Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 12, No. 4, December 2024, pp. 768~779
ISSN: 2089-3272, DOI: 10.52549/ijeei.v12i4.5787

An Approach for Improving Accuracy and Optimizing Resource Usage for Violence Detection in Surveillance Cameras in IoT Systems

Hoang-Tu Vo, Phuc Pham Tien , Nhon Nguyen Thien , Kheo Chau Mui
Information Technology Department, FPT University, Can Tho 94000, Vietnam

Article Info

Article history:

Received Jul 29, 2024
Revised Sep 26, 2024
Accepted Oct 8, 2024

Keywords:

Violence detection
Surveillance camera
Fine-tune model
BiLSTM
Deep learning

ABSTRACT

Violence is a serious issue that can happen in many places, like streets, schools, or homes, where people hurt each other or damage things. To help prevent violence, some places use special cameras called surveillance cameras. These cameras watch over areas and look for signs of violence, like people fighting or breaking things. When they see something, they can send an alert to the police or others who can help. However, building models to detect violence in videos or surveillance cameras can be challenging. Current models may not be lightweight, fast, or use fewer resources, which means they may not work well on all devices or in all situations. So, there is a need for new models that are better at detecting violence while still being fast and using fewer resources. In this study, we focus on training multiple models to detect violence. Specifically, we fine-tune various models, including MobileNet, MobileNetV2, DenseNet121, and ResNet50V2, combined with Bidirectional Long Short-Term Memory (BiLSTM) networks. Among these models, MobileNetV2 for spatial feature extraction combined with BiLSTM for temporal feature extraction stands out for its compact size, quick processing time, and ability to achieve satisfactory results. This combination offers a lightweight solution that can efficiently detect violence in videos or surveillance footage while maintaining good performance levels in IoT systems.

Corresponding Author:

Hoang-Tu Vo
Information Technology Department, FPT University,
Can Tho 94000, Vietnam.
Email: tuvh6@fe.edu.vn

1. INTRODUCTION

Violence in society is a serious issue that affects individuals, families, and communities. It can take many forms, such as physical harm, verbal abuse, or emotional trauma [1], [2], [3], etc. Violence can occur in various settings, including homes, schools, workplaces, and public spaces. The effects of violence on individuals are profound and can include physical injuries, psychological distress, and long-term trauma. Families also suffer from the impact of violence, as it can lead to broken relationships, financial strain, and disrupted family dynamics. In addition, violence has broader societal consequences, contributing to social unrest, economic instability, and a breakdown of trust within communities.

Addressing the root causes of violence requires a multifaceted approach, including education, prevention programs, and access to support services for victims. Intelligent video surveillance systems are extensively utilized in various settings to enhance security and monitor activities effectively [4], [5], [6], etc. These systems employ advanced technologies such as artificial intelligence and machine learning to analyze video footage in real-time. By detecting anomalies, recognizing patterns, and identifying specific events or behaviors, intelligent surveillance systems can alert security personnel to potential threats or suspicious activities promptly. Additionally, these systems can automate tasks such as facial recognition, object tracking, and crowd monitoring, thereby improving efficiency and accuracy in surveillance operations. The widespread

adoption of intelligent video surveillance systems underscores their importance in safeguarding public spaces, businesses, and critical infrastructure, contributing to enhanced safety and security for individuals and communities alike.

In recent years, there have been significant advancements in recognizing human actions [7], [8], [9], etc, with particular emphasis on detecting violence, which remains one of the most complex areas of study within computer vision. Detecting instances of violence from surveillance cameras, whether in public or private settings, presents unique challenges. Early detection of violence from surveillance cameras faces several challenges beyond the need for human cooperation in monitoring violent incidents. One major difficulty lies in accurately distinguishing between normal and potentially violent behavior, as some actions may appear aggressive but are actually harmless. Additionally, variations in lighting conditions, camera angles, and image quality can impact the reliability of violence detection algorithms.

Moreover, the diversity of environments where surveillance cameras are deployed presents challenges in developing models that can adapt to different settings effectively. The primary objective of this study is to introduce a model that utilizes MobileNetV2 [10] for spatial feature extraction combined with BiLSTM [1] for temporal feature extraction, that maintains high performance comparable to the latest violence detection models while also reducing computational complexity.

This reduction in complexity is particularly crucial for deploying the model on devices with limited processing capabilities, such as smartphones or embedded systems. By developing a model that strikes a balance between performance and computational efficiency, this research aims to make violence detection more accessible and practical in real-world scenarios where resources are constrained. This approach ensures that the model can effectively identify violent behaviors while conserving computational resources, thereby enabling its deployment across a wide range of devices and applications.

The main contributions of this paper are as follows:

- The study centers on training multiple models specifically for detecting violence.
- The study fine-tunes several existing models, including MobileNet, MobileNetV2, DenseNet121, and ResNet50V2 combined with BiLSTM.
- Among the various models explored, the combination of MobileNetV2 and BiLSTM stands out.
- The proposed model provides a lightweight solution for efficiently detecting violence in videos or surveillance footage in IoT systems.

The structure of the study is as follows: In Part 2, we present a thorough related work. Part 3 outlines the overall methodology for detecting violence in surveillance video, including details about the dataset, data preparation, and evaluation metrics for the model. Moving on to Part 4, we delve into the experimental system and present the final results. Finally, Part 5 summarizes our study's findings and offers concluding remarks.

2. RELATED WORK

In recent years, the integration of diverse artificial intelligence (AI) techniques within computer vision has significantly advanced the capability to recognize and analyze violent behaviors depicted in video datasets. This amalgamation of AI methodologies has revolutionized the field, allowing for more nuanced detection of aggression and harmful actions in visual data. Moreover, these advancements have facilitated the development of complex algorithms capable of identifying violent activities, contributing to enhanced accuracy and efficiency in violence detection systems. The action recognition field primarily focuses on simple actions like clapping or walking, with less attention given to detecting fights or aggressive behaviors, which are crucial in settings such as prisons or elderly centers. To address this gap, the authors Bermejo Nievas, Enrique, et al. in [12] evaluates the Bag-of-Words framework along with STIP and MoSIFT descriptors for fight detection, introducing a new database with 1000 sequences to facilitate research on violence detection, achieving nearly 90% accuracy in detecting fights. This paper [13] presents an algorithm for detecting violent scenes in movies, decomposing the task into action scene and bloody frame detection. By analyzing the semantic-complete scene structure, the algorithm extracts features from segmented scenes and uses SVM classification, integrating face, blood, and motion information to determine violent content, yielding promising results in detecting most violent scenes. In this study [14] introduces two main contributions: a novel feature extraction method called Oriented Violent Flows (OViF) for violence detection in videos, which outperforms baseline approaches on public databases, and the adoption of feature combination and multi-classifier strategies, achieving superior performance with AdaBoost+Linear-SVM compared to existing methods on the Violent-Flows benchmark. The authors in this research [15] presents a rapid and reliable framework for violence detection and localization in surveillance scenes, utilizing a Gaussian Model of Optical Flow (GMOF) to extract potential violence regions based on deviations from normal crowd behavior, followed by violence detection using video volumes densely sampled from these regions and classification with a novel descriptor called Orientation Histogram of Optical Flow (OHOF) via linear SVM. In this study [16], the authors proposed a model For the purpose of

identifying violence in video surveillance footage. The model combines a spatial feature extractor resembling a U-Net network with MobileNet V2 as an encoder, coupled with LSTM for temporal feature extraction and classification. Khan, Samee Ullah, et al. in this research [17] presents a violence detection scheme for movies, involving three steps: segmenting the movie into shots, selecting representative frames based on saliency, and passing them through a lightweight deep learning (DL) model, fine-tuned via transfer learning (TL) to distinguish between shots depicting violence and those without violent content, followed by merging non-violence scenes to generate a violence-free video. The authors in this study [18] introduces a novel feature descriptor called Histogram of Optical Flow Magnitude and Orientation (HOMO), which involves converting input frames to grayscale, computing optical flow between consecutive frames, comparing flow magnitude and orientation changes, applying threshold values to obtain binary indicators, and using these indicators to derive the HOMO descriptor, subsequently employed to train an SVM classifier. Ullah, Fath U. Min, et al. in this study [19] presents a triple-staged end-to-end DL violence detection framework, involving the detection of persons in video surveillance using a lightweight CNN model to filter out unnecessary frames, followed by the extraction of spatiotemporal features from sequences of 16 frames with detected persons using a 3D CNN, which are then fed to a Softmax classifier. The authors Sumon, Shakil Ahmed, et al. in this paper [20], various strategies are explored to determine feature saliency in violence detection from videos, utilizing three pretrained ImageNet models (VGG16, VGG19, ResNet50) to extract features from video frames. These features are then input into different networks, including a fully connected network and an LSTM network, while attention mechanisms via spatial transformer networks are applied. Among the models tested, features extracted by ResNet50, when combined with LSTM, achieved the highest accuracy of 97.06% in violence detection. This paper [21] introduces a lightweight computational model aimed at improving the categorization of violent and non-violent activities, employing a CNN-based Bidirectional LSTM for detecting violent behaviors and comparing it with other existing methods. The proposed model achieves high classification accuracies of 99.27%, 100%, and 98.64% on standard video datasets including Hockey Fights, Movies, and Violent-Flows, respectively.In the research [22], Asad, Mujtaba, et al. presents a new method for detecting fights or violent behaviors by learning spatial and temporal features from equally spaced sequential video frames. Utilizing multi-level features extracted from CNN layers and a proposed feature fusion method, motion information is considered, with a Wide-Dense Residual Block introduced to learn combined spatial features. These features are then concatenated and input to LSTM units to capture temporal dependencies. The authors in this study [23] proposed a neural architecture designed for violence detection via surveillance cameras, leveraging a pre-trained ResNet-50 model to extract features from video frames, which are subsequently fed into a ConvLSTM block. By employing short-term differences between video frames, the model enhances robustness to address occlusions and discrepancies, while convolutional neural networks facilitate the extraction of concentrated spatio-temporal features, complementing the sequential nature of videos for input into LSTMs. This work [24], Accattoli, Simone, et al. suggests utilizing a pre-trained 3D CNN, known as C3D, in conjunction with a SVM classifier to develop an automated system for detecting violence in videos.

Although these studies present models with high accuracy, they often require significant computational resources. This makes it challenging to deploy such models on resource-constrained devices, such as those in IoT systems. To address this limitation, this study proposes a model that utilizes MobileNetV2 for spatial feature extraction combined with BiLSTM for temporal feature extraction, which requires fewer computational resources, making it more suitable for deployment on IoT devices, while still maintaining strong performance in violence detection.

3. MATERIALS AND METHODS

3.1. The process of gathering and preparing data

In our study, we utilized the Real Life Violence Situations Dataset [25], [26], which contains several videos depicting both violent and non-violent behavior. The Real-Life Violence Situations Dataset comprises 2000 short videos, with 1000 videos depicting violent situations and 1000 videos showing non-violent scenes. The average length of the video clips is 5 seconds. Sample video clips from the dataset are shown in Figure 1. From the video database, we extracted 16 frames from each video. The process of extracting frames from each video was conducted systematically. Specifically, frames were sampled at regular intervals throughout the duration of the video to ensure representation across various time points. This method of frame extraction allowed us to capture a diverse range of visual information from each video, enabling comprehensive analysis and processing for subsequent tasks such as feature extraction and classification. The general formula for setting the current frame position of the video can be expressed as:

$$SkipFramesWindow = \frac{VideoFramesCount}{SequenceLength} \quad (1)$$

$$\text{FramePosition} = \text{FrameCounter} * \text{SkipFramesWindow} \quad (2)$$

Where:

- VideoFramesCount: the total number of frames in the video.
- SequenceLength: The number of frames in a video is extracted.
- SkipFramesWindow: the number of frames to skip in each iteration.
- FramePosition: the position of the current frame in the video.
- FrameCounter: the index of the current frame in the sequence.

After following the above steps to extract frames from the video database, we obtained a dataset consisting of 2000 images, ready for training and testing. We divided this dataset into three different categories using a training, validation, and testing ratio of 60:20:20. Before training the model, the data processing process includes resizing images to 64 x 64 and rescaling them to ensure that pixel values are normalized to the range [0, 1].

Figure 1. Sample Video clips from Real Life Violence Situations Dataset.

3.2. Overall Methodology

The overall methodology for detecting violence in surveillance video, as shown in Figure 2, involves several sequential steps: collecting data (video), extracting images from the video dataset to create a dataset for the task in this study, preprocessing images extracted from videos, training classification models, testing the trained model, and finally performing violence detection in videos. Firstly, we utilized the Real Life Violence Situations Dataset [25], [26], which comprises 2000 short videos capturing both violent and non-violent behavior. Each category includes 1000 videos, with an average clip duration of 5 seconds. From each video, we extracted 16 frames systematically as outlined in the process of gathering and preparing data section. This process creates a dataset of 2000 images for training and testing. Before training the model, the preprocessing steps include resizing to 64 x 64 and normalizing the pixel values to the range [0, 1]. Subsequently, We divided this dataset into three different categories using a training, validation, and testing ratio of 60:20:20, respectively. Then, the training set and validation set are utilized to train the four CNN architectures: MobileNet [27], MobileNetV2 [10], DenseNet121 [28], and ResNet50V2 [29], in combination with BiLSTM networks. Subsequently, the test set is utilized to evaluate the performance of the models, after which the model with the best performance is selected. Finally, the selected model is used to identify violence on surveillance cameras or surveillance videos.

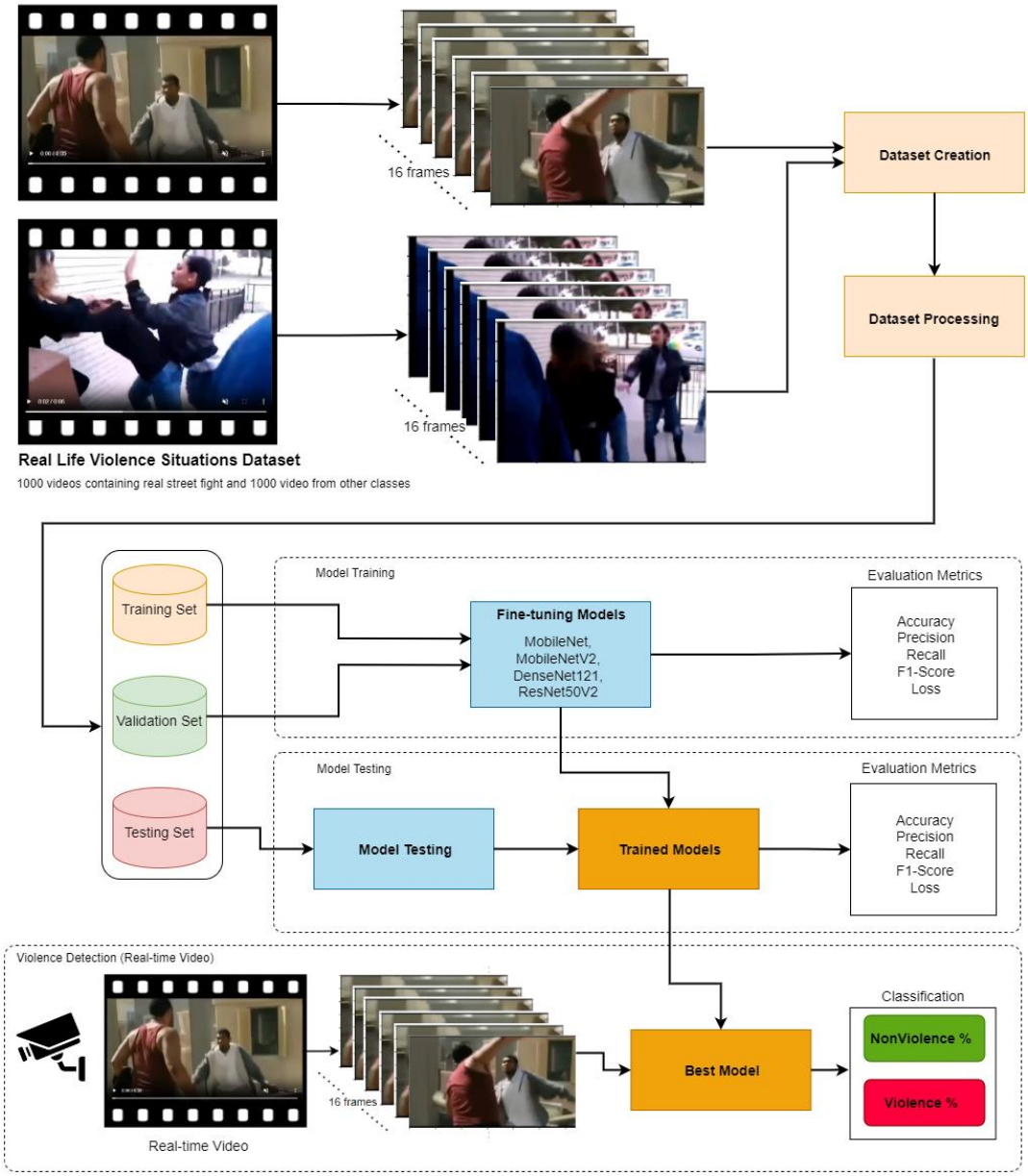

Figure 2. The overall methodology for detecting violence in surveillance video.

3.3. Proposed Model

Our research focuses on developing a model for violence detection in surveillance videos aimed at maintaining performance levels comparable to cutting-edge models, while also reducing computational complexity. This optimization enables the deployment of the model on low-resource edge devices, ensuring efficient operation in various real-world scenarios. High-level view of violence detection model is shown in Figure 3.

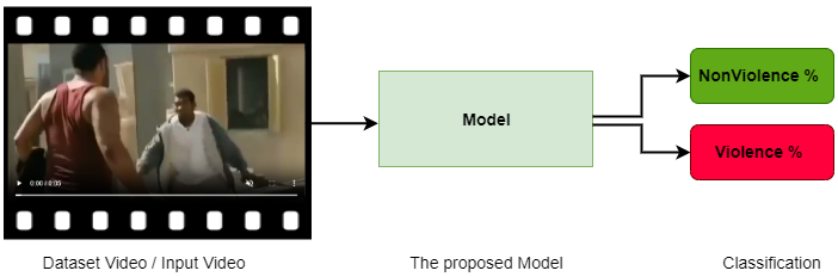

Figure 3. High-level View of Violence Detection Model

We propose the utilization of Time Distributed to enable the convolutional neural network to process multiple images simultaneously (frames of the surveillance videos). The output is then forwarded to BiLSTM network. Finally, the results are fed into Dense layers for classifying the sequence of frames. The basic structure of the proposed method for detecting violence in surveillance video is shown in Figure 4.

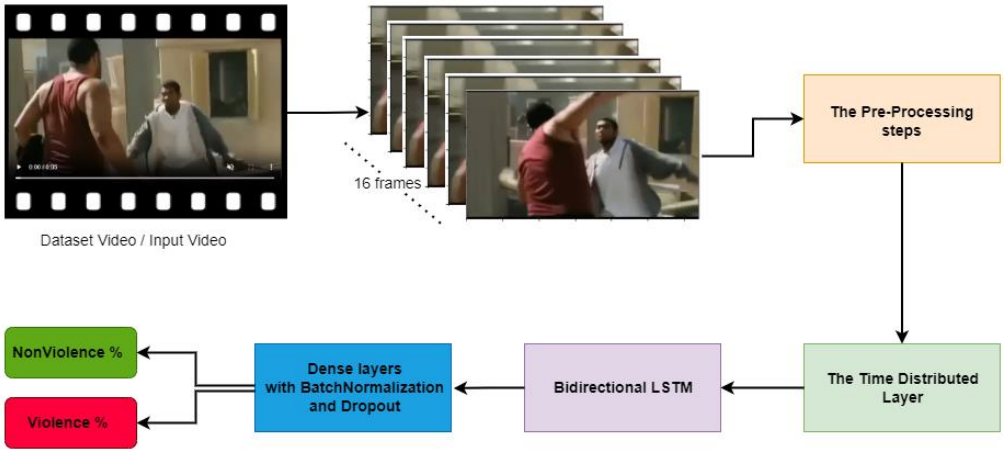

Figure 4. The basic structure of the proposed method for detecting violence in surveillance video

The model processes batches of 16 frames extracted from surveillance videos. It employs a TimeDistributed layer to apply a pre-trained convolutional neural network to each frame, extracting spatial features. The output of the previous layer is then passed to the Bidirectional LSTM layers to capture temporal features, followed by dense layers with activation functions, batch normalization, and dropout for feature extraction and classification. Finally, the output layer, using softmax activation, predicts the probability distribution across different classes. Figure 5 presents the proposed model architecture using Time Distributed and BiLSTM for detecting violence in surveillance video.

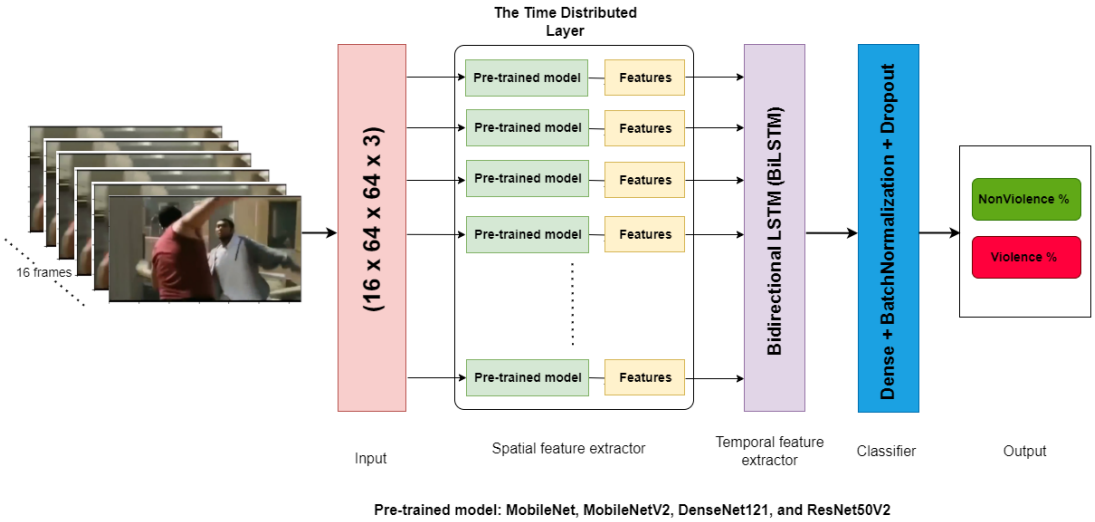

Figure 5. The proposed model architecture for detecting violence in surveillance video

3.4. Performance Evaluation Measures

In the realm of machine learning and data science research, evaluating model performance is crucial. Several metrics serve as yardsticks to assess the effectiveness of models. Accuracy measures the proportion of correctly predicted instances out of the total. Precision quantifies the true positive rate among the predicted positive instances, while Recall (also known as sensitivity) gauges the true positive rate among actual positive instances. The f1-score, which balances precision and recall, provides a comprehensive view of model performance.

Additionally, loss functions play a pivotal role during model training, guiding optimization by quantifying the discrepancy between predicted and actual values.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recal = \frac{TP}{TP+FN} \quad (5)$$

$$F_1 - Score = \frac{Precision*Recall}{Precision+Recall} \quad (6)$$

$$Loss = -\sum_{j=1}^{k} y_i \log(\hat{y}_i) \quad (7)$$

In which, FP represent False Positive, TN denote True Negative, TP signify True Positive, and FN indicate False Negative. The variable k represents the number of classes, while y corresponds to the actual value, ŷ is prediction value.

4. RESULTS AND DISCUSSION

4.1. Environmental settings

Our experimental procedures were conducted on the Kaggle platform to acquire the experimental outcomes. The research employed a Tesla P100-PCIE GPU with 16GB of memory, while the system itself possessed 29GB of RAM. GPU information is presented in Figure 6.

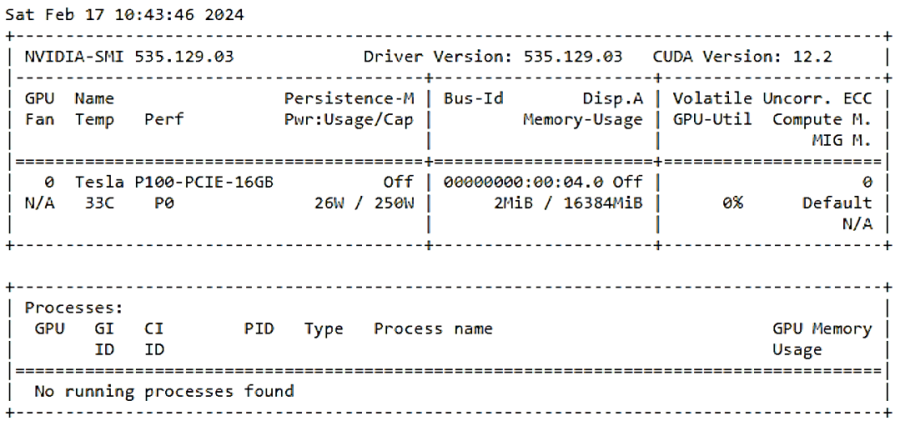

Figure. 6. GPU information

4.2. Experiments

4.2.1 Experiments 01: Confusion matrix of the MobileNetV2BiLSTM model for violence detection on the test set.

Table 1. Hyperparameters Of The Proposed Deep Learning Methodology For Training A Violence Detection Model

Hyperparameters	Value
Batch Size	16
Number of Epochs	100
Optimizer	SGD
Loss Function	Categorical Crossentropy
Activation Function in Hidden Layer	Relu
Activation Function in Output Layer	Softmax
EarlyStopping Monitor	Val Accuracy
EarlyStopping Patience	10
Learning Rate	ReduceLROnPlateau

The hyperparameters of the proposed models for training a violence detection model are shown in the Table 1. The table 2 presents the results of different proposed models for violence detection based on various evaluation metrics. Each row corresponds to a different proposed model, including MobileNetBiLSTM, MobileNetV2BiLSTM, DenseNet121BiLSTM, and ResNet50V2BiLSTM. The "Loss" column indicates the average loss, while the "Precision", "Recall", "F1-score", and "Accuracy" columns represent performance metrics related to the model's ability to correctly classify violent and non-violent images.

Higher values in these metrics indicate better performance. For instance, MobileNetV2BiLSTM shows the lowest loss and highest accuracy among the models, achieving a precision, recall, and F1-score of 95.00%. These results suggest that MobileNetV2BiLSTM performs the best among the models evaluated in terms of accuracy and overall performance in violence detection.

Table 2. The table presents the results of the proposed models on the testing set for detecting violence.

Model	Loss	Precision(%)	Recall(%)	F1-Score(%)	Accuracy(%)
MobileNetBiLSTM	0.2869	93.00	93.00	93.00	92.75
MobileNetV2BiLSTM	**0.2127**	**95.00**	**95.00**	**95.00**	**95.00**
DenseNet121BiLSTM	0.6833	91.00	91.00	91.00	91.00
ResNet50V2BiLSTM	0.6389	93.50	94.00	94.00	93.75

The training/validation accuracy and loss of the MobileNetV2BiLSTM model are displayed in Figure 7. Comparison results of the proposed models for violence detection on the test set are shown in Figure 8. Confusion matrix of the MobileNetV2BiLSTM model for violence detection on the test set are shown in Figure 9.

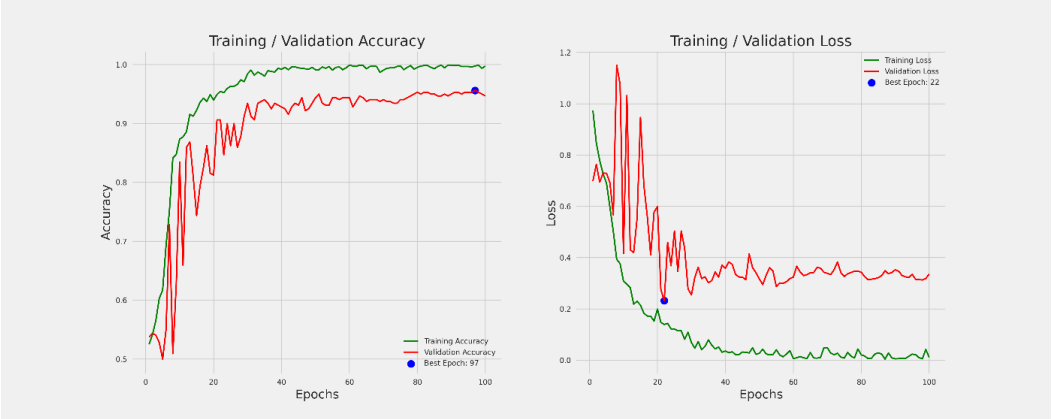

Figure 7. The Training / Validation Accuracy and Loss of MobileNetV2BiLSTM model

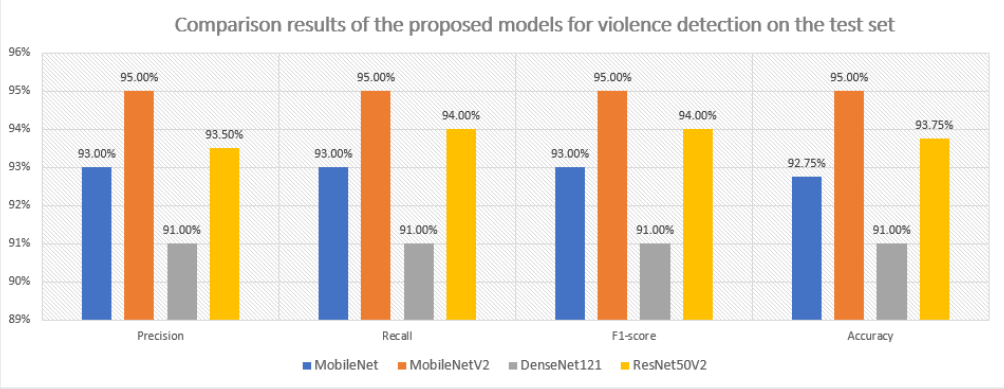

Figure 8. Comparison results of the proposed models for violence detection on the test set

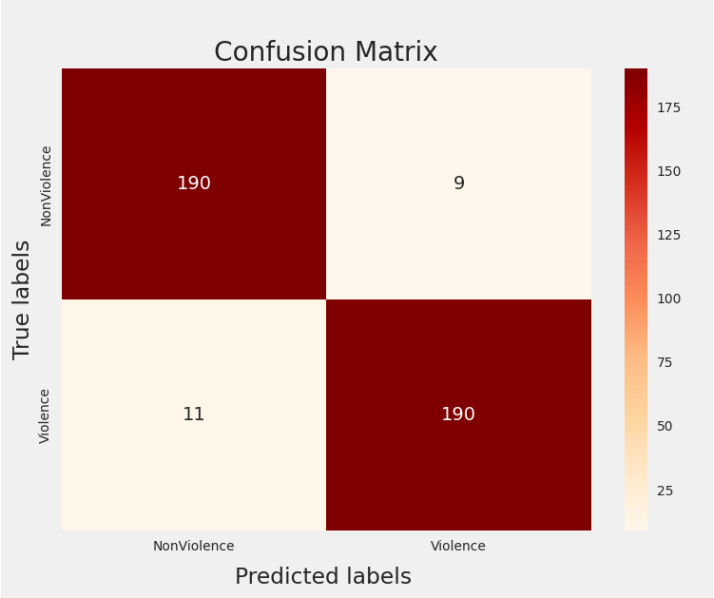

Figure 9. Confusion matrix of the MobileNetV2BiLSTM model for violence detection on the test set

Table 3. The number of model parameters in the proposed model is compared to previous research.

Paper	Num of Parameters
Vijeikis, et al. [16]	4M
Rendón-Segador, Fernando J., et al. [30]	4.5M
Sudhakaran, Swathikiran, and Oswald Lanz [31]	9.6M
Aktı, Şeymanur, Gözde Ayşe Tataroğlu, and Hazım Kemal Ekenel [32]	9M
Li, Ji, et al. [33]	7.4M
Proposed model	**3.6M**

Table 3 presents the number of parameters for various models. The models evaluated include those proposed by Vijeikis et al. [16], which contains 4 million parameters, and Rendón-Segador et al. [30], with a total of 4.5 million parameters. Sudhakaran et al. [31] introduced a model with 9.6 million parameters, while Aktı et al. [32] reported a model comprising 9 million parameters. Li et al. [33] proposed a model with 7.4 million parameters. In contrast, the proposed model in this study significantly reduces the number of parameters to 3.6 million, highlighting its resource efficiency and suitability for deployment on devices with limited computational capabilities, such as in IoT systems.

4.2.2 Experiments 02: Selected model for violence detection on real-time video and input video.

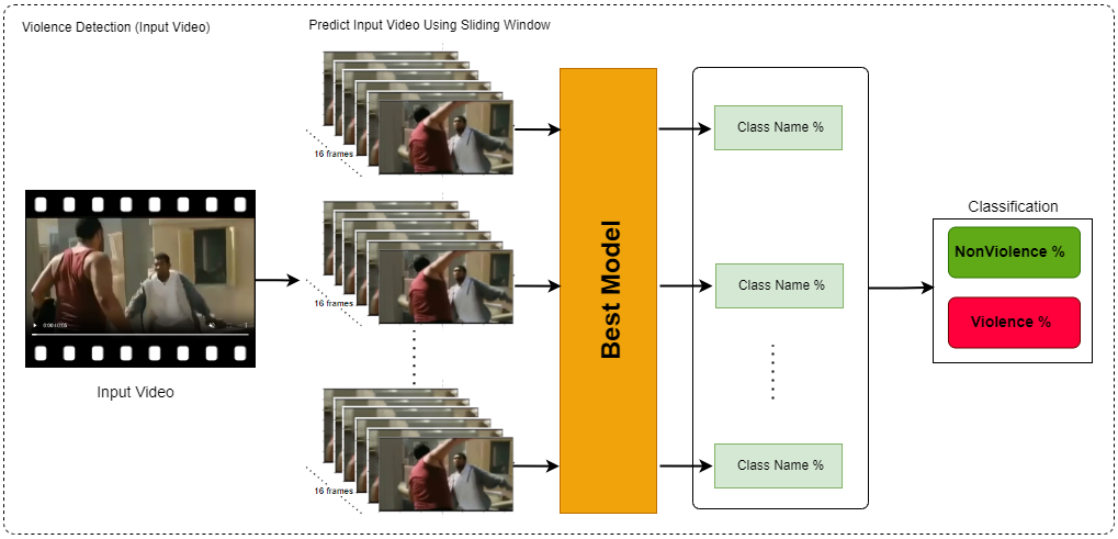

Figure 10. Predict Input Video Using Sliding Window

Based on the results reported in Experiment 1, where various proposed models were trained for violence detection, the MobileNetV2BiLSTM model demonstrated superior performance compared to others. Moreover, our selection of the pretrained MobileNetV2 was based on its compact size in computational requirements and learned parameters, which allows for effective real-time operations while maintaining good accuracy. Therefore, it was selected for the task of detecting violence in both input and real-time videos.

The task of detecting violence in real-time videos: From the surveillance camera, every sequence of 16 frames will be sent through the model to identify violence. This means that the model will analyze a group of 16 consecutive frames at a time to determine if any violent behavior is present in the video footage.

The task of detecting violence in input videos: The processing involves the following steps: Firstly, the input video undergoes a sliding window technique, dividing it into sequential frames (e.g., 16 frames per window). Then, for each window, the frames are read, resized, and normalized before predicting the class label using a selected model (MobileNetV2BiLSTM). Subsequently, the predicted class name and probabilities for each window in the video are stored. Lastly, the process identifies the class with the highest number of predictions and computes its average probability from the prediction data. Prediction of input video using sliding window is presented in Figure 10.

5. CONCLUSION AND FUTURE WORK

In conclusion, the study aimed to develop effective models for violence detection in surveillance videos, with a focus on balancing performance and computational efficiency. Experiment 1 involved training various models, among which MobileNetV2BiLSTM (MobileNetV2 combined with BiLSTM networks) achieved high-performance results, demonstrating superior accuracy and overall performance. Experiment 2 selected the MobileNetV2BiLSTM model for violence detection in both input and real-time videos due to its outstanding performance. For real-time videos, the model analyzes sequences of 16 frames at a time, while for input videos, a sliding window technique is employed to process sequential frames.

This selected model shows promising potential for practical deployment in real-world scenarios, offering efficient violence detection while conserving computational resources. These findings emphasize the importance of leveraging advanced models like MobileNetV2BiLSTM to enhance security measures and public safety through IoT systems.

Future research in this task could explore several directions. Firstly, incorporating advanced techniques such as attention mechanisms or reinforcement learning could assist the models in better focusing on relevant regions or behaviors within the video footage. Secondly, there could be a focus on optimizing existing models to enhance both performance and computational efficiency. This could involve fine-tuning hyperparameters, improving training procedures, and exploring new optimization methods. Third, research could explore multi-task models, which are capable of performing multiple tasks simultaneously, such as violence detection alongside object recognition or other threat behavior recognition. Lastly, In the current scope of our research, we focus on developing a violence detection model that can be deployed on devices with limited computational resources, such as those in IoT systems. The development of a full IoT system, including applications such as alarms or notifications to police or authorities, will be part of our future research. We plan to not only stop at detection but also incorporate response measures, such as automated alerts, alarms, and emergency management, ensuring a more comprehensive solution in future developments.

REFERENCES

[1] Dario Bacchini and Concetta Esposito. Growing up in violent contexts: differential effects of community, family, and school violence on child adjustment. *Children and peace: From research to action*, pages 157–171, 2020.

[2] James A Mercy, Susan D Hillis, Alexander Butchart, Mark A Bellis, Catherine L Ward, Xiangming Fang, and Mark L Rosenberg. Interpersonal violence: global impact and paths to prevention. *Injury prevention and environmental health. 3rd edition*, 2017.

[3] Linda L Dahlberg and Etienne G Krug. Violence a global public health problem. *Ciencia & Saude Coletiva*, 11(2):277–292, 2006.

[4] Duarte Duque, Henrique Santos, and Paulo Cortez. Prediction of abnormal behaviors for intelligent video surveillance systems. *In 2007 IEEE Symposium on Computational Intelligence and Data Mining*, pages 362–367. IEEE, 2007.

[5] Jong Sun Kim, Dong Hae Yeom, and Young Hoon Joo. Fast and robust algorithm of tracking multiple moving objects for intelligent video surveillance systems. *IEEE Transactions on Consumer Electronics*, 57(3):1165–1170, 2011.

[6] Chao Huang, Zhihao Wu, Jie Wen, Yong Xu, Qiuping Jiang, and Yaowei Wang. Abnormal event detection using deep contrastive learning for intelligent video surveillance system. *IEEE Transactions on Industrial Informatics*, 18(8):5171–5179, 2021.

[7] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. *In 2011 International conference on computer vision*, pages 1331–1338. IEEE, 2011.

[8] Josephine Sullivan and Stefan Carlsson. Recognizing and tracking human action. *In Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part I 7*, pages 629–644. Springer, 2002.
[9] Muhammad Attique Khan, Kashif Javed, Sajid Ali Khan, Tanzila Saba, Usman Habib, Junaid Ali Khan, and Aaqif Afzaal Abbasi. Human action recognition using fusion of multiview and deep features: an application to video surveillance. *Multimedia tools and applications*, pages 1–27, 2020.
[10] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
[11] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.
[12] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno Garcıa, and Rahul Sukthankar. Violence detection in video using computer vision techniques. *In Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14*, pages 332–339. Springer, 2011.
[13] Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang, and Chih-Wen Su. Violence detection in movies. *In 2011 Eighth International Conference Computer Graphics, Imaging and Visualization*, pages 119–124. IEEE, 2011.
[14] Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, and Yi Liu. Violence detection using oriented violent flows. *Image and vision computing*, 48:37–41, 2016.
[15] Tao Zhang, Zhijie Yang, Wenjing Jia, Baoqing Yang, Jie Yang, and Xiangjian He. A new method for violence detection in surveillance scenes. *Multimedia Tools and Applications*, 75:7327–7349, 2016.
[16] Romas Vijeikis, Vidas Raudonis, and Gintaras Dervinis. Efficient violence detection in surveillance. *Sensors*, 22(6):2216, 2022.
[17] Samee Ullah Khan, Ijaz Ul Haq, Seungmin Rho, Sung Wook Baik, and Mi Young Lee. Cover the violence: A novel deep-learning-based approach towards violence-detection in movies. *Applied Sciences*, 9(22):4963, 2019.
[18] Javad Mahmoodi and Afsane Salajeghe. A classification method based on optical flow for violence detection. *Expert systems with applications*, 127:121–127, 2019.
[19] Fath U Min Ullah, Amin Ullah, Khan Muhammad, Ijaz Ul Haq, and Sung Wook Baik. Violence detection using spatiotemporal features with 3d convolutional neural network. *Sensors*, 19(11):2472, 2019.
[20] Shakil Ahmed Sumon, Raihan Goni, Niyaz Bin Hashem, Tanzil Shahria, and Rashedur M Rahman. Violence detection by pretrained modules with different deep learning approaches. *Vietnam Journal of Computer Science*, 7(01):19–40, 2020.
[21] Rohit Halder and Rajdeep Chatterjee. Cnn-bilstm model for violence detection in smart surveillance. *SN Computer science*, 1(4):201, 2020.
[22] Mujtaba Asad, Jie Yang, Jiang He, Pourya Shamsolmoali, and Xiangjian He. Multi-frame feature-fusion-based model for violence detection. *The Visual Computer*, 37:1415–1431, 2021.
[23] Manan Sharma and Rishabh Baghel. Video surveillance for violence detection using deep learning. *In Advances in Data Science and Management: Proceedings of ICDSM 2019*, pages 411–420. Springer, 2020.
[24] Simone Accattoli, Paolo Sernani, Nicola Falcionelli, Dagmawi Neway Mekuria, and Aldo Franco Dragoni. Violence detection in videos by combining 3d convolutional neural networks and support vector machines. *Applied Artificial Intelligence*, 34(4):329–344, 2020.
[25] Real life violence situations dataset, available online: https://www.kaggle.com/datasets/mohamedmustafa/real-life-violencesituations-dataset.
[26] Mohamed Mostafa Soliman, Mohamed Hussein Kamal, Mina Abd ElMassih Nashed, Youssef Mohamed Mostafa, Bassel Safwat Chawky, and Dina Khattab. Violence recognition from videos using deep learning techniques. *In 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 80–85, 2019.
[27] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
[28] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016
[30] Fernando J Rend´on-Segador, Juan A ´Alvarez-Garc´ıa, Fernando Enr´ıquez, and Oscar Deniz. Violencenet: Dense multi-head self-attention with bidirectional convolutional lstm for detecting violence. *Electronics*, 10(13):1601, 2021
[31] Swathikiran Sudhakaran and Oswald Lanz. Learning to detect violent videos using convolutional long short-term memory. *In 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2017.
[32] S¸eymanur Aktı, G¨ozde Ays¸e Tataro˘glu, and Hazım Kemal Ekenel. Vision-based fight detection from surveillance cameras. *In 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2019.

[33] Ji Li, Xinghao Jiang, Tanfeng Sun, and Ke Xu. Efficient violence detection using 3d convolutional neural networks. *In 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.

BIOGRAPHY OF AUTHORS

HoangTu-Vo holds a Bachelor of Information Systems from Can Tho University, Vietnam, in 2011. In 2013, He graduated with a master's degree in Information Systems from Can Tho University in Vietnam. Currently, He is working as a lecturer in Information Technology Department at FPT University, Can Tho campus in Vietnam. His research interests include Machine learning, Deep learning, Image processing, and Computer vision. He can be contacted at email: tuvh6@fe.edu.vn

Phuc Pham Tien earned his Bachelor's degree in Information Systems from Can Tho University, Vietnam, in 2003. In 2010, he completed a Master's degree in Information Technology at the same university. Currently, he is a lecturer in Information Technology Department at FPT University, Can Tho campus, Vietnam. His research interests include machine learning and deep learning. He can be reached at phucpt@fe.edu.vn

Nhon Nguyen Thien holds a Bachelor of Information Systems from Can Tho University, Vietnam, in 2013. In 2017, He graduated with a master's degree in Information Systems from Can
Tho University in Vietnam. Currently, He is working as a lecturer in Information Technology Department at FPT University, Can Tho campus in Vietnam. His research areas of interest include Data science, Machine learning, Deep learning, and Web application development. He can be contacted at email: nhonnt9@fe.edu.vn

Kheo Chau Mui holds a Bachelor of Information Systems from Can Tho University, Vietnam, in 2010. In 2020, She graduated with a master's degree in Computer Science from Can
Tho University in Vietnam. Currently, She is working as a lecturer in Information Technology Department at FPT University, Can Tho campus in Vietnam. Her research areas of interest include
Image processing, Image classification, Data science, Object detection, Deep learning, and Machine learning. She can be contacted at email: kheocm@fe.edu.vn