AI-based Violent Incident Detection in Surveillance Videos to Enhance Public Safety

Khaled Merit and Mohammed Beladgham

Tahri Mohammed University of Bechar, Algeria

https://doi.org/10.26636/jtit.2025.4.2328

Abstract — Acts of violence may occur at any moment, even in densely populated areas, making it important to monitor human activities to ensure public safety. Although surveillance cameras are capable of detecting the activity of people, around-the-clock monitoring still requires human support. As such, an automated framework capable of detecting violence, issuing early alerts, and facilitating quick reactions is required. However, automation of the entire process is challenging due to issues such as low video resolution and blind spots. This study focuses on detecting acts of violence using three video data sets (movies, hockey game and crowd) by applying and comparing advanced ResNet architectures (ResNet50V2, ResNet101V2, ResNet152V2) with the use of the bidirectional gated recurrent unit (BiGRU) algorithm. Spatial features of each video frame sequence are extracted using these pre-trained deep transfer learning models and classified by means of an optimized BiGRU model. The experimental results were then compared with those achieved by wavelet feature extraction approaches and other classification models, including CNN and LSTM. Such an analysis indicates that the combination of ResNet152V2 and BiGRU offers decent performance in terms of higher accuracy, recall, precision, and F1 score across the different datasets. Furthermore, the results indicate that deeper ResNet models significantly improve overall performance of the model in terms of violence detection scores, relative to shallower ResNet models. ResNet152V2 was found to be the ultimate model across the datasets when it comes to a high degree of accuracy in detecting acts of violence.

Keywords — bidirectional gated recurrent unit, deep learning, deep transfer learning, video processing, violence detection

1. Introduction

Violence is defined as deliberate exertion of physical force intended to harm, dominate, or manipulate individuals or groups. Actions of this type are considered to constitute criminal conduct that transgresses legal boundaries, societal expectations, and moral principles followed by global communities. The effects of violence are multifaceted, encompassing not only bodily injury, but also deep emotional and psychological distress which can, in extreme situations, lead to fatal outcomes.

Currently, the use of closed-circuit television (CCTV) is increasing because the solution is capable of providing non-stop surveillance – a task humans cannot accomplish. Cameras record all the events from various angles. As a result, a large amount of video data still requires a human to identify unwelcome types of activity, including violence. If performed manually, this video-monitoring process requires significant amounts of time and effort. Therefore, it is necessary to employ an automatic detection system that will accelerate the entire procedure. One of the challenges faced when performing automatic detection is low resolution of the video feed generated by CCTV cameras [1] (resulting from poor lighting, ambient conditions, distance, and hardware constraints).

Automatic event detection has been possible for several years now, and the process of detecting acts of violence is similar to that of recognizing actions [2]. The difference is that violence detection focuses not only on movement, but also on the intention of that movement. In this case, the speed of movement that occurs will determine whether a given action is categorized as an act of violence or just an ordinary movement.

The authors of [3]–[8] detect objects in CCTV video data. However, not all acts of violence, such as hand-to-hand fights or altercations, involve weapons. Therefore, it is necessary to detect acts of violence that do not depend on a suspicious object.

Several studies have been conducted that focused on detecting acts violence, with various approaches relied upon in the process. The author of [9] used a histogram of optical flow (HOF) to extract valuable features from videos, while in [10], HOF magnitude and orientation (HOMO) are used. In [11], motion features are extracted from dynamic RGB images, while in [12] convolutional neural network (CNN) models (namely VGG-19, ResNet50 and Xception) are employed, with each of them trained using the ImageNet dataset. The results they achieve are reasonably good.

Studies [13] and [14] used VGG-16 for feature extraction and a simple SVM classification algorithm. Better results were obtained in 0, which used ResNet50 as the backbone for three-dimensional CNNs and dense optical flow for the region of interest.

Another difficulty encountered while detecting acts of violence with the use of surveillance cameras stems from the presence of crowds in public places. The violent flow dataset, also known as the crowd dataset, is one example of a dataset containing videos of public crowds. Several studies have re-

lied upon the crowd data set. For example, the author of [16] used a violent flow (ViF) descriptor and then classified the output using a linear SVM, achieving a precision score of 81.3%.

Using the same classification algorithm combined with HOF, the researchers in [17] obtained an accuracy of 83.37%. That result still needs to be improved to create a precise detection system. This dataset is challenging because acts of violence are sometimes not visible due to the density of the crowd.

On the other hand, crowded conditions often lead to false positives. Therefore, in this study, acts of violence were detected using the crowd dataset, with the overall aim of improving the quality of the model in terms of its performance and lead time.

To classify data into appropriate classes, a powerful classification model is needed. the following solutions were used: bidirectional gated recurrent unit model (BiGRU), long-short-term memory model (LSTM), CNN, etc. The LSTM model used images and also accepted other data types, such as text, and achieved very good accuracy levels [18].

The main contribution of this work is a benchmark of advanced ResNet architectures (ResNet50V2, ResNet101V2, ResNet152V2) against classical and other deep learning-based feature extractors, when combined with various temporal classifiers (CNN, LSTM, BiGRU). Our work provides a clear evidence-based pathway for selecting model components, demonstrating that the synergistic combination of ResNet152V2 and BiGRU delivers superior and consistent state-of-the-art performance across diverse benchmark datasets.

We used ResNet50V2, ResNet101V2, and ResNet152V2 to extract vital features from video and wavelets. BiGRU was selected as a classification algorithm, as it offers better results than LSTM in terms of predicting the condition of a pulp paper press [19]. In the classification of emotions in noisy speech, BiGRU provides a shorter run time and a lower error rate while removing noise, compared to LSTM [20]. For comparison, this research also uses the LSTM and CNN algorithms.

The structure of this paper is as follows. The description and video pre-processing stages are detailed in Section 2. The methods used for extracting violence-specific features are explained in Section 3. Violence classification algorithms are presented in Section 4. Experimental results and discussion, including computational efficiency analysis, are provided in Section 5. Ethical considerations are discussed in Section 6. Conclusions and future work are described in Section 7.

2. Video Detection

A general scheme for detecting violent acts is illustrated in Fig. 1. Initially, it is essential to pre-process the video data, followed by a systematic categorization into training and testing datasets utilizing k-fold validation. Subsequently, the feature extraction stage is performed using ResNet50V2, ResNet101V2, and ResNet152V2. We also compared the features extracted using several methods: principal component analysis (PCA), discrete wavelet transforms (DWT), VGG-16 and VGG-19.

The most effective method of extracting features from the training data were used to develop a violence detection model employing the BiGRU algorithm. We compared the classification model with several algorithms such as CNN and LSTM. In the final stage, the model was assessed using the test dataset. This evaluation utilized metrics such as accuracy, recall, specificity, G-mean, and CPU time to thoroughly gauge the effectiveness of the model.

In addition to ResNetXV2, we also compared wavelet feature extraction methods and non-feature extraction to compare performance in terms of violence detection and extraction processing time.

2.1. Dataset

Data from three datasets were used in this research to assess the performance of the model in detecting violence in a video: movies [21], hockey game [21], and crowd [16] (Tab. 1).

The videos in the movies dataset contain several movie scenes and consist of 200 clips divided into 100 fight and 100 non-fight sequences. The hockey dataset contains 1000 video recordings of matches from the National Hockey League, divided into 500 violent and 500 non-violent clips. The crowd dataset is a real-time video recording of violence in a crowd, containing 246 videos with 123 violent and 123 non-violent clips. Each dataset was divided into training and test datasets using k-fold validation.
Figures 2, 3 present sample frames from each dataset.

2.2. Pre-processing

Pre-processing phase is the preliminary step in building a violence-detection system. In this step, each video is converted into a series of RGB format images. These images are subsequently resized to 224 × 224 pixels to align with the input specifications of the ResNet models.

The next phase involves extracting the pixel intensities from each set of images. In this scenario, we obtained a matrix with dimensions $m \times n \times 224 \times 224 \times 3$. In this scenario, m signifies the total number of clips, n indicates the number of images captured per recording session, and $224 \times 224 \times 3$ specifies the dimensions of an RGB image in bytes.

Tab. 1. Brief description of datasets used to detect violence in video footage.

Datasets	Frame size	No. of clips	Violence	No violence	Format
Movies [21]	576 × 720	200	100	100	.mpg .mp4
Hockey [21]	288 × 360	1000	500	500	.avi
Crowd [16]	240 × 320	246	123	123	.avi

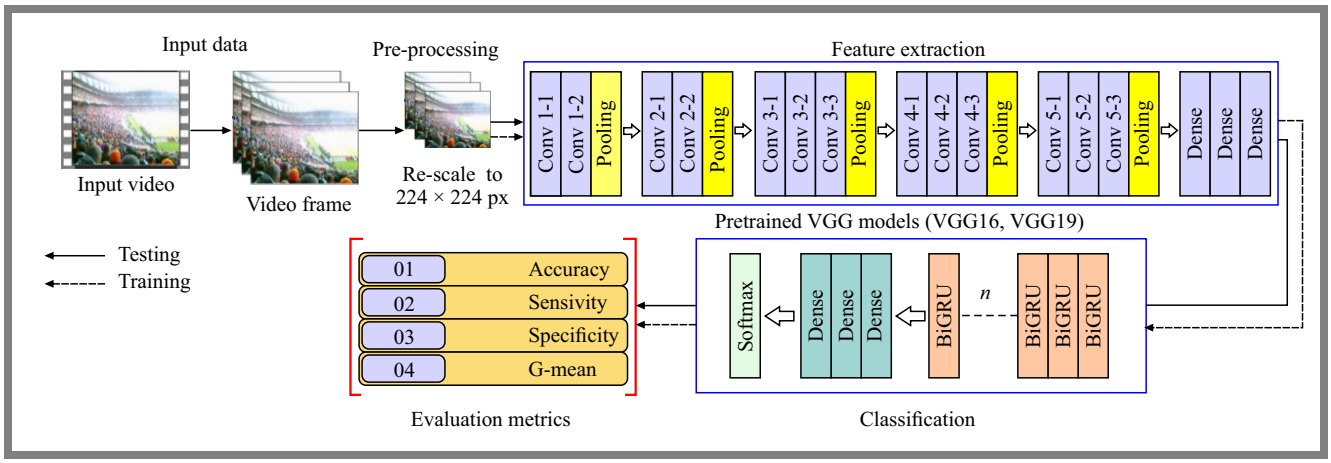

Fig. 1. Schematic of the violence detection system.

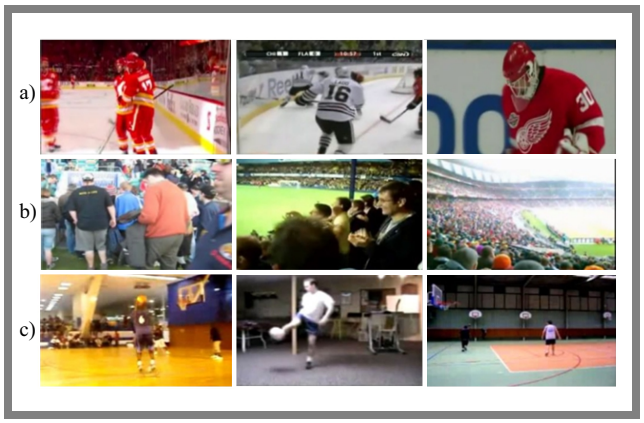

Fig. 2. Sample frames of non-violent video benchmark datasets: a) hockey dataset, b) crowd dataset, and c) movie dataset.

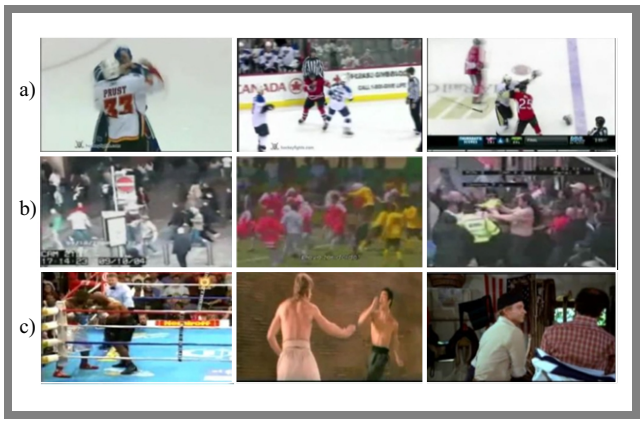

Fig. 3. Sample frames of violent video benchmark datasets: a) hockey dataset, b) crowd dataset, and c) movie dataset.

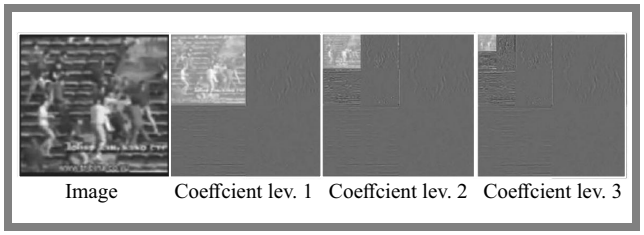

Fig. 4. Feature extraction process using DWT.

3. Violence Feature Extraction

3.1. Discrete Wavelet Transform

In this research, level 3 wavelet decomposition was used to compare the feature extraction methods. The mother wavelet uses Daubechies 8, N (in Db), where N represents the Daubechies polynomial order. The wavelet of the Daubechies order $N \geqslant 2$ has $2N$ vanishing moments and a small-scale support with an interval of $[0, 2N-1]$ [22]. The Daubechies polynomial order $N-1$ is defined as:

$$P_{N-1}(y) = \sum_{k=0}^{N-1} \binom{2N-1}{k} y^k (1-my)^{N-1-k} . \qquad (1)$$

After obtaining a grayscale image, level 1 wavelet decomposition is performed and then LL, LH, HL, and HH sub-bands are obtained. The LL sub-band contains the approximate value of the image and is the input for the next decomposition level. The sub-band used during the classification process is the approximate value of the level 3 wavelet decomposition. In this process, a matrix measuring $m \times n \times 41 \times 41$ is produced. The matrix is reshaped to adjust the input dimensions in the classification process. The results of feature extraction using the DWT are shown in Fig. 4.

3.2. Principal Component Analysis

Principal component analysis (PCA) is a transformation technique that converts and decomposes a large set of correlated variables into a smaller set of uncorrelated variables. This method effectively reduces the dimensionality of the data while preserving essential information. Each image frame is converted to grayscale and dimensioned into a row vector with dimension $(1 \times m)$, where m is $n \times n$, and n is the size of the image. For each dataset, all vectors were aggregated into a size matrix of size $(N \times 50\,176)$, where N is the number of images. The next step is to select the value of the principal component with k percent of the total eigenvalues. The results of the feature extraction using PCA are shown in Fig. 5.

3.3. Residual Networks (ResNet)

ResNet is a deep learning approach and is an evolution of a CNN. In the learning process, ResNet implements residual connections that can connect layers to other layers by skipping some middle layers. It is claimed to avoid the vanishing gradient problems that occur during the training process [23]. More than the use of a deep learning architecture alone is needed to increase the accuracy of the learning process. Therefore, to improve recognition accuracy, transfer learning is used.

Transfer learning is an approach to deep learning (and machine learning) in which knowledge is transferred from one model to another. A common misconception regarding transfer learning is that training and test datasets must come from the same source or have the same distribution. In practice, however, the transferred tasks may differ in the same domain. In common deep neural networks, models learn only from existing data. With limited data, it will be difficult for the model to obtain optimal recognition results. Deep transfer learning, on the contrary, using pre-trained models trained on other datasets in the same domain, can boost classification performance [24].

ResNet50V2, ResNet101V2, and ResNet152V2 are improved versions of their respective ResNet families, incorporating identity mapping and applying batch normalization and ReLU activation before the weight layers, which enhances gradient flow and training stability [25]. The key characteristics of these architectures, which form the basis of our feature extraction comparison, are summarized in Tab. 2.

The ResNetXV2 architecture is illustrated in Fig. 6. In this study, we used these models as feature extractors for the input video. Subsequently, we advanced the learning process by employing additional deep learning techniques, including CNN, LSTM, and BiGRU, as classification methods. The matrix resulting from block 5 for each ResNet variant is characterized by dimensions of $m \times n \times 7 \times 7 \times 512$.

The matrix is subsequently processed through the flatten and dense layers, resulting in a matrix of size of $m \times n \times 4096$. The characteristic extraction procedure utilizing ResNet ends at the dense layer and further processing is conducted using an alternative classifier.

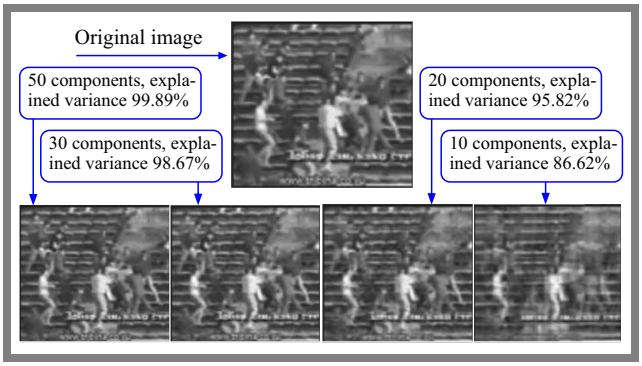

Fig. 5. Feature extraction process using PCA.

Tab. 2. Architectural comparison of the ResNetV2 models used to extract features.

Model	Depth (layers)	Parameters (millions)	Bottleneck blocks	Complexity
ResNet50V2	50	25.6	16 3×[3, 4, 6, 3]	Medium
ResNet101V2	101	44.6	33 3×[3, 4, 23, 3]	High
ResNet152V2	152	60.2	50 3×[3, 8, 36, 3]	Highest

3.4. VGG

VGG represents a convolutional neural network (CNN) framework that was developed using the ImageNet database [24]. VGG can handle massive datasets, as it contains several weighted layers with millions of parameters. The difference between VGG-16 and VGG-19 networks is the depth of the weight layers, as shown in Fig. 7. In VGG-16, the number of weight layers is 16, whereas VGG-19 has a layer depth of 19. We used VGG-16 and VGG-19 as a comparison feature extractor for the input video. The output matrix from block 5 of the VGG-16 model has dimensions of $m \times n \times 7 \times 7 \times 512$. After being processed through both the flatten and dense layers, the matrix is reconfigured to have dimensions of $m \times n \times 4096$.

4. Violance Classification

After acquiring the feature set from the trained ResNet models, we compared several deep learning methods to detect acts of violence in a given. The CNN in this study consists of three convolution and max-pooling layers. The CNN architecture is shown in Fig. 8a. The hyperparameter settings for the CNN were an initial learning rate of 0.1, a batch size of 100, 200 epochs, a dense kernel size of 100, a loss function based on mean squared error, and SGD optimizers.

LSTM is an advanced recurrent neural network that solves the vanishing gradient problem [26]. Each LSTM cell has three gates, namely a forget gate, an input gate, and an output gate (Fig. 8b). In this study, the hyperparameter settings for LSTM were as follows: an initial learning rate of 0.1, a batch size of 100, 100 epochs, a dense kernel size of 100, a loss function based on mean squared error, and the Adam optimizer.

GRU was introduced in [27], with its design similar to that of the LSTM but using a more straightforward memory unit to simplify training and implementation. In this study, classification was performed using a bidirectional GRU (BiGRU) to better capture contextual information from both past and future frames (Fig. 8c). The result of the feature extraction stage using ResNet passes through the BiGRU layer in this process. Furthermore, the resulting BiGRU matrix goes through three dense layers and the last output goes through a dense layer with two units using the softmax activation function. This layer maps the classification results into two class labels: violence or non-violence. The hyperparameter settings for the BiGRU were an initial learning rate of 0.1, a batch size of 100, 100 epochs, a dense kernel size of 100, a loss function based on mean squared error and the Adam optimizer.

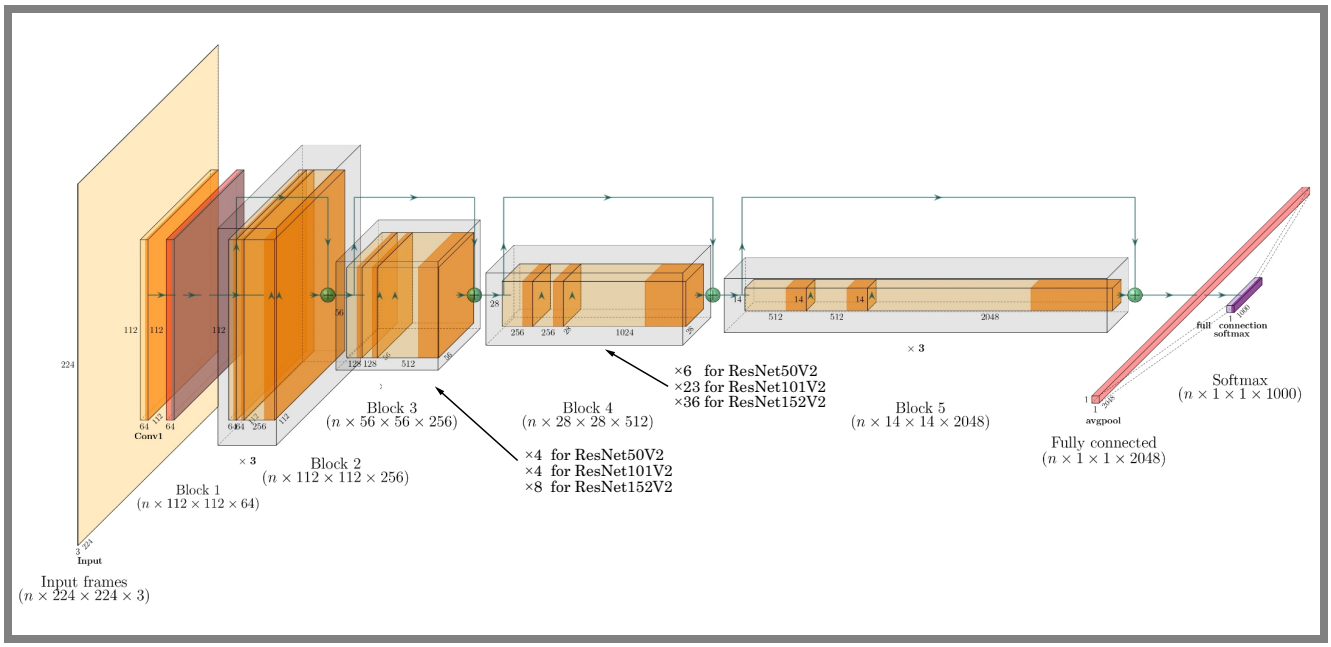

Fig. 6. Three-dimensional ResNetXV2 architecture.

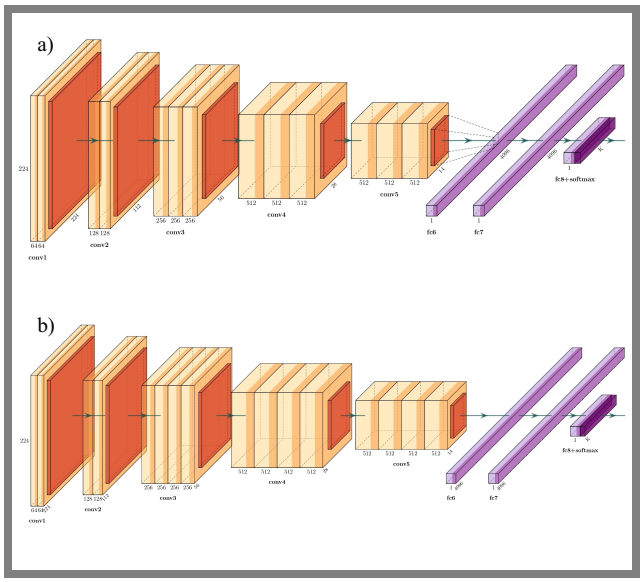

Fig. 7. VGG architecture: a) VGG16 and b) VGG19.

The bidirectional architecture enables the model to capture more comprehensive temporal patterns, which is particularly beneficial for violence detection in video sequences, where context from both preceding and subsequent frames is crucial for accurate classification.

5. Results and Discussion

In this study, acts of violence were classified using the following publicly available datasets: movies, hockey game, and crowd. We used a deep transfer learning approach based on ResNet50V2, ResNet101V2, and ResNet152V2 to extract essential features from the data. Furthermore, we compared the experimental results using Daubechies-8 wavelet and PCA as classical feature extraction methods, and VGG-16 and VGG-19 as deep transfer learning-based feature extraction methods. The pre-trained weights obtained from the ImageNet dataset were used, since the images in ImageNet have a resolution of 224 × 224, which matches the CCTV image frame input. Additionally, ImageNet has approximately 14 million images grouped into 1000 various categories. The use of a model pre-trained on ImageNet certainly improves learning outcomes on violence datasets and ensures good recognition performance.

We divided the training and test data using 10-fold cross-validation. CNN, LSTM, and BiGRU classification algorithms were used. The parameters used for the evaluation of the model included the following: accuracy, recall, precision, and F1 score. We also considered performance of the model in terms of the time required for feature extraction, training, and testing for each dataset. The experimental results are listed in Tabs. 3, 4, and 5.

Table 4 indicates that a combination of ResNet152V2 and BiGRU produces the maximum accuracy of 1.000 in the hockey dataset. In addition to achieving the highest degree of precision, the ResNet152V2-BiGRU combination produced the best precision, recall and F1 score values of 1.000 in each metric. This shows that the model's ability to classify the two classes is better than that of other algorithm combinations.

As in the hockey dataset, the best accuracy on the crowd data set (1.000) is also obtained when the ResNet152V2-BiGRU combination is used (Tab. 5). If reviewed further, the use of ResNet152V2 for feature extraction improved model performance, as evidenced by the increase in precision, recall, precision, and F1 score, compared with classical and older feature extraction approaches.

However, using deep transfer learning features yields significantly better results when compared with classical feature

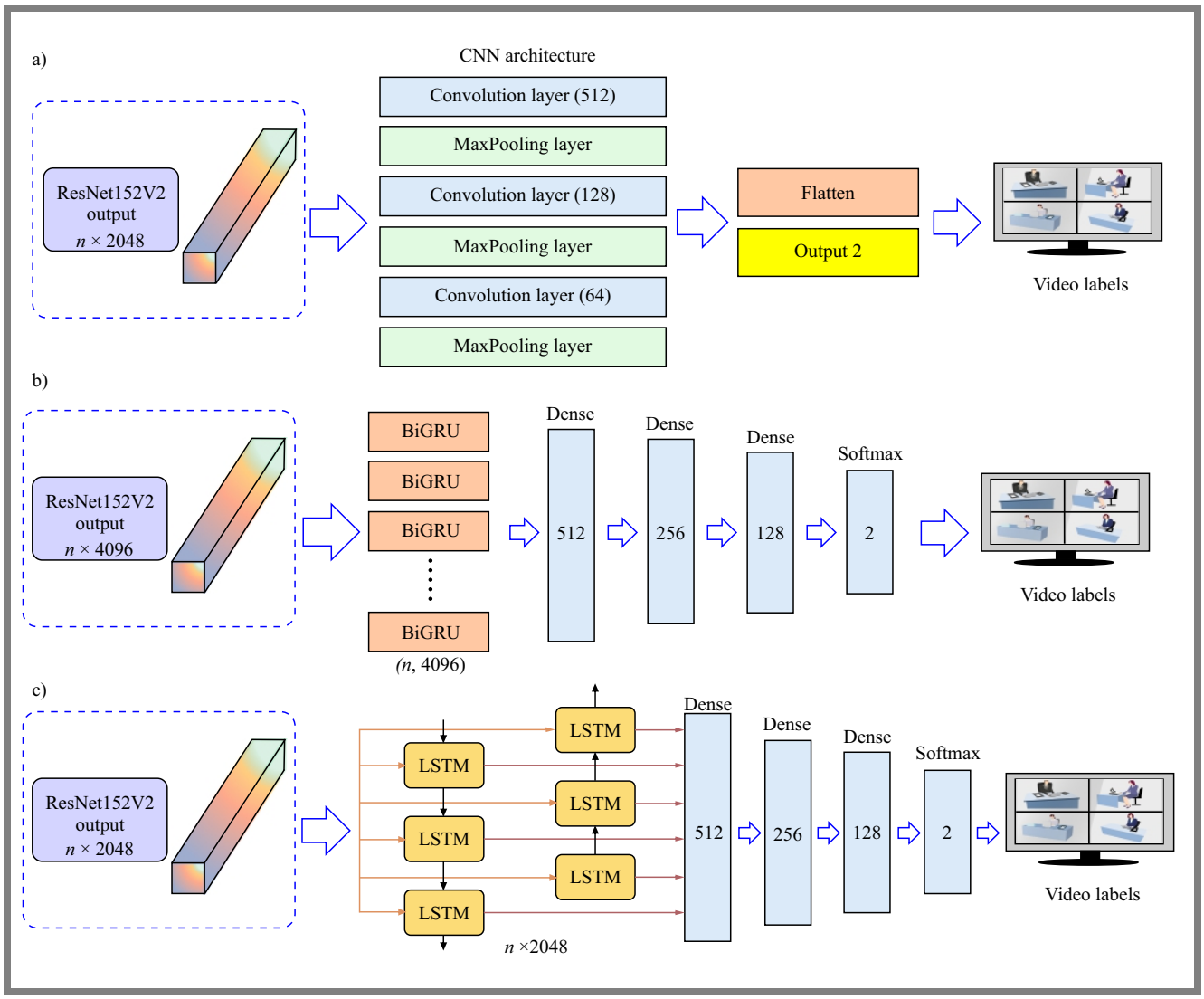

Fig. 8. a) CNN, b) BiGRU, and c) LSTM architectures.

extraction. It can be found that the model built with the crowd dataset using the ResNet152V2-BiGRU combination obtained the best performance, as it achieves the best accuracy and obtains the best metric results (all 1.000).

In contrast to the previous two datasets, the experimental results on the movie dataset were 1.000 for most classification methods and metrics. These excellent metric scores were achieved by all combinations of algorithms, except for BiGRU and its combination with Daubechies-8 wavelets and PCA.

This occurrence can be attributed to the fact that the video in this dataset represents a particular instance of a film scene, where the lighting and camera angles have been deliberately configured. Therefore, the video is clear and does not contain much noise. This differs from the hockey and crowd datasets, which were obtained from surveillance cameras.

Tables 3 – 5 also present the time required to perform feature extraction on a data set and the classification time. One may notice that feature extraction using ResNet152V2 takes longer, but for the training process, ResNet152V2 is faster than VGG-16 and VGG-19. Table 3 also presents the CPU time required to process one test video. The fastest time was obtained using VGG-19. For the crowd data set, ResNet152V2-based feature extraction improves model performance. However, this also increases the time required to process the test data. Upon further analysis, for an increase in accuracy of up to 0.25, a time difference of 0.1 to 0.6 s can be tolerated.

Furthermore, one of the advantages of BiGRU is that in terms of time, it is faster than LSTM, as fewer parameters are used in BiGRU. Consequently, BiGRU is more efficient in terms of memory and time. The results show that BiGRU can perform successfully on all datasets in this study.

Although the proposed ResNet152V2-BiGRU model achieved perfect evaluation metrics (1.000) on the hockey and crowd datasets, it is important to contextualize these results. Its good performance can be attributed to the model's strong capacity for spatio-temporal feature learning on these specific benchmarks. We employed a 10-fold cross-validation strategy to minimize the risk of overfitting and data leakage, and the convergence of training and validation curves (Fig. 9) supports the model's generalization within these datasets.

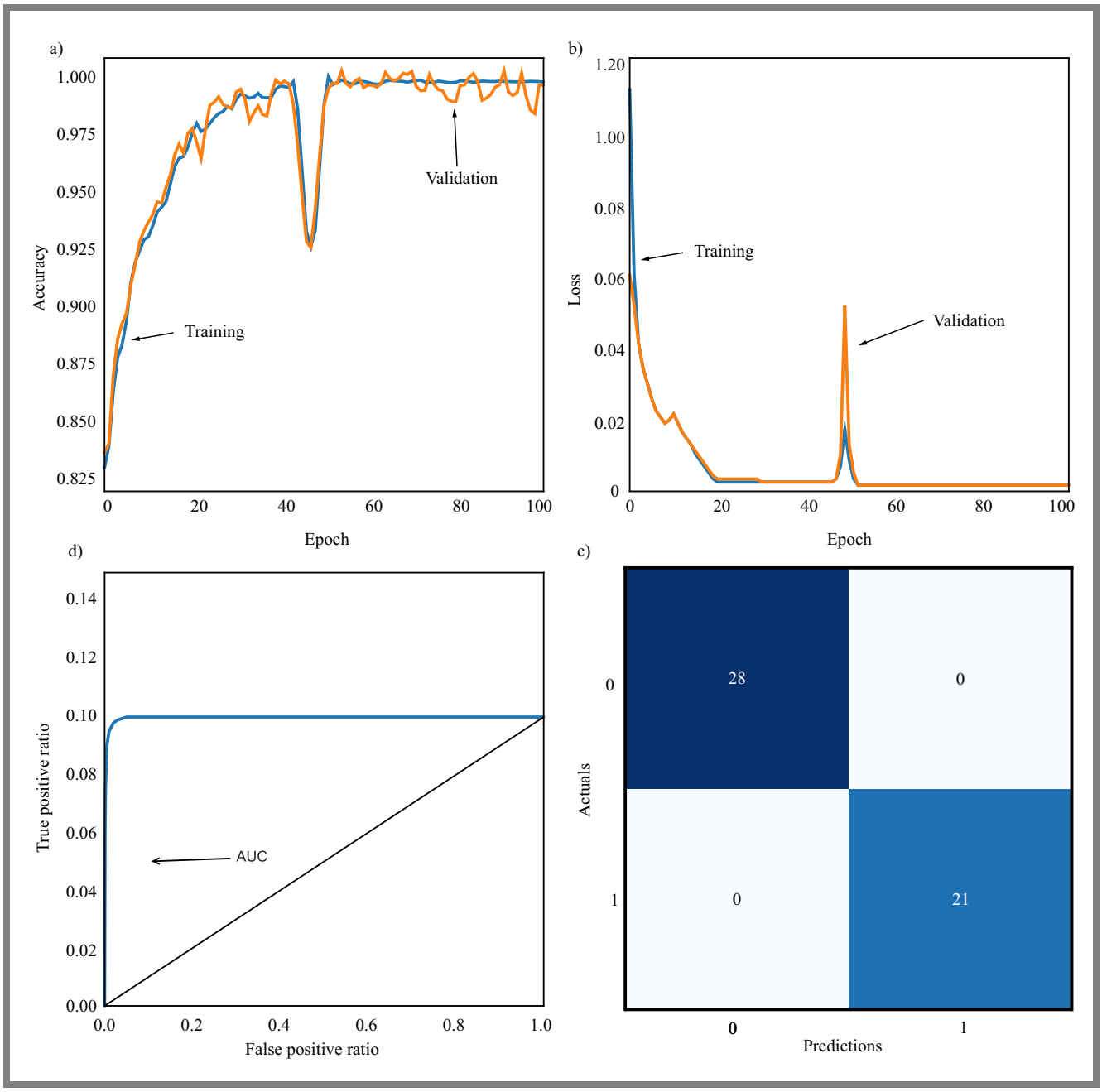

Fig. 9. Performance evaluation of the ResNet152V2 model: a) training accuracy, b) model loss, c) confusion matrix, and d) receiver operating characteristic (ROC) curve.

However, these results, while indicative of a high degree of effectiveness, should be interpreted with the understanding that real-world surveillance footage poses additional unmodeled challenges. The following section discusses computational trade-offs and the need for future validation on larger, more complex real-world streams. The perfect scores show that the ResNet152V2+ BiGRU model can learn optimally on a violent data set to recognize the patterns in each category very effectively.

Figure 9 shows the accuracy and loss results during training and validation. It can be seen that the performance of the model decreased at the 50th epoch but stabilized by the 100th epoch and did not experience overfitting when the results between training and validation almost overlapped and were not significantly different.

In addition, we compared the accuracy of the proposed method with other studies that also used data sets from movies, hockey games, and crowds. Violent event detection using deep transfer learning provides excellent recognition, and almost all models obtained perfect evaluation metrics.

However, not all classifier models correctly detect every relevant class. In Fig. 10, we present a scatter plot of the recognition results for each data instance in the crowd dataset.

Tables 3 – 5 reveal that the best recognition results were obtained using the ResNet152V2 transfer learning model and

Tab. 3. Experimental feature extraction results on the movies datasets – hockey.

Classifier	Feature extraction	Extraction time [s]	Training time [s]	Testing time [s]	Accuracy	Recall	Precision	F1 score
LSTM	PCA	0.043	12.046	0.522	0.810	0.764	0.862	0.811
	Wavelet	0.013	11.514	0.582	0.940	0.915	0.968	0.941
	VGG-16	0.112	43.465	0.536	0.950	0.950	0.951	0.950
	VGG-19	0.106	27.753	0.438	0.970	0.970	0.970	0.970
	ResNet50V2	0.231	22.164	0.415	1.000	1.000	1.000	1.000
	ResNet101V2	0.378	22.300	0.420	1.000	1.000	1.000	1.000
	ResNet152V2	0.492	22.500	0.425	1.000	1.000	1.000	1.000
BiGRU	PCA	0.043	2.079	0.264	0.755	0.806	0.708	0.756
	Wavelet	0.013	3.266	0.900	0.865	0.783	0.957	0.866
	VGG-16	0.112	27.863	0.873	0.975	0.975	0.975	0.975
	VGG-19	0.106	26.900	0.388	0.965	0.965	0.965	0.965
	ResNet50V2	0.231	22.430	0.697	1.000	1.000	1.000	1.000
	ResNet101V2	0.378	22.600	0.700	1.000	1.000	1.000	1.000
	ResNet152V2	0.492	22.800	0.705	1.000	1.000	1.000	1.000
CNN	PCA	0.043	3.700	0.438	0.810	0.886	0.736	0.811
	Wavelet	0.013	15.453	1.060	0.945	0.934	0.957	0.946
	VGG-16	0.112	262.702	0.885	0.915	0.915	0.915	0.915
	VGG-19	0.106	263.022	0.318	0.900	0.900	0.901	0.900
	ResNet50V2	0.231	142.571	0.751	0.990	0.990	0.990	0.990
	ResNet101V2	0.378	143.000	0.760	0.995	0.995	0.995	0.995
	ResNet152V2	0.492	144.000	0.770	1.000	1.000	1.000	1.000

Tab. 4. Experimental feature extraction results on the movies datasets.

Classifier	Feature extraction	Extraction time [s]	Training time [s]	Testing time [s]	Accuracy	Recall	Precision	F1 score
LSTM	PCA	0.043	2.525	0.535	0.825	0.882	0.750	0.822
	Wavelet	0.013	2.904	0.459	1.000	1.000	1.000	1.000
	VGG-16	0.112	13.067	0.592	1.000	1.000	1.000	1.000
	VGG-19	0.106	12.243	0.423	1.000	1.000	1.000	1.000
	ResNet50V2	0.231	12.050	0.429	1.000	1.000	1.000	1.000
	ResNet101V2	0.378	12.180	0.435	1.000	1.000	1.000	1.000
	ResNet152V2	0.492	12.350	0.440	1.000	1.000	1.000	1.000
BiGRU	PCA	0.043	5.178	0.530	0.825	0.842	0.800	0.825
	Wavelet	0.013	5.484	0.440	0.975	0.950	1.000	0.975
	VGG-16	0.112	22.319	0.526	1.000	1.000	1.000	1.000
	VGG-19	0.106	13.120	0.362	1.000	1.000	1.000	1.000
	ResNet50V2	0.231	10.503	0.394	1.000	1.000	1.000	1.000
	ResNet101V2	0.378	10.650	0.402	1.000	1.000	1.000	1.000
	ResNet152V2	0.492	10.800	0.408	1.000	1.000	1.000	1.000
CNN	PCA	0.043	3.324	0.267	1.000	1.000	1.000	1.000
	Wavelet	0.013	4.519	0.902	1.000	1.000	1.000	1.000
	VGG-16	0.112	49.161	0.225	1.000	1.000	1.000	1.000
	VGG-19	0.106	82.547	0.157	1.000	1.000	1.000	1.000
	ResNet50V2	0.231	30.721	0.540	1.000	1.000	1.000	1.000
	ResNet101V2	0.378	31.000	0.550	1.000	1.000	1.000	1.000
	ResNet152V2	0.492	31.500	0.560	1.000	1.000	1.000	1.000

Tab. 5. Experimental feature extraction results on the movies datasets – crowd.

Classifier	Feature extraction	Extraction time [s]	Training time [s]	Testing time [s]	Accuracy	Recall	Precision	F1 score
LSTM	PCA	0.043	3.438	0.554	0.490	0.474	0.375	0.474
	Wavelet	0.013	2.797	0.526	0.625	0.583	0.667	0.624
	VGG-16	0.112	23.860	0.435	0.980	0.980	0.981	0.980
	VGG-19	0.106	23.441	0.402	0.939	0.939	0.946	0.939
	ResNet50V2	0.231	11.406	0.757	1.000	1.000	1.000	1.000
	ResNet101V2	0.378	11.550	0.760	1.000	1.000	1.000	1.000
	ResNet152V2	0.492	11.700	0.765	1.000	1.000	1.000	1.000
BiGRU	PCA	0.043	2.120	0.522	0.500	0.500	0.250	0.433
	Wavelet	0.013	2.537	0.511	0.656	0.690	0.625	0.657
	VGG-16	0.112	14.332	0.360	1.000	1.000	1.000	1.000
	VGG-19	0.106	22.837	0.360	1.000	1.000	1.000	1.000
	ResNet50V2	0.231	11.975	0.368	1.000	1.000	1.000	1.000
	ResNet101V2	0.378	12.100	0.375	1.000	1.000	1.000	1.000
	ResNet152V2	0.492	12.250	0.380	1.000	1.000	1.000	1.000
CNN	PCA	0.043	3.302	0.355	0.592	0.563	0.750	0.574
	Wavelet	0.013	3.752	0.329	0.667	0.750	0.583	0.661
	VGG-16	0.112	82.546	1.384	0.898	0.898	0.915	0.897
	VGG-19	0.106	57.631	0.164	0.694	0.694	0.691	0.690
	ResNet50V2	0.231	41.594	0.937	0.980	0.980	0.980	0.980
	ResNet101V2	0.378	42.000	0.945	0.985	0.985	0.985	0.985
	ResNet152V2	0.492	42.500	0.950	1.000	1.000	1.000	1.000

recognition comparisons were performed using the BiGRU, LSTM, and CNN models. On the 49th test data, four were miss-classified when the CNN+ ResNet152V2 model was used, while neither the BiGRU+ ResNet152V2 model nor the LSTM+ ResNet152V2 model output any misclassifications.

Figure 11 shows the detection results for each video in the video test data, by including the probability of recognizing violence and non-violence. The recognition results show the prediction results of the BiGRU+ ResNet152V2 combination, which is the best of the compared models. This model was then tested in the crowd, movies, and hockey datasets. Each image in the left column has a ground truth class of "violence" and each image in the right column has a ground truth class of "non-violence". The prediction results for each video show that the detection results are the same as the ground truth, with a high confidence rate for each class.

5.1. Computational Efficiency and Real-time Feasibility

Computational efficiency is a critical consideration for the deployment of AI models in real world systems. As shown in Tab. 3, there is a clear trade-off between model performance and processing time. Although ResNet152V2 has the longest feature extraction time (0.492 s per image), it yields the highest accuracy. To assess real-time feasibility, we consider the processing time per video clip. For the crowd dataset, the total test time for the ResNet152V2-BiGRU model was 0.38 s per video. Assuming a standard video clip length of a few seconds, this demonstrates good potential for near-real-time analysis in a processed clip-based system.

However, for true real-time streaming at standard frame rates (e.g., 25 – 30 fps), the current model requires optimization. Future work will focus on employing more efficient feature extractors (e.g., MobileNet, EfficientNet) and model compression techniques (e.g., pruning, quantization) to bridge this gap without a significant sacrifice in accuracy. BiGRU's faster training and testing time compared to LSTM, due to its simpler gating mechanism, is a positive step towards this goal.

In terms of the complexity and time consumption of the proposed model, it can be seen in Tabs. 3 – 5 that each deep-transfer learning model has a different extraction time. The longest feature extraction time was obtained using ResNet152V2 with an execution time of 0.492 s for each image, while the fastest feature extraction execution time was achieved for the wavelet, with an execution time of 0.013 s. ResNet152V2 has the longest extraction time, where the transfer learning process is quite complex because it uses many residual networks, causing the learning process to take longer than in the case of other transfer learning models.

The longest training process was that of CNN with VGG-19 feature extraction (263.022 s). A comparison with other studies is presented in Tab. 6. In the movies dataset, the proposed method outperforms the other methods with the highest scoring accuracy of 100%.

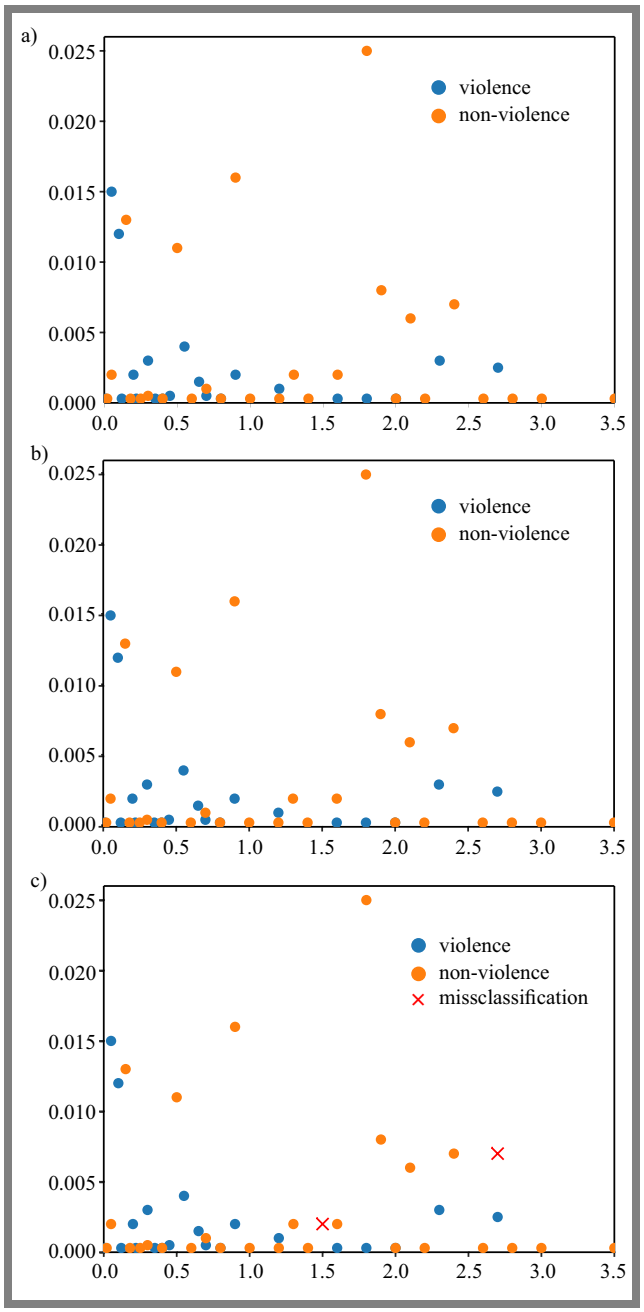

Fig. 10. Scatter plot of: a) BiGRU + ResNet152V2, b) LSTM + ResNet152V2, and c) CNN +ResNet152V2.

6. Ethical Considerations

The deployment of AI-based violence detection systems in public spaces requires a serious discussion focusing on ethical implications. Continuous video monitoring and analysis inherently raise privacy concerns. It is imperative that such systems are deployed in compliance with data protection regulations (e.g., GDPR). Strategies such as on-edge processing, where video data is analyzed locally without being stored or transmitted, can help mitigate privacy risks.

AI models can perpetuate and amplify societal biases if trained on non-representative data. Future work must include rigorous bias auditing across different demographics to ensure that the model does not disproportionately target specific groups.

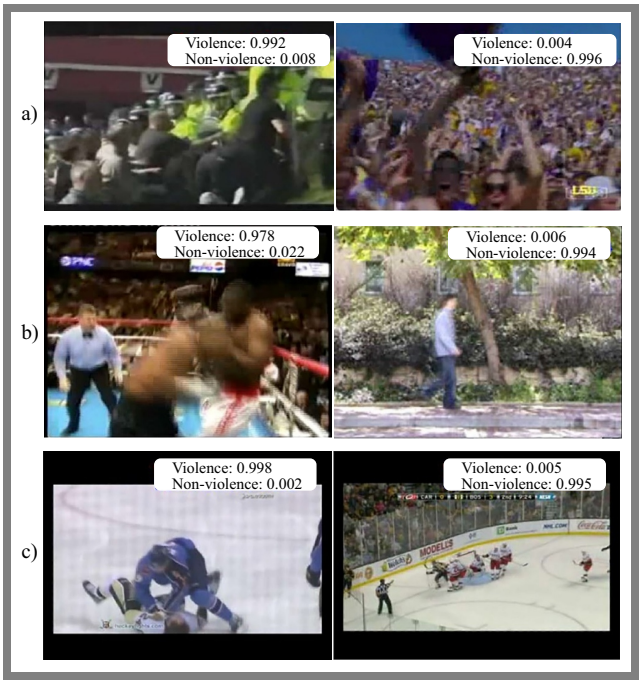

Fig. 11. Detection performance of BiGRU + ResNet152V2 in the: a) crowd, b) movies, and c) hockey datasets.

A false positive, where a non-violent act is flagged as violent, could lead to unnecessary alarm, wasted security resources, and potentially serious confrontations. Therefore, achieving high precision is not just a technical goal but an ethical imperative. In practice, such systems should function as an assistive tool for human operators who make the final decision, rather than as a fully autonomous response trigger.

Transparency in system capabilities and limitations, along with clear governance frameworks, is essential for the responsible development and deployment of this technology.

7. Conclusions

In this study, a comparative analysis of various solutions capable of detecting acts of violence in videos was conducted. The key finding is that the combination of ResNet152V2 for spatial feature extraction and BiGRU for temporal modeling represents a highly effective and efficient architecture, as validated by its top-tier performance with the use of the movies, hockey, and crowd data sets. ResNet50V2, ResNet101V2 and ResNet152V2 were used for feature extraction, while classical (wavelet and PCA), and other deep transfer learning methods (VGG-16 and VGG-19) were used as comparison methods.

Furthermore, CNN, LSTM, and BiGRU algorithms were used for classification. The best precision results in the hockey dataset were obtained when using the ResNet152V2-BiGRU combination. Furthermore, in the movies dataset, all combinations of algorithms achieved excellent performance (1.000). Similarly to the hockey dataset, the best accuracy on the crowd data set was achieved using the ResNet152V2-BiGRU combination. Furthermore, ResNet152V2-BiGRU provides the best accuracy, recall, precision, and F1 score performance.

Tab. 6. Comparison of the proposed violence detection system with state-of-the-art approaches.

		Accuracy [%]		
Ref.	Method	Movies	Hockey	Crowd
[10]	HOMO	–	89.3	76.8
[15]	Violence4D (4D-CNN)	100	100	97.29
[16]	ViF	96.7	81.6	81.2
[28]	MoWLD + BoW	–	91.9	82.5
	MoWLD + SparseCoding	–	93.7	86.3
[29]	ConFeat	96.5	94.4	80.9
[30]	sHOT	–	–	82.9
[31]	DIMOLIF	–	88.6	85.8
[32]	LHOF + BoW	–	–	86.5
[33]	BoW (MoBSIFT) + MF	98.9	90.3	–
[34]	AlexNet + 3D-CNN	98.7	92.9	88.0
[35]	Xception + LSTM	–	91	88
	InceptionV3 + LSTM	–	90	89
[36]	OViF	–	84.2	76.8
[37]	3D CNN + interest frames	100	99.4	97.49
[38]	Hough forest + 2D CNN	99	94.6	–
[39]	Modified 3D CNN	99.97	98.96	–
[40]	Object detection + LSTM	–	98	98.21
[41]	Object detection + 3D CNN	99.9	96	98
[42]	Two-cascade TSM	–	98.995	97.959
[43]	Dual-stream CNN + echo state network	–	99	99.01
[44]	Vision-based fight detection	100	98	–
[45]	Edge Vision	–	98.5	–
[46]	EvoKeyNet + DeepkeyFrm	–	98.98	99.29
[47]	2D CNN + ESM + STA	100	99.7	98.53
Proposed	ResNet152V2 + BiGRU	100	100	100

The experimental results obtained in the course of this study show that BiGRU performs better in terms of time than LSTM. BiGRU also achieves good performance on all data sets used in this study. The ResNet152V2-BiGRU combination achieves the best accuracy and F1 score values on all datasets.

In general, using ResNet152V2 for feature extraction improves the performance of the model on all datasets, but this increases the time required to process the test data. A difference of approximately 6 s can still be tolerated for the crowd dataset considering that the accuracy obtained increased to 0.263.

Limitations and Future Work

Despite the promising results, this study is limited by the scale and scope of the benchmark datasets used. The models were trained and tested on controlled datasets which may not fully capture the challenges of real-world surveillance, such as severe occlusions, extreme lighting variations, dynamic camera angles, as well as dense and complex crowd behavior. Consequently, the performance reported here might not directly translate to operational environments.

A primary direction for future research is to validate and retrain the proposed model on larger, more diverse, and more challenging real-world video datasets. Furthermore, exploring the model's robustness to attacks and its performance in low-resolution, long-duration video streams will be essential for practical public safety applications.

References

[1] S. Natha *et al.*, "Deep BiLSTM Attention Model for Spatial and Temporal Anomaly Detection in Video Surveillance", *Sensors*, vol. 25, art. no. 251, 2025 (https://doi.org/10.3390/s25010251).

[2] M. Karim *et al.*, "Human Action Recognition Systems: A Review of the Trends and State-of-the-art", *IEEE Access*, vol. 12, pp. 36372–36390, 2024 (https://doi.org/10.1109/access.2024.3373199).

[3] Ragedhaksha, Darshini, Shahil, and J. Arunnehru, "Deep Learning-based Real-world Object Detection and Improved Anomaly Detection for Surveillance Videos", *Materials Today: Proceedings*, vol. 80, pp. 2911–2916, 2023 (https://doi.org/10.1016/j.matpr.2021.07.064).

[4] S. Singla and R. Chadha, "Detecting Criminal Activities from CCTV by Using Object Detection and Machine Learning Algorithms", *2023 3rd International Conference on Intelligent Technologies (CONIT)*, Hubli, India, 2023 (https://doi.org/10.1109/conit59222.2023.10205699).

[5] V. Payghode *et al.*, "Object Detection and Activity Recognition in Video Surveillance Using Neural Networks", *International Journal of Web Information Systems*, vol. 19, pp. 123–138, 2023 (https://doi.org/10.1108/ijwis-01-2023-0006).

[6] P. Negre *et al.*, "Literature Review of Deep-learning-based Detection of Violence in Video", *Sensors*, vol. 24, art. no. 4016, 2024 (https://doi.org/10.3390/s24124016).

[7] T. Santos, H. Oliveira, and A. Cunha, "Systematic Review on Weapon Detection in Surveillance Footage through Deep Learning", *Computer Science Review*, vol. 51, art. no. 100612, 2024 (https://doi.org/10.1016/j.cosrev.2023.100612).

[8] J. Ruiz-Santaquiteria *et al.*, "Improving Handgun Detection through a Combination of Visual Features and Body Pose-based Data", *Pattern Recognition*, vol. 136, art. no. 109252, 2023 (https://doi.org/10.1016/j.patcog.2022.109252).

[9] L.M. Salim and Y. Celik, "Detection of Dangerous Human Behavior by Using Optical Flow and Hybrid Deep Learning", *Electronics*, vol. 13, art. no. 2116, 2024 (https://doi.org/10.3390/electronics13112116).

[10] J. Mahmoodi and A. Salajegheh, "A Classification Method Based on Optical Flow for Violence Detection", *Expert Systems with Applications*, vol. 127, pp. 121–127, 2019 (https://doi.org/10.1016/j.eswa.2019.02.032).

[11] X. Wang, J. Yang, and N. K. Kasabov, "Integrating Spatial and Temporal Information for Violent Activity Detection from Video Using Deep Spiking Neural Networks", *Sensors*, vol. 23, art. no. 4532, 2023 (https://doi.org/10.3390/s23094532).

[12] L. Hsairi, S.M. Alosaimi, and G.A. Alharaz, "Violence Detection Using Deep Learning", *Arabian Journal for Science and Engineering*, vol. 50, pp. 11669–11679, 2024 (https://doi.org/10.1007/s13369-024-09536-y).

[13] S.G. Jaiswal, S.W. Mohod, D. Sharma, and A. Hinge, "Violent Video Classification with Transfer Learning Approach Using Inception-V3 and Support Vector Machine", *Indian Journal of Science and Technology*, vol. 16, pp. 3018–3026, 2023 (https://doi.org/10.17485/ijst/v16i37.1972).

[14] M.Q. Khan, S.N. Sabir, F. Malik, and M. Khan, "Deep Convolutional Network for Automatic Violence Detection in Surveillance Videos Using Transfer Learning", *Kashf Journal of Multidisciplinary Research*, vol. 2, pp. 251–275, 2025 (https://doi.org/10.71146/kjmr270).

[15] M. Magdy, M.W. Fakhr, and F.A. Maghraby, "Violence 4D: Violence Detection in Surveillance Using 4D Convolutional Neural Networks", *IET Computer Vision*, vol. 17, pp. 282–294, 2023 (https://doi.org/10.1049/cvi2.12162).

[16] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent Flows: Real-time Detection of Violent Crowd Behavior", *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Providence, USA, 2012 (https://doi.org/10.1109/cvprw.2012.6239348).

[17] S. Roshan, G. Srivathsan, K. Deepak, and S. Chandrakala, "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*, Academic Press, pp. 249–270, 2020 (https://doi.org/10.1016/b978-0-12-816385-6.00011-8).

[18] A. Gupta, A. Mittal, and R. Jain, "A Novel Sarcasm Detection Approach for Text-image Data: Leveraging Multimodal Fusion and Weighted Latent Factors", *Information Fusion*, vol. 103, art. no. 103266, 2025 (https://doi.org/10.1016/j.inffus.2025.103266).

[19] N. Sutranggono and R. Sarno, "Detection and Sentiment Analysis Based on Mental Disorders Aspects Using Bidirectional Gated Recurrent Unit and Semantic Similarity", *International Journal of Intelligent Engineering and Systems*, vol. 17, pp. 1–12, 2024 (https://doi.org/10.22266/ijies2024.0831.01).

[20] U. Jaishankar, J.H. Nirmal, and G. Gidaye, "Robust Time Domain Scalogram Filter Bank Feature Learning Model for Speech Depression Detection with Metaheuristic Spatio Temporal Residual BIGRU Model", *International Journal of Biomedical Engineering and Technology*, vol. 47, pp. 348–382, 2025 (https://doi.org/10.1504/ijbet.2025.145219).

[21] E.B. Nievas, O.D. Suarez, G.B. García, and R. Sukthankar, "Violence Detection in Video Using Computer Vision Techniques", *Lecture Notes in Computer Science*, vol. 6855, pp. 332–339, 2011 (https://doi.org/10.1007/978-3-642-23678-5_39).

[22] C.M. Akujuobi, "Wavelets", in: *Wavelets and Wavelet Transform Systems and Their Applications: A Digital Signal Processing Approach*, Switzerland: Springer, pp. 13–44, 2022 (https://doi.org/10.1007/978-3-030-87528-2_2).

[23] M. Shafiq and Z. Gu, "Deep Residual Learning for Image Recognition: A Survey", *Applied Science*, vol. 12, art. no. 8972, 2022 (https://doi.org/10.3390/app12188972).

[24] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition", *arXiv*, 2014 (https://doi.org/10.48550/arXiv.1409.1556).

[25] X. Yu, Z. Yu, and S. Ramalingam, "Learning Strict Identity Mappings in Deep Residual Networks", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018 (https://doi.org/10.1109/cvpr.2018.00466).

[26] R.S. Kızıltepe, J.Q. Gan, and J.J. Escobar, "A Novel Keyframe Extraction Method for Video Classification Using Deep Neural Networks", *Neural Computing and Applications*, vol. 35, pp. 24513–24524, 2023 (https://doi.org/10.1007/s00521-021-06322-x).

[27] S.H. Hendi, H.B. Taher, and K.Q. Hussein, "Automated Video Events Detection and Classification Using CNN-GRU Model", *Wasit Journal of Computer and Mathematics Science*, vol. 2, pp. 77–86, 2023 (https://doi.org/10.31185/wjcms.188).

[28] T. Zhang *et al.*, "MoWLD: A Robust Motion Image Descriptor for Violence Detection", *Multimedia Tools and Applications*, vol. 76, pp. 1419–1438, 2017 (https://doi.org/10.1007/s11042-015-3133-0).

[29] S. Keceli and A.Y. Kaya, "Violent Activity Detection with Transfer Learning Method", *Electronics Letters*, vol. 53, pp. 1047–1048, 2017 (https://doi.org/10.1049/el.2017.0970).

[30] H. Rabiee, H. Mousavi, M. Nabi, and M. Ravanbakhsh, "Detection and Localization of Crowd Behavior Using a Novel Tracklet-based Model", *International Journal of Machine Learning and Cybernetics*, vol. 9, pp. 1999–2010, 2018 (https://doi.org/10.1007/s13042-017-0682-8).

[31] A.B. Mabrouk and E. Zagrouba, "Spatio-temporal Feature Using Optical Flow Based Distribution for Violence Detection", *Pattern Recognition Letters*, vol. 92, pp. 62–67, 2017 (https://doi.org/10.1016/j.patrec.2017.04.015).

[32] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violence Detection in Surveillance Video Using Low-level Features", *PLOS.One*, vol. 13, art. no. e0203668, 2018 (https://doi.org/10.1371/journal.pone.0203668).

[33] I.P. Febin, K. Jayasree, and P.T. Joy, "Violence Detection in Videos for an Intelligent Surveillance System Using MoBSIFT and Movement Filtering Algorithm", *Pattern Analysis and Applications*, vol. 23, pp. 611–623, 2020 (https://doi.org/10.1007/s10044-019-00821-3).

[34] A.S. Keceli and A. Kaya, "Violent Activity Classification with Transferred Deep Features and 3D-CNN", *Signal, Image and Video Processing*, vol. 17, pp. 139–146, 2023 (https://doi.org/10.1007/s11760-022-02213-3).

[35] M.A. Soeleman, C. Supriyanto, and D.P. Prabowo, "An Empirical Study of CNN-LSTM on Class Imbalance Datasets for Violence Video Detection", *Proc. of the 2021 International Conference on Computer, Control, Informatics and Its Applications*, pp. 81–85, 2021 (https://doi.org/10.1145/3489088.3489126).

[36] Y. Gao *et al.*, "Violence Detection Using Oriented Violent Flows", *Image and Vision Computing*, vol. 48, pp. 37–41, 2016 (https://doi.org/10.1016/j.imavis.2016.01.006).

[37] J. Mahmoodi, H. Nezamabadi-pour, and D. Abbasi-Moghadam, "Violence Detection in Videos Using Interest Frame Extraction and 3D Convolutional Neural Network", *Multimedia Tools and Applications*, vol. 81, pp. 20945–20961, 2022 (https://doi.org/10.1007/s11042-022-12532-9).

[38] I. Serrano, O. Deniz, J.L. Espinosa-Aranda, and G. Bueno, "Fight Recognition in Video Using Hough Forests and 2D Convolutional Neural Network", *IEEE Transactions on Image Processing*, vol. 27, pp. 4787–4797, 2018 (https://doi.org/10.1109/TIP.2018.2845742).

[39] W. Song *et al.*, "A Novel Violent Video Detection Scheme Based on Modified 3D Convolutional Neural Networks", *IEEE Access*, vol. 7, pp. 39172–39179, 2019 (https://doi.org/10.1109/access.2019.2906275).

[40] F.U.M. Ullah *et al.*, "An Intelligent System for Complex Violence Pattern Analysis and Detection", *International Journal of Intelligent Systems*, vol. 37, pp. 10400–10422, 2022 (https://doi.org/10.1002/int.22537).

[41] F.U.M. Ullah *et al.*, "Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network", *Sensors*, vol. 19, art. no. 2472, 2019 (https://doi.org/10.3390/s19112472).

[42] Q. Liang, Y. Li, B. Chen, and K. Yang, "Violence Behavior Recognition of Two-cascade Temporal Shift Module with Attention Mechanism", *Journal of Electronic Imaging*, vol. 30, art. no. 043009, 2021 (https://doi.org/10.1117/1.jei.30.4.043009).

[43] W. Ullah *et al.*, "Intelligent Dual Stream CNN and Echo State Network for Anomaly Detection", *Knowledge-Based Systems*, vol. 253, art. no. 109456, 2022 (https://doi.org/10.1016/j.knosys.2022.109456).

[44] Ş. Aktı, G.A. Tataroğlu, and H.K. Ekenel, "Vision-based Fight Detection from Surveillance Cameras", *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Istanbul, Turkey, 2019 (https://doi.org/10.1109/IPTA.2019.8936070).

[45] F.U.M. Ullah *et al.*, "AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks", *IEEE Transactions on Industrial Informatics*, vol. 18, pp. 5359–5370, 2021 (https://doi.org/10.1109/TII.2021.3116377).

[46] M. Shoaib *et al.*, "Augmenting the Robustness and Efficiency of Violence Detection Systems for Surveillance and Non-surveillance Scenarios", *IEEE Access*, vol. 11, pp. 123295–123313, 2023 (https://doi.org/10.1109/access.2023.3329062).

[47] J. Mahmoodi and H. Nezamabadi-pour, "A Spatio-temporal Model for Violence Detection Based on Spatial and Temporal Attention Modules and 2D CNNs", *Pattern Analysis and Applications*, vol. 27, art. no. 46, 2024 (https://doi.org/10.1007/s10044-024-01265-0).

Khaled Merit, Ph.D.
Laboratory of TIT, Department of Electrical Engineering
https://orcid.org/0000-0002-7762-1898
E-mail: merit.khaled@univ-bechar.dz
Tahri Mohammed University of Bechar, Algeria
https://www.univ-bechar.dz

Mohammed Beladgham, Ph.D., Full Professor
Laboratory of TIT, Department of Electrical Engineering
https://orcid.org/0000-0002-2371-6859
E-mail: beladgham.mohammed@univ-bechar.dz
Tahri Mohammed University of Bechar, Algeria
https://www.univ-bechar.dz