Received: 5 June 2023 | Revised: 19 October 2023 | Accepted: 8 January 2024

DOI: 10.4218/etrij.2023-0222

ORIGINAL ARTICLE

ETRI Journal WILEY

Violent crowd flow detection from surveillance cameras using deep transfer learning–gated recurrent unit

Elly Matul Imah | **Riskyana Dewi Intan Puspitasari**

Data Science Department, Universitas Negeri Surabaya, Surabaya, Indonesia

Correspondence
Elly Matul Imah, Data Science Department, Universitas Negeri Surabaya, Ketintang, Surabaya, East Java, Indonesia.
Email: ellymatul@unesa.ac.id

Funding information
Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia, Grant/Award Numbers: 205/E5/PG.02.00.PM/2023, B/65354/UN38.III.1/LK.04.00/2023

Abstract

Violence can be committed anywhere, even in crowded places. It is hence necessary to monitor human activities for public safety. Surveillance cameras can monitor surrounding activities but require human assistance to continuously monitor every incident. Automatic violence detection is needed for early warning and fast response. However, such automation is still challenging because of low video resolution and blind spots. This paper uses ResNet50v2 and the gated recurrent unit (GRU) algorithm to detect violence in the Movies, Hockey, and Crowd video datasets. Spatial features were extracted from each frame sequence of the video using a pretrained model from ResNet50V2, which was then classified using the optimal trained model on the GRU architecture. The experimental results were then compared with wavelet feature extraction methods and classification models, such as the convolutional neural network and long short-term memory. The results show that the proposed combination of ResNet50V2 and GRU is robust and delivers the best performance in terms of accuracy, recall, precision, and F1-score. The use of ResNet50V2 for feature extraction can improve model performance.

KEYWORDS
deep learning, deep transfer learning, video processing, violence detection

1 | INTRODUCTION

Violence is the intentional use of physical force to injure, abuse, or threaten a person or group. Violence is a crime that violates the laws, regulations, norms, and values that prevail in society. These crimes can harm someone both physically and mentally or can kill someone. WHO reported 1.25 million deaths caused by injuries related to violence in 2019. There have been 23 294 cases of violence in Indonesia, including 3822 male and 21 201 female victims. Furthermore, from the data obtained, the most significant number of those who experienced this violence were 13–17 years old, that is, children or teenagers. The large amount of violence experienced by children and adolescents is a major concern because it will impact their future. Thus, it is necessary to prevent violence by reducing the injuries and fatalities caused by it. Currently, the use of closed-circuit television (CCTV) is increasing because it can carry out 24-h surveillance, which humans cannot do. The camera records all the events from various angles. As a result, a large amount of video data still requires a human to identify anomalous activity such as violence. This video-monitoring process requires significant time and effort if performed

manually. Therefore, it is necessary to have an automatic detection system that will accelerate the monitoring process. One of the challenges faced when performing automatic detection is the low resolution of the video produced by CCTV [1]. For several years, research on automatic detection has been carried out, and violence detection is similar to action recognition [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 influence whether an action is categorized as an act of violence or just an ordinary movement. The authors of [3–5] detect objects in CCTV video data. The author of [6] detects the use of weapons, which indicates the occurrence of acts of violence in a video. The authors of [7, 8] conducted similar research. However, not all acts of violence, such as hand-to-hand fights or beatings, use weapons. Therefore, it is necessary to detect acts of violence that do not depend on weapons. Several studies have been conducted on the detection of violence, using various approaches. The author of [9] used a histogram of optical flow (HOF) to extract valuable features from videos, whereas [10] used HOF magnitude and orientation (HOMO). The researchers of [11] extracted motion features from RGB dynamic images. A different approach was taken by [12], who used residual network 50 (ResNet50), VGG-19, and Xception, which are convolutional neural network (CNN) architectures trained on the ImageNet dataset. Their results are reasonably good. The studies [13, 14] used VGG-16 for feature extraction and a simple classification algorithm, namely, SVM. Better results were obtained in [15], which used ResNet50 as the backbone for three-dimensional CNNs and dense optical flow for the region of interest. The detection of acts of violence from surveillance cameras faces difficulties such as video quality, because surveillance cameras are sometimes placed in low-light places. Another difficulty occurs with surveillance cameras in public places, especially in crowds. One dataset containing videos of public crowds is the Crowd dataset, also called the Violent Flow dataset. Several studies have used the Crowd dataset. The author of [16] used a Violent Flow (ViF) descriptor and then classified the output using a linear SVM; the accuracy obtained was 81.3%. Using the same classification algorithm combined with HOF, the researchers of [17] yielded an accuracy of 83.37%. That accuracy still needs to be improved to create an accurate detection system. This dataset is challenging because acts of violence are sometimes not visible because of the crowd; on the other hand, crowded conditions often lead to false positives. Therefore, in this study, violence was detected using the Crowd dataset to improve the quality of the model in terms of performance and time. A powerful classification model is needed to classify data into appropriate classes, such as long short-term memory (LSTM), gated recurrent units (GRUs), or CNNs. The LSTM model is intended for images and other data such as text data, and it produces good accuracy [18]. This discussion reveals that there are still many challenges in violence detection research, and the need for such detection is increasing. Therefore, in this study, acts of violence were detected using a deep learning approach. We used ResNet50V2 to extract significant features in video. For comparison, this study also used wavelets for feature extraction. This study used a GRU as a classification algorithm. A GRU offers better results than LSTM in predicting the condition of a pulp paper press [19]. In the classification of emotions in noisy speech, the GRU provides a faster runtime and lower error for washing noise than LSTM [20]. For comparison, this research also uses the LSTM and CNN algorithms. The composition of this paper is as follows: violent video preprocessing and feature extraction are described in Section 2. An explanation of the violence classification can be found in Section 3. In Section 4, the dataset used, results, and discussion of the experiments are described. Finally, Section 5 presents the conclusions.

2 | VIOLENCE VIDEO DETECTION

A general scheme for detecting violent acts is illustrated in Figure 1. The first step is to preprocess the video data and then divide them into training data and test data using k-fold validation. Subsequently, feature extraction is performed using ResNet50V2. We also compared the feature extraction obtained using several methods: discrete wavelet transform (DWT), principal component analysis (PCA), VGG-16, and VGG-19. The best feature extraction results from the training data were used to build a violence detection model using the GRU algorithm. We compared the classification model with several algorithms such as LSTM and CNN. The last step was to evaluate the model using the test data. The measurement parameters used to evaluate the model include accuracy, recall, specificity, G-Mean, and CPU time. In addition to ResNet50V2, we also compared wavelet feature extraction methods and no feature extraction to compare the performance in violence detection and extraction time.

2.1 | Dataset

In this study, we used three datasets to evaluate the model's performance in detecting the occurrence of violence in a video. The three datasets included the Movies dataset [21], Hockey dataset [21], and Crowd dataset

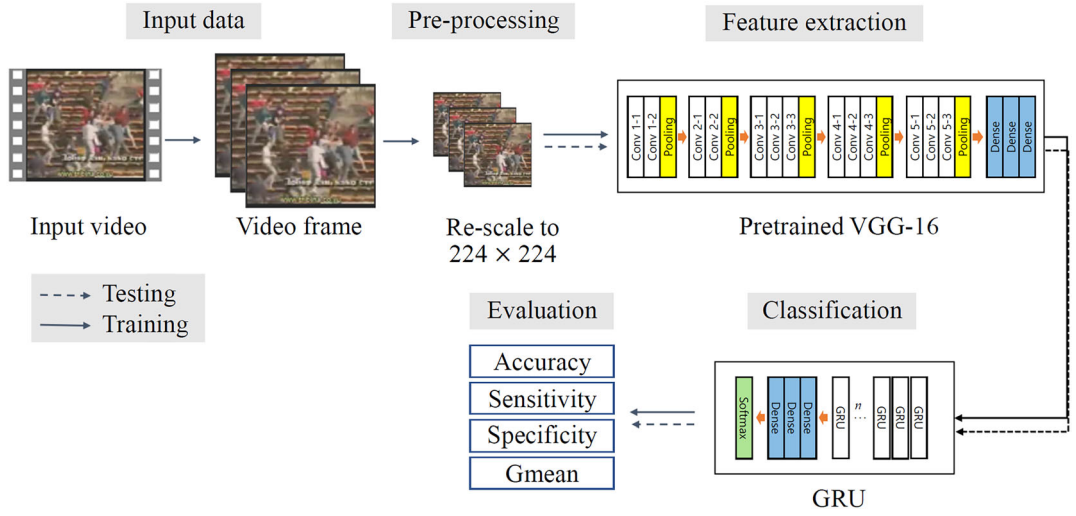

FIGURE 1 Violence detection system.

TABLE 1 Summary of each dataset for violence detection.

Datasets	Format	Average resolution	Total videos	Violent	No violent
Movies dataset	.mpg, .mp4	360 × 250	200	100	100
Hockey dataset	.avi	360 × 288	1000	500	500
Crowd dataset	.avi	360 × 240	246	123	123

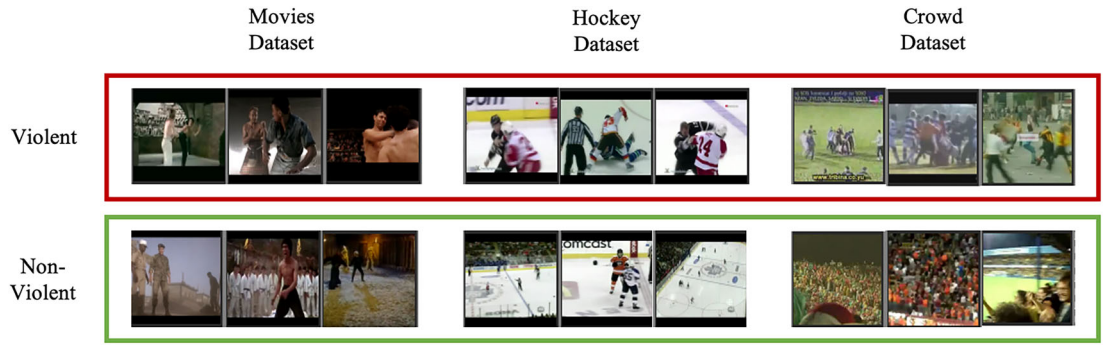

FIGURE 2 Sample frames of violent and non-violent videos from the Movies, Hockey, and Crowd datasets.

note [16]. The videos in the Movies dataset contain several movie scenes and consist of 200 videos divided into 100 fight videos and 100 non-fight videos. The Hockey dataset contains 1000 video recordings of matches from the National Hockey League, divided into 500 violent and 500 nonviolent videos. The Crowd dataset is a real-time video recording of violence in a crowd, containing 246 videos with 123 violent and 123 nonviolent videos. Each dataset was divided into training and test datasets using k-fold validation. A summary of each dataset is presented in Table 1, and the snippets of each dataset are shown in Figure 2.

2.2 | Preprocessing

The first step in building a violence-detection model is preprocessing. In this step, each video is extracted into several images in RGB format. The image set was then resized to 224 × 224 pixels to match the input shape from ResNet50V2. The next step is to obtain the pixel values from each image set. Therefore, in this process, we obtained a matrix measuring $m \times n \times 224 \times 224 \times 3$, where m is the number of videos, n is the number of images captured in each video, and $224 \times 224 \times 3$ is an RGB image 224 × 224 in size.

3 | VIDEO FEATURE EXTRACTION

3.1 | DWT

In this research, wavelet decomposition level 3 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 Daubechies order $N \geq 2$ has $2N$ vanishing moments and a small-scale support with an interval $[0, 2N-1]$ [22]. Daubechies polynomial order $N-1$ is defined in (1).

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

After obtaining a grayscale image, level 1 wavelet decomposition is performed. From the level 1 decomposition, the sub-bands LL, LH, HL, and HH 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. Then, the matrix is reshaped to adjust the input dimensions in the classification process. The results of feature extraction using the DWT are shown in Figure 3.

3.2 | PCA

PCA is a transformation that changes and decomposes a large number of correlated variables into a small number of uncorrelated variables and can reduce the dimensions of the data without eliminating important information in the data. A two-dimensional image can be handled as a one-dimensional vector. If the length of the image is w and the image width is h, then the number of components of the one-dimensional vector is $(w \times h)$. The overall image space is not optimal for representing image information. The basis vector of this feature space of the image is called the principal component and is taken from the eigenvalue decomposition process. PCA produces a feature matrix containing the eigenvectors with the highest eigenvalues that capture the highest data variation. 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 image size. For each dataset, all vectors were aggregated into a matrix of size $(N \times 50\,176)$, where N is the number of images. The next step is to select the principal component value with k% of the total eigenvalues. The results of the feature extraction using PCA are shown in Figure 4.

3.3 | ResNet

ResNet is a method in deep learning and is a development of CNNs. In the learning process, ResNet implements residual connections that can connect layers to other layers by skipping some layers in between. It is claimed that this avoids 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 the results of recognition accuracy, transfer learning is used. Transfer learning is an approach in deep

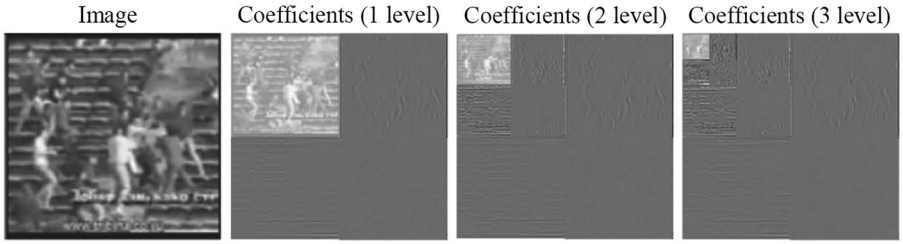

FIGURE 3 Feature extraction process using DWT.

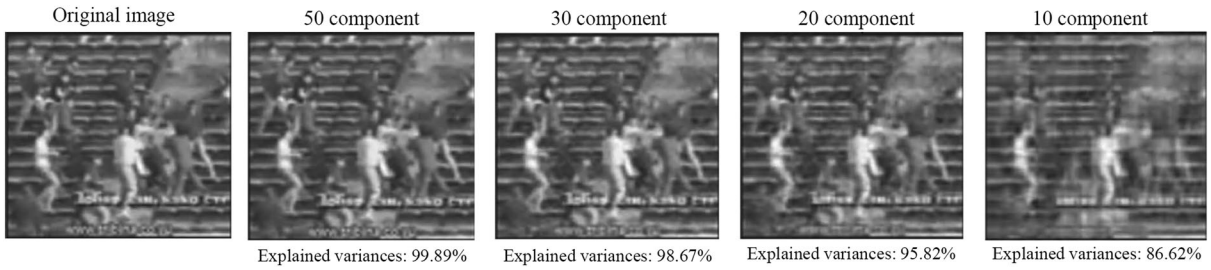

FIGURE 4 Feature extraction process using PCA.

learning (and machine learning) in which knowledge is transferred from one model to another. A common misconception regarding transfer learning is that the training and test datasets must come from the same source or have the same distribution. However, in practice, 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. On the other hand, deep transfer learning using pretrained models trained on other datasets in the same domain can boost classification performance [24]. ResNet50V2 is an improved version of ResNet50 that performs better than the previous version on the ImageNet dataset. Modifications to the ResNet50V2 network include changes to connections between blocks, where the last nonlinearity is removed. In addition, batch normalization and ReLU activation are applied to the input before multiplication with the weight matrix [25]. The ResNet50V2 architecture is illustrated in Figure 5.

In this study, we used ResNet50V2 as the feature extractor for the input video. We then continued the learning process using other deep learning methods, such as LSTM, GRU, and CNN, as classification methods. Figure 5 shows the 3D architecture of the ResNet50V2 network. According to Figure 5, ResNet50V2 consists of five convolution blocks. In the first block, the preprocessed data pass through two convolution layers, followed by a maxpooling layer. In the first block, a matrix with dimensions $m \times n \times 112 \times 112 \times 128$ is generated. As in the first block, the data pass through two convolution layers and one maxpooling layer in the second block. Then, in the third to fifth blocks, the data pass through four convolution layers and one maxpooling layer. The resulting matrix of block 5 ResNet50V2 is $m \times n \times 7 \times 7 \times 512$, which then enter the flatten layer and dense layer to produce a matrix with dimensions of $m \times n \times 4096$. The feature extraction process using ResNet50V2 only stops at the dense layer because the following process is performed using another classifier algorithm.

3.4 | VGG

The VGG is a CNN that has been trained using the ImageNet dataset and was first introduced by [24]. The VGG can handle massive datasets because it contains several weighted layers with millions of parameters. The difference between the VGG-16 and VGG-19 networks is the depth of the weight layers. In VGG-16, the number of weight layers is 16, whereas the VGG-19 has a layer depth of 19. VGG is unique in studying data because it only focuses on convolution layers of 3×3 filters with a stride of 1 and always uses the same padding and maxpool layer of 2×2 filter with a stride of 2. The main advantage of VGG-19 over VGG-16 is that it has more layers, enabling it to learn more complex data representations. Because of this, VGG-19 is heavier and requires additional memory and computational resources. However, in some cases, VGG-16 is more accurate than VGG-19. We used VGG-16 and VGG-19 as a comparison feature extractor for the input video. We then continued the learning process using other deep learning methods, such as LSTM, GRU, and CNN, as classification methods. The resulting matrix of block 5 VGG-16 is $m \times n \times 7 \times 7 \times 512$ in size, and it then enters the flatten layer and dense layer to produce a matrix with dimensions of $m \times n \times 4096$.

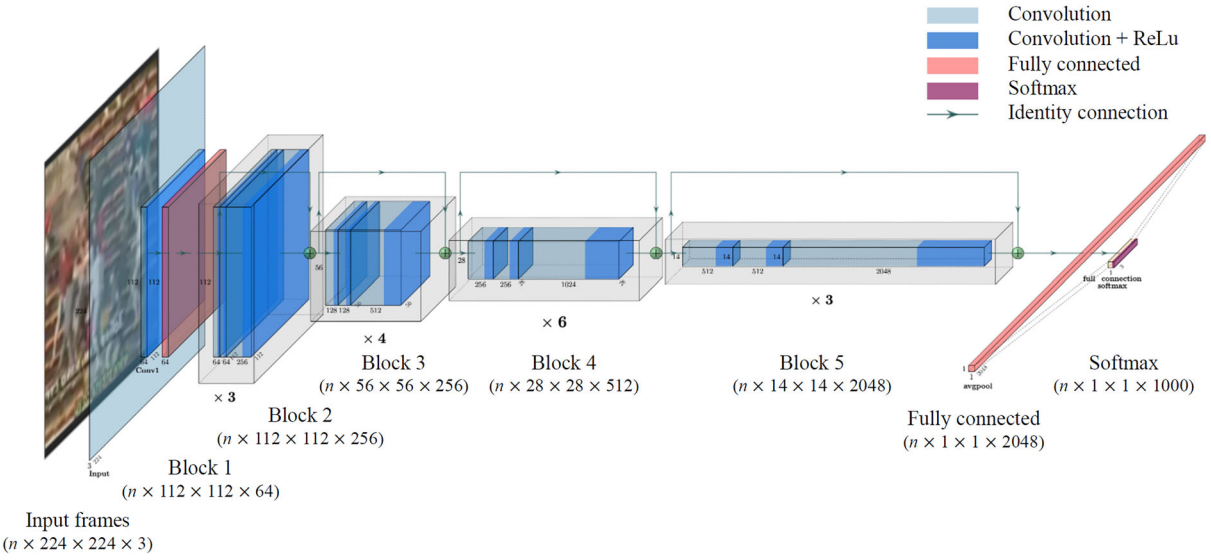

FIGURE 5 Three-dimensional ResNet50V2 architecture.

4 | VIOLENCE CLASSIFICATION

4.1 | CNN

After acquiring the feature set from the pretrained ResNet50V2, we compared several deep learning methods to detect scene violence. One such method is CNN. In addition to using a pretrained CNN, this study also used a self-trained CNN to compare the results obtained when using ResNet50V2. The CNN in this study consists of three convolution and maxpooling layers. The CNN architecture is shown in Figure 6. As shown in the figure, the set of images obtained from preprocessing enters the convolution layer with 512 units. The data are then entered into the maxpooling layer, and so on, until the last convolution and maxpooling layers. Furthermore, the features are flattened with a fully connected layer for classification into two classes. In this study, 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.

4.2 | LSTM

LSTM is an advanced recurrent neural network that solves the vanishing gradient problem. LSTM overcomes the vanishing gradient problem commonly observed in recurrent neural networks by inserting gating functions into the state dynamics [26]. Each LSTM cell has three gates, namely a forget gate, an input gate, and an output gate. The LSTM architecture can be seen on Figure 7. The forget gate determines what information will be stored and forgotten. The input gate is in charge of updating the contents or the contents of memory cells. In this study, the hyperparameter settings for LSTM 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.

4.3 | GRU

The GRU was introduced by [27] with a design similar to that of the LSTM but using a more straightforward memory unit to simplify training and implementation. The implementation of the GRU algorithm begins by determining how much information from the previous unit will be passed to the next unit. The result of the Hadamard operation of the reset gate and the weights determine what information from the previous time step will be removed. The last step is to calculate the output at time step t using (1). In this study, classification was carried out using a GRU. The result of feature extraction using ResNet50V2 passes through the GRU layer in this process. Furthermore, the resulting matrix from the GRU 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

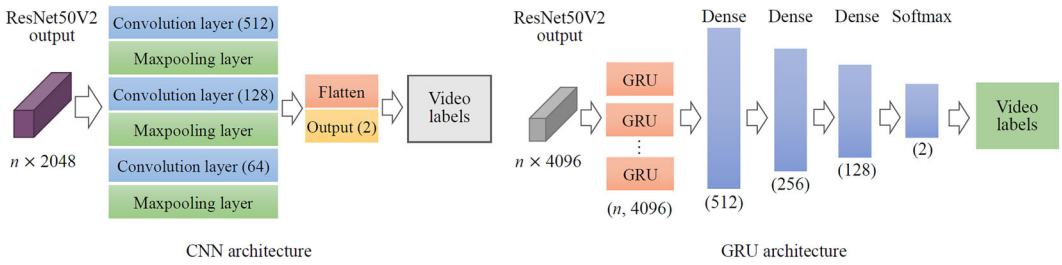

FIGURE 6 CNN and GRU architecture.

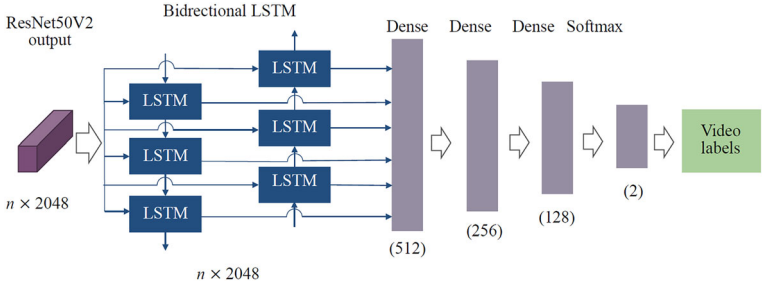

FIGURE 7 LSTM architecture.

results into two class labels: violence or non-violence. This process is illustrated in Figure 6. The hyperparameter settings for GRU 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.

5 | RESULTS AND DISCUSSION

In this study, acts of violence were classified on the following publicly available datasets: Movies, Hockey, and Crowd. We used a deep transfer learning approach based on ResNet50V2 to extract essential features from the data. In addition, 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. This study used pretrained weights obtained from training on the ImageNet dataset. We used ImageNet weights as pretrained weights because the images in ImageNet have a resolution of 224 × 224, which matches the CCTV image frame input. In addition, ImageNet has approximately 14 million images in 1000 various categories. The use of a model pretrained on ImageNet certainly improves learning outcomes on violence datasets and produces good recognition performance. We divided the training and test data using 10-fold cross-validation. The classification algorithms CNN, LSTM, and GRU were used. The parameters used for model evaluation were accuracy, recall, precision, and F1-score; we also considered model performance in terms of the time required for feature extraction, training, and testing for each dataset. The experimental results obtained on the Movies, Hockey, and Crowd datasets are listed in Table 2.

Based on Table 2, the highest accuracy on the Hockey dataset (1.000) is obtained when the ResNet50V2–GRU combination. In addition to obtaining the highest accuracy, the ResNet50V2–GRU combination produced the best precision, recall, and F1-score values of 1.000 on each metric. This shows that the model's ability to classify the two classes is better than the abilities of the other algorithm combinations. As on the Hockey dataset, the best accuracy on the Crowd dataset (1.000) is also obtained when the ResNet50V2–GRU combination is used. If further reviewed, the use of ResNet50V2 for feature extraction improved model performance, as evidenced by the increase in accuracy, recall, precision, and F1-score, when compared with classical and older feature extraction approaches. On the other hand, using deep transfer learning features yields significantly better results when compared with classical feature extraction. This may indicate that using level-3 Daubechies-8 wavelets and PCA is inappropriate because they can eliminate features that are important for classification. It can be found that the model built with the Crowd dataset using the ResNet50V2–GRU combination obtained the best performance because it produces 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 GRU and its combination with Daubechies-8 wavelets and PCA. This may have happened because the video in this dataset is a snapshot of a scene in a film in which the lighting and shooting angles are set. Hence, the video is clear and does not contain much noise. This differs from the Hockey and Crowd datasets, which were obtained from surveillance cameras.

Table 2 also presents the time required to perform feature extraction on a dataset and the classification time in seconds. Based on Table 2, it can be seen that feature extraction using ResNet50V2 takes longer, but for the training process, ResNet50V2 is faster than VGG-16 and VGG-19. Table 2 also presents the CPU time required to process one test video in seconds. The fastest time was obtained using VGG-19. For the Crowd dataset, ResNet50V2-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 0.1–0.6 s can be tolerated. In addition, one of the advantages of the GRU, as shown in Table 2, is that in terms of time, GRU is faster than LSTM because fewer parameters are used in the GRU. As a result, the GRU is more efficient in terms of memory and time. The results show that GRU can provide good performance on all the datasets in this study. The perfect scores show that the ResNet50V2 + GRU model can learn optimally on a violence dataset to recognize the patterns in each category very well. Figure 8 shows plots of 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 the Movies, Hockey, and Crowd datasets. 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 Figure 9, we display a scatter plot of the recognition results for each data instance in the Crowd dataset.

TABLE 2 Experimental results of feature extraction on the Movie, Hockey, and Crowd datasets.

Dataset	Classifier	Feature extraction	Extraction time	Training time	Testing time	Acc.	Rec.	Prec.	F1-score
Movies	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**
	GRU	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**
	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**
Hockey	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**
	GRU	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**
	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**
Crowd	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**
	GRU	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**
	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

TABLE 2 (Continued)

Dataset	Classifier	Feature extraction	Extraction time	Training time	Testing time	Acc.	Rec.	Prec.	F1-score
		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**

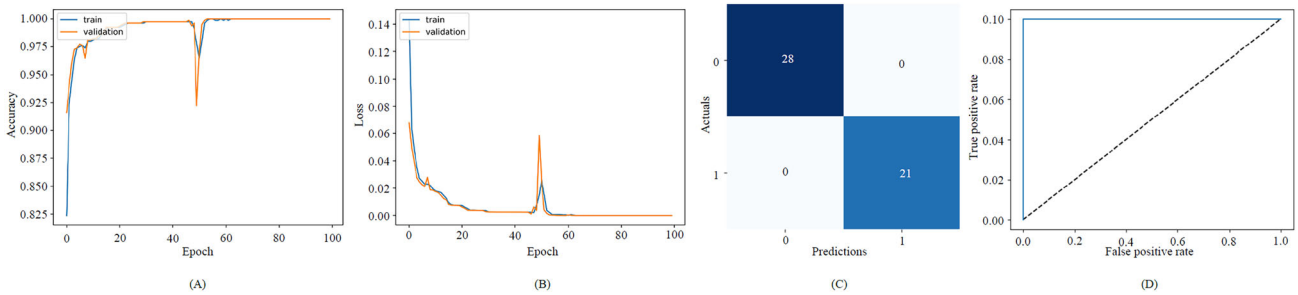

FIGURE 8 Plots of (A) training accuracy, (B) model loss, (C) the confusion matrix, and (D) the ROC for ResNet50V2.

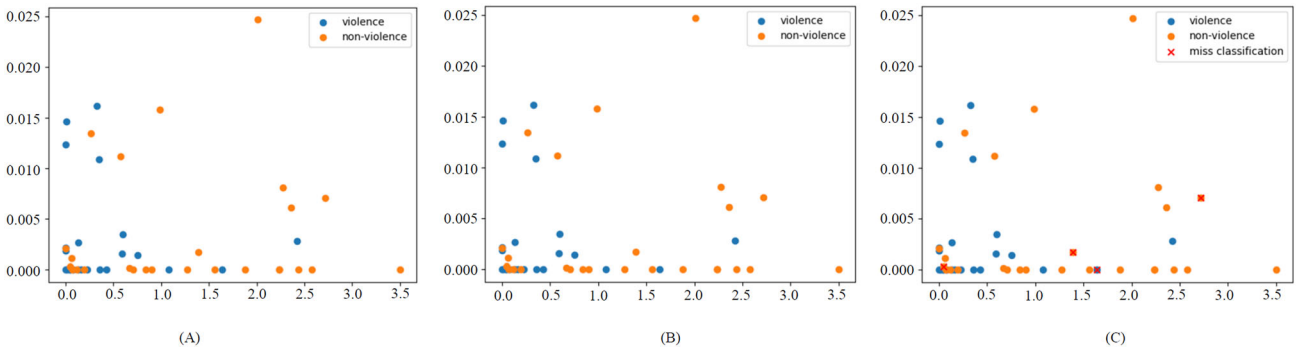

FIGURE 9 Scatter plot of (A) GRU + ResNet50V2, (B) LSTM + ResNet50V2, and (C) CNN + ResNet50V2.

Table 2 reveals that the best recognition results were obtained using the ResNet50V2 transfer learning model, and recognition comparisons using the GRU, LSTM, and CNN models were performed. In the graph, it can be seen that of the 49 test data, four were miss-classified when the CNN + ResNet50V2 model was used, whereas neither the GRU + ResNet50V2 model nor the LSTM + ResNet50V2 model output miss-classifications. In addition, to the graphs of the evaluation metric results, we also show the detection results in the video test data for each video by including the probability value of the results of recognizing violence and non-violence. The results of this detection are shown in Figure 10. In this figure, the recognition results show the prediction results of the GRU + ResNet50V2 model, which is the best of the compared models. This model was then tested on 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. In terms of the complexity and time consumption of the proposed model, it can be seen in Table 2 that each deep transfer learning model has a different extraction time. The longest feature extraction time was obtained using ResNet50V2 with an execution time of 0.231 s for each image, whereas the fastest feature extraction execution time was for the wavelet, with an execution time of 0.13 s. ResNet50V2 has the longest extraction time, where the transfer learning process is quite complex because it utilizes many residual networks, causing the learning process to take longer than that in other transfer learning models. The longest training process was that of the CNN with VGG-19 feature extraction (263 022 s). A comparison with other studies is presented in Table 3. In the Movies dataset, the proposed method outperforms the other methods with the highest scoring accuracy of

FIGURE 10 Detection results of GRU + ResNet50V2 on the (A) Crowd, (B) Movies, and (C) Hockey datasets.

TABLE 3 Comparison of the violence detection system proposed in this study with previous similar systems.

		Accuracy (%)		
Researcher	**Method**	**Movie dataset**	**Hockey dataset**	**Crowd dataset**
Zhang [28]	MoWLD + BoW	-	91.9	82.5
Zhang [28]	MoWLD + SparseCoding	-	93.7	86.3
Keceli [29]	ConFeat	96.5	94.4	80.9
Rabiee [30]	sHOT	-	-	82.9
Mabrouk [31]	DIMOLIF	-	88.6	85.8
Zhou [32]	LHOF + BoW	-	-	86.5
Mahmoodi [10]	HOMO	-	89.3	76.8
Febin [33]	BoW (MoBSIFT) + MF	98.9	90.3	-
Keceli [34]	AlexNet + 3D-CNN	98.7	92.9	88.0
Proposed system	**ResNet50V2 + GRU**	**100.0**	**100.0**	**100.0**

100%. This demonstrates that our proposed method performs well at detecting violence in videos.

6 | CONCLUSION

In this study, violence was detected in video data from the Movie, Hockey, and Crowd datasets. ResNet50V2 was used for feature extraction and classical (Wavelet and PCA), and other deep transfer learning methods (VGG-16 and VGG-19) were used as comparison methods. Furthermore, the CNN, LSTM, and GRU algorithms were used for classification. The best accuracy results on the Hockey dataset were obtained when using the ResNet50V2–GRU combination. Furthermore, on the Movies dataset, all combinations of algorithms provided excellent performance (1.000). Similar to the Hockey dataset, the best accuracy on the Crowd dataset was achieved using the ResNet50V2–GRU combination. Furthermore, ResNet50V2–GRU provides the best accuracy, recall, precision, and F1-score performance. The experimental results in this study show that GRU performs better in terms of time than LSTM. GRU also provides good performance on all the datasets used in this study. The ResNet50V2–GRU combination achieves the best accuracy and F1-score values on all datasets. In general, using ResNet50V2 for feature extraction improves the model performance 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. Furthermore, reducing testing time for real-time violence detection remains a challenge.

ACKNOWLEDGMENTS

This research was supported by a Prototype Output Grant (Approval No. B/65354/UN38.III.1/LK.04.00/2023, No. 205/E5/PG.02.00.PM/2023), funded by the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

ORCID

Elly Matul Imah https://orcid.org/0000-0003-1008-4837
Riskyana Dewi Intan Puspitasari https://orcid.org/0000-0002-6065-6090

REFERENCES

1. M. Asad, J. Yang, J. He, P. Shamsolmoali, and X. He, *Multi-frame feature-fusion-based model for violence detection*, Vis. Comput. **37** (2021), 1415–1431.
2. M. Ullah, M. M. Yamin, A. Mohammed, S. D. Khan, H. Ullah, and F. A. Cheikh, *Attention-based lstm network for action recognition in sports*, (Proc. IS&T Int'l. Symp. on Electronic Imaging: Intelligent Robotics and Industrial Applications using Computer Vision), 2021, pp. 302–1–302–6.
3. M. Nohara and H. Nishi, *Video object detection method using single-frame detection and motion vector tracking*, (IEEE 18th International Conference on Industrial Informatics, Warwick, UK), 2020, pp. 119–25.
4. N. Venkatesvara Rao, D. Venkatavara Prasad, and M. Sugumaran, *Real-time video object detection and classification using hybrid texture feature extraction*, Int. J. Comput. Appl. **43** (2021), no. 2, 119–126.
5. G. Heo, J. Jeon, and B. Son, *Crack automatic detection of cctv video of sewer inspection with low resolution*, KSCE J. Civ. Eng. **23** (2019), 1219–1227.
6. J. L. Salazar González, C. Zaccaro, J. A. Álvarez-García, L. M. Soria Morillo, and F. Sancho Caparrini, *Real-time gun detection in cctv: an open problem*, Neural Netw. **132** (2020), 297–308.
7. J. Lim, M.I. Al Jobayer, V.M. Baskaran, J.M. Lim, K. Wong, and J. See, *Gun detection in surveillance videos using deep neural networks*, (Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Lanzhou, China), 2019, pp. 1998–2002.
8. R. Debnath and M.K. Bhowmik, *Automatic visual gun detection carried by a moving person*, (IEEE 15th International Conference on Industrial and Information Systems, Rupnagar, India), 2020, DOI 10.1109/ICIIS51140.2020.9342681.
9. S. Das, A. Sarker, and T. Mahmud, *Violence detection from videos using hog features*, (4th International Conference on Electrical Information and Communication Technology, Khulna, Bangladesh), 2019, DOI 10.1109/EICT48899.2019.9068754.
10. J. Mahmoodi and A. Salajeghe, *A classification method based on optical flow for violence detection*, Expert Syst. Appl. **127** (2019), 121–127.
11. A. Jain and D. K. Vishwakarma, *Deep neuralnet for violence detection using motion features from dynamic images*, (Third International Conference on Smart Systems and Inventive Technology, Tirunelveli, India), 2020, pp. 826–31.
12. M. A. Soeleman, C. Supriyanto, and D. P. Prabowo, *An empirical study of cnn-lstm on class imbalance datasets for violence video detection*, (The 2021 International Conference on Computer, Control, Informatics and Its Applications), 2022, pp. 81–85.
13. I. E.M. Karisma and A. Wintarti, *Violence classification using support vector machine and deep transfer learning feature extraction*, (International Seminar on Intelligent Technology and Its Applications, Surabaya, Indonesia), 2021, pp. 337–42.
14. K. Karisma, E. M. Imah, I. K. Laksono, and A. Wintarti, *Detecting violent scenes in movies using gated recurrent units and discrete wavelet transform*, Regist: J. Ilm. Teknol. Sist. Inf. **8** (2022), no. 2, 94–103.
15. M. W. Fakhr, F. A. Maghraby, and M. Magdy, *Violence 4d: violence detection in surveillance using 4d convolutional neural networks*, IET Comput. Vis. **17** (2023), no. 3, 282–294.
16. T. Hassner, Y. Itcher, and O. Kliper-Gross, *Violent flows: Real-time detection of violent crowd behavior*, (3rd IEEE International Workshop on Socially Intelligent Surveillance and Monitoring at the IEEE Conf on Computer Vision and

Pattern Recognition, Providence, RI, USA), 2012, DOI 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*, The Cognitive Approach in Cloud Computing and Internet of Things Technologies for Surveillance Tracking Systems, Academic Press, 2020, DOI 10.1016/B978-0-12-816385-6.00011-8.
18. A. Onan and M. A. Toçoğlu, *A term weighted neural language model and stacked bidirectional lstm based framework for sarcasm identification*, IEEE Access **9** (2021), 7701–7722.
19. B. C. Mateus, M. Mendes, J. T. Farinha, R. Assis, and A. M. Cardoso, *Comparing lstm and gru models to predict the condition of a pulp paper press*, Energies **22** (2021), no. 21, 6958.
20. R. Rana, *Gated recurrent unit (GRU) for emotion classification from noisy speech*, arXiv preprint, 2016, DOI 10.48550/arXiv.1612.07778.
21. E.B. Nievas, O.D. Suarez, G.B. Garcia, and R. Sukthankar, *Hockey fight detection dataset*, 2011, DOI 10.1371/journal.pone.0120448.
22. S. Bahri, L. Awalushaumi, and M. Susanto, *The approximation of nonlinear function using daubechies and symlets wavelets*, (Proceedings of the International Conference on Mathematics and Islam), 2018, pp. 300–306.
23. S. Ren J. Sun, K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, arXiv preprint, 2015, DOI 10.48550/arXiv.1512.03385.
24. K. Simonyan, and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, arXiv preprint, 2014, DOI 10.48550/arXiv.1409.1556.
25. S. Ren, J. Sun, K. He, and X. Zhang, *Identity mappings in deep residual networks*, In *European conference on computer vision*, Springer, 2016, 630–645.
26. J. Q. Gan, E. J. J. Savran, and R. Kızıltepe, *A novel keyframe extraction method for video classification using deep neural networks*, Neural. Comput. Applic. **35** (2021), 24513–24524.
27. X. Xiang and L. Zhang, *Video event classification based on two-stage neural network*, Multimed. Tools Appl. **79** (2020), 21471–21486.
28. B. Yang, J. Yang, X. He, Z. Zheng, T. Zhang, and W. Jia, *Mowld: a robust motion image descriptor for violence detection*, Multimed. Tools Appl. **76** (2017), no. 1, 1419–1438.
29. A. S. Keçeli and A. Y. Kaya, *Violent activity detection with transfer learning method*, Electron. Lett. **53** (2017), no. 15, 1047–1048.
30. H. Mousavi, M. Nabi, M. Ravanbakhsh, and H. Rabiee, *Detection and localization of crowd behavior using a novel tracklet-based model*, Int. J. Mach. Learn. Cybern. **9** (2018), 1999–2010.
31. A. B. Mabrouk and E. Zagrouba, *Spatio-temporal feature using optical flow based distribution for violence detection*, Pattern. Recogn. Lett. **92** (2017), 62–67.
32. H. Luo, X. Hou, P. Zhou, and Q. Ding, *Violence detection in surveillance video using low-level features*, PLoS ONE **13** (2018), no. 10, DOI 10.1007/s12205-019-0980-7.
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. Anal. Appl. **23** (2020), no. 2, 611–623.
34. A. S. Keceli and A. Kaya, *Violent activity classification with transferred deep features and 3d-Cnn*, Signal Image Video Process. **17** (2022), no. 1, 139–146.

AUTHOR BIOGRAPHIES

Elly Matul Imah received her B.S. degree in mathematics from Sepuluh Nopember Institute of Technology, in 2004, and her M.Sc. and Ph.D. degrees in computer science from Universitas Indonesia in 2009 and 2014, respectively. She is a Lecturer and Researcher in the Data Science Department at Universitas Negeri Surabaya. She is also the Secretary of the Indonesian Association for Pattern Recognition and the Head of the Data Engineering Laboratory of the Data Science Department at Universitas Negeri Surabaya. Her research interests include machine learning, pattern recognition, biomedical engineering, computer vision, and neurocognitive science.

Riskyana Dewi Intan Puspitasari received her B.S. degree in mathematics from Universitas Negeri Surabaya, in 2016, and her M.Sc. degree in computer science from Universitas Indonesia in 2020. She is a Lecturer and Researcher in the Data Science Department at Universitas Negeri Surabaya. Her research interests include machine learning, deep learning, biomedical engineering, and signal processing.

How to cite this article: E. M. Imah and R. D. I. Puspitasari, *Violent crowd flow detection from surveillance cameras using deep transfer learning–gated recurrent unit*, ETRI Journal **46** (2024), 671–682, DOI 10.4218/etrij.2023-0222