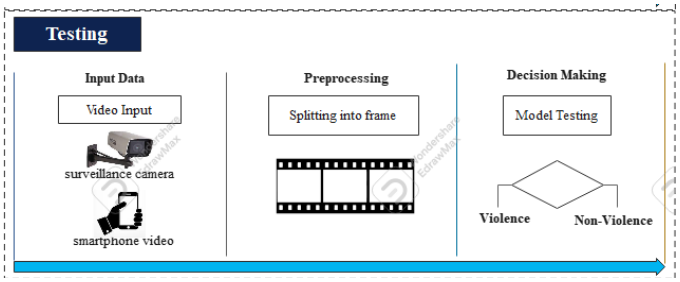

Figure 2. Testing Process with Application

RESULTS AND DISCUSSION

Performance evaluation was carried out with a 5-fold cross-validation scheme, widely used in the existing literature (Honarjoo et al., 2021b). Selecting a model architecture involves assessing the performance of several models using a data set, like a hockey battle (Ullah et al., 2019). In the hockey fight dataset, violence detection occurs in players' unusual movements, such as hitting, kicking, and pushing with both hands, to determine violations. Classification accuracy was achieved for two scenarios, where video and differences between frames were input.

As stated previously, this research conceptualizes violent behavior. One of the main challenges associated with applying this concept arises when applied to the field of sport. In the hockey fight dataset, fight videos depict players engaging in physical combat involving collisions and intentional contact with each other. One direct method for identifying violent scenes is to combine the proximity of one participant to another participant. However, non-violent videos also include instances where players engage in physical acts of affection, such as hugging or exchanging news, to celebrate. These videos could be mistakenly considered to depict violent content. However, the proposed methodology effectively circumvents this problem by demonstrating its ability to code the movements of specific parts, such as limbs, and the individuals' reactions. However, the proposed method should surpass previous state-of-the-art strategies in hardness datasets, achieving second place in accuracy. Additional research is needed to develop methodologies to reduce robustness associated with video density. One potential approach is to partition the framework into different subsections and independently estimate the output of each region. The video is then classified as violent if the network designates one of the regions as violent.

In previous research (Patel, 2021), LSTM, part of the RNN structure, has produced a stable and robust performance in the long-term dependency modeling process. LSTM mainly performs as an accumulator of state information in its memory cells. LSTM has a controller gate as a parameterization function to access, write, and clear parameters—accumulating information into cells through activating input gates. Additionally, the previous cell process ct-1 can be "forgotten" in this process if the forget gate ft is active. The advantage of using memory cells and gates is that they control the information flow. This process allows gradients to be trapped within cells in what is known as a constant error carousel and prevents too rapid a loss of information. This problem is a critical problem of LSTM. Apart from handling temporal correlation, there is also redundancy in spatial data. So, it is necessary to expand LSTM by exploiting the advantages of the convolutional structure used in input-to-state and state-to-state transitions. By stacking multiple ConvLSTM layers and forming a violence detection coding structure, we can build a network model for more general spatiotemporal problems.

To evaluate the superiority of convLSTM over regular LSTM, an alternative model incorporating LSTM is trained and tested using the hockey fights dataset. The novel framework comprises the VGG16 architecture, subsequently augmented with an LSTM RNN layer. The output of the final fully connected layer, fc7, in the VGG16 architecture is utilized as the input for a Long Short-Term Memory (LSTM) network of 1000 units. The remaining architectural components exhibit similarities to the convLSTM-based approach. The outcomes derived from this model and the corresponding count of trainable parameters are contrasted with the suggested model in Table 1. The table presented demonstrates the benefits of employing convLSTM instead of LSTM, highlighting the convLSTM's capacity for generating valuable video representations. Notably, the convLSTM requires significantly fewer parameters to be optimized than LSTM, with a difference of 77.5 million versus 72 million parameters. This phenomenon aids in enhancing the network's ability to generalize effectively, particularly when confronted with a scarcity of data while mitigating the risk of overfitting. The resulting model can process 30 frames per second using an NVIDIA K80s, T4s, P4s, and P100s GPU with 13GB RAM.

Table 1. Accuracy and parameter results from the ConvLSTM and LSTM models

Model	Accuracy	No. of Parameter
convLSTM	93,4%	77.5M
LSTM	92 %	72 M

Once the model finishes training and testing, the subsequent phase involves using the model within an intelligent application designed to detect instances of violence. The program is designed to autonomously identify cases of violence and deliver notifications based on video footage seamlessly integrated into the application. The graphic depicts the intelligent application's visual representation and operational procedure in the violence detection system. The following figure explains the results of designing a physical violence detection system.

1. Physical Violence Reporting page

Figure 3 is a violence reporting page where reporters can upload videos that indicate violence has occurred. On this page, the detection of acts of violence in uploaded videos will be processed. In the process, it integrates violence detection models, deep learning methods, and intelligent applications.

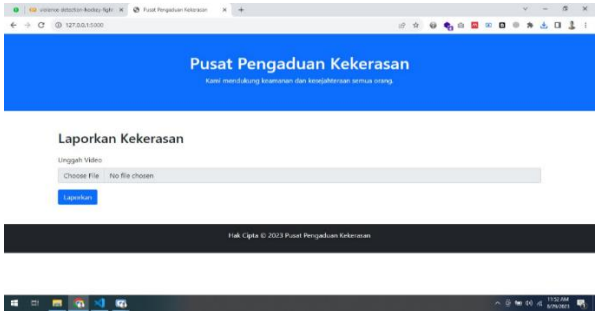

Figure 3. Physical Violence Reporting page

2. Violence Validation Process page

Figure 4 shows the results of detecting acts of violence. In this process, the model training and testing process is carried out, where uploaded videos are processed based on the results of training and testing models.

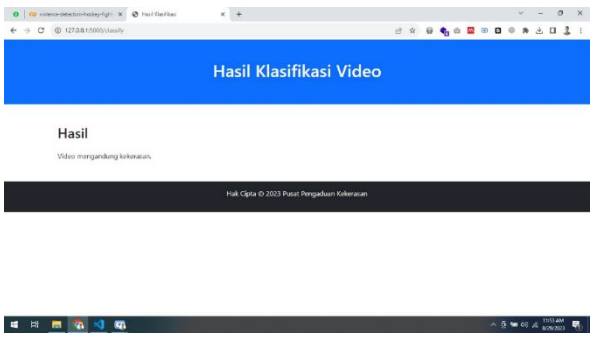

Figure 4. Violence Validation Process page

CONCLUSION

In this research, we apply machine learning approaches, specifically deep learning, to a challenging problem of violence detection that is still under development in utilizing state-of-the-art machine learning techniques. This research presents a new end-to-end deep neural network model that can be drilled to detect video violence. Convolutional neural networks (CNN) are used in the proposed model to extract framing-level features, followed by feature aggregation in the temporal domain using convolutional short-term memory (convLSTM). The proposed method is tested on three different datasets, and the results show that it performs better than state-of-the-art techniques. Additionally, it is shown that networks drilled to model changes in frames outperform networks prepared using frames as input. The findings of a comparative study between conventional fully connected LSTM and convLSTM show that the convLSTM model utilizes spatiotemporal data due to its convolutional structure, resulting in better video representation than LSTM with fewer parameters, thereby preventing overfitting. One idea is to add ConvLSTM to the spatial feature maps generated by a convolutional neural network and use the ConvLSTM hidden states for final classification. The ConvLSTM model produces an accuracy of 93%, which can be applied to general cases with video data characteristics that match the video data trained and tested on the ConvLSTM model. The convLSTM model produced can be applied and triggered in a physical violence detection information system used to detect physical violence in society to reduce bullying in the academic environment.

ACKNOWLEDGE

We want to thank the Directorate General of Higher Education, Research and Technology through the Directorate of Research, Technology and Community Service (DRTPM), Ministry Of Education, Culture, Research And Technology Of The Republic Of Indonesia, for providing grants and supporting the implementation of this research.

REFERENCE

Bahar, H. (2013). Pengembangan Pembelajaran Terpadu Dalam Pendidikan Karakter. *Jurnal Teknodik*, *17*(2), 209–225.

Bianculli, M., Falcionelli, N., Sernani, P., Tomassini, S., Contardo, P., Lombardi, M., & Dragoni, A. F. (2020). A dataset for automatic violence detection in videos. *Data in Brief*, *33*, 106587.

https://doi.org/10.1016/j.dib.2020.106587

Dinesh Jackson, S. R., Fenil, E., Gunasekaran, M., Vivekananda, G. N., Thanjaivadivel, T., Jeeva, S., & Ahilan, A. (2019). Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM. *Computer Networks*, *151*, 191–200. https://doi.org/10.1016/j.comnet.2019.01.028

Dong, Z., Qin, J., & Wang, Y. (2016). Multi-stream deep networks for person to person violence detection in videos. *Communications in Computer and Information Science*, *662*, 517–531. https://doi.org/10.1007/978-981-10-3002-4_43

Gkountakos, K., Ioannidis, K., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2020). A crowd analysis framework for detecting violence scenes. *ICMR 2020 - Proceedings of the 2020 International Conference on Multimedia Retrieval*, *June*, 276–280. https://doi.org/10.1145/3372278.3390725

Honarjoo, N., Abdari, A., & Mansouri, A. (2021a). Violence Detection Using One-Dimensional Convolutional Networks. *2021 12th International Conference on Information and Knowledge Technology, IKT 2021*, *December*, 188–191. https://doi.org/10.1109/IKT54664.2021.9685835

Honarjoo, N., Abdari, A., & Mansouri, A. (2021b). Violence detection using pre-trained models. *Proceedings of the 5th International Conference on Pattern Recognition and Image Analysis, IPRIA 2021*, *April*. https://doi.org/10.1109/IPRIA53572.2021.9483558

Majd, M., & Safabakhsh, R. (2020). Correlational Convolutional LSTM for human action recognition. *Neurocomputing*, *396*(xxxx), 224–229. https://doi.org/10.1016/j.neucom.2018.10.095

Patel, M. (2021). *Real-Time Violence Detection Using CNN-LSTM*. http://arxiv.org/abs/2107.07578

Pei, Z., Qi, X., Zhang, Y., Ma, M., & Yang, Y. H. (2019). Human trajectory prediction in crowded scene using social-affinity Long Short-Term Memory. *Pattern Recognition*, *93*, 273–282. https://doi.org/10.1016/j.patcog.2019.04.025

Ramzan, M., Abid, A., Khan, H. U., Awan, S. M., Ismail, A., Ahmed, M., Ilyas, M., & Mahmood, A. (2019). A Review on State-of-the-Art Violence Detection Techniques. *IEEE Access*, *7*, 107560–107575. https://doi.org/10.1109/ACCESS.2019.2932114

Roshan, S., Srivathsan, G., Deepak, K., & Chandrakala, S. (2020). Violence Detection in Automated Video Surveillance: Recent Trends and Comparative Studies. In *The Cognitive Approach in Cloud Computing and Internet of Things Technologies for Surveillance Tracking Systems*. INC. https://doi.org/10.1016/b978-0-12-816385-6.00011-8

Sharma, S., Sudharsan, B., Naraharisetti, S., Trehan, V., & Jayavel, K. (2021). A fully integrated violence detection system using CNN and LSTM. *International Journal of Electrical and Computer Engineering*, *11*(4), 3374–3380. https://doi.org/10.11591/ijece.v11i4.pp3374-3380

Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, *2015-Janua*(July), 802–810.

Soeleman, M. A., Supriyanto, C., Prabowo, D. P., & Andono, P. N. (2022). Video Violence Detection Using LSTM and Transformer Networks Through Grid Search-Based Hyperparameters Optimization. *International Journal of Safety and Security Engineering*, *12*(05), 615–622. https://doi.org/10.18280/ijsse.120510

Sudhakaran, S., & Lanz, O. (2017). Learning to detect violent videos using convolutional long short-term memory. *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017*. https://doi.org/10.1109/AVSS.2017.8078468

Ullah, F. U. M., Ullah, A., Muhammad, K., Haq, I. U., & Baik, S. W. (2019). Violence detection using spatiotemporal features with 3D convolutional neural network. *Sensors*, *19*(11), 1–15. https://doi.org/10.3390/s19112472

Wu, P., Liu, J., & Shen, F. (2020). A Deep One-Class Neural Network for Anomalous Event Detection in Complex Scenes. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(7), 2609–2622. https://doi.org/10.1109/TNNLS.2019.2933554