IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 18, 2021, accepted October 27, 2021, date of publication November 30, 2021,
date of current version December 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3131315

Deep Learning for Automatic Violence Detection: Tests on the AIRTLab Dataset

PAOLO SERNANI[1], NICOLA FALCIONELLI[1], SELENE TOMASSINI[1], (Student Member, IEEE), PAOLO CONTARDO[1,2], AND ALDO FRANCO DRAGONI[1]

[1]Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche, 60131 Ancona, Italy
[2]Gabinetto Interregionale di Polizia Scientifica per le Marche e l'Abruzzo, 60129 Ancona, Italy

Corresponding author: Paolo Sernani (p.sernani@univpm.it)

ABSTRACT Following the growing availability of video surveillance cameras and the need for techniques to automatically identify events in video footages, there is an increasing interest towards automatic violence detection in videos. Deep learning-based architectures, such as 3D Convolutional Neural Networks, demonstrated their capability of extracting spatio-temporal features from videos, being effective in violence detection. However, friendly behaviours or fast moves such as hugs, small hits, claps, high fives, etc., can still cause false positives, interpreting a harmless action as violent. To this end, we present three deep learning-based models for violence detection and test them on the AIRTLab dataset, a novel dataset designed to check the robustness of algorithms against false positives. The objective is twofold: on one hand, we compute accuracy metrics on the three proposed models (two are based on transfer learning and one is trained from scratch), building a baseline of metrics for the AIRTLab dataset; on the other hand, we validate the capability of the proposed dataset of challenging the robustness to false positives. The results of the proposed models are in line with the scientific literature, in terms of accuracy, with transfer learning-based networks exhibiting better generalization capabilities than the trained from scratch network. Moreover, the tests highlighted that most of the classification errors concern the identification of non-violent clips, validating the design of the proposed dataset. Finally, to demonstrate the significance of the proposed models, the paper presents a comparison with the related literature, as well as with models based on well-established pre-trained 2D Convolutional Neural Networks (2D CNNs). Such comparison highlights that 3D models get better accuracy performance than time distributed 2D CNNs (merged with a recurrent module) in processing the spatio-temporal features of video clips. The source code of the experiments and the AIRTLab dataset are available in public repositories.

INDEX TERMS Convolutional long short-term memory, convolutional neural network, deep learning, support vector machine, violence detection.

I. INTRODUCTION

Public video surveillance systems are common all over the world, being capable of providing accurate and rich information in many security applications [1]. However, the need of watching hours of video footages undermines the chance to take decisions in a short time, which is essential in video surveillance for crime and violence prevention [2]. In this regard, several studies about automatic detection of violent scenes in videos have been presented, with the aim to unburden authorities from the need of watching hours of videos to identify events lasting few seconds. Whilst early research works used hand-crafted features and flow descriptors typical of traditional action recognition methods [3]–[5], recent works highlighted the accuracy of deep learning-based approaches in violence detection [6]–[8]. In fact, deep learning techniques have been proven effective in extracting spatio-temporal features from videos [9], [10], i.e. features that represent the motion information contained in a sequence of frames, in addition to the spatial information contained in a single frame.

Among the deep learning-based techniques for violence detection, in our previous research [11] we showed the effectiveness of the combination of 3D Convolutional Neural Networks (3D CNN) and Support Vector Machines (SVM) to detect both person-to-person fights and crowd violence in

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Sharif.

videos. Nevertheless, we highlighted that there are still false positives, detecting friendly behaviours or rapid moves such as hugs, small hits, claps, and high fives as violent. To further investigate in such direction, this paper presents a comparison of three different deep learning models on a novel dataset, the AIRTLab dataset [12], that we built to include, as non-violent samples, video clips that can cause false positives. Specifically, this paper adds the following contributions to the state of the art of automatic violence detection in videos:

- it describes a new dataset intended to train and benchmark techniques for automatic violence detection in videos. The dataset is specifically tailored to test the performance against false positives;
- it proposes two transfer learning-based models and one ''trained from scratch'' model, testing them on the presented dataset. The results serve as a baseline to benchmark the performance of violence detection techniques applied to the proposed dataset;
- it compares the performance of the proposed models with well-defined pre-trained 2D Convolutional Neural Networks (2D CNN) on the task of violence detection. Specifically, it tests the performance of VGG16 and VGG19 [13], ResNet50 version 2 [14], Xception [15], and NASNet Mobile [16]. Being 2D, these networks have been adapted to be applied frame-by-frame to videos and process spatio-temporal information, in order to be compared with the proposed models. As a side effect of such comparison, this paper provides a benchmark of well-established networks on the task of violence detection;
- it provides the implementation of all the described models and experiments, as the source code of the tests is publicly available in a GitHub repository,[1] to ensure the reproducibility of the experiments. Moreover, the AIRTLab dataset is also available in a public repository.[2]

In fact, the datasets traditionally used to compare violence detection techniques, such as the Hockey Fight Dataset [17], the Movie Fight Dataset [17], and the Crowd Violence Dataset [3], usually include few videos, recorded at a low resolution; in many cases such videos are registered in too specific environments (such as hockey arenas and football stadiums). Instead, the dataset proposed in this paper includes 350 Full HD clips, at 30 frame per seconds.

Moreover, following our previous research, C3D, an existing 3D CNN pre-trained to classify sport categories in videos [18], is used as a feature extractor in the transfer learning models. In the first model, the classification task is performed by a SVM classifier. In the second model, the classification is done by fully connected layers. Instead, the model which was trained from scratch is based on the Convolutional Long Short-Term Memory (ConvLSTM) architecture [19] to extract the spatio-temporal features of the videos; the classification task is performed by fully connected layers following the ConvLSTM. We tested these three models on the AIRTLab dataset, in order to compare their performance. Rather than proposing novel recognition architectures and models, we want to validate that the proposed dataset is capable of challenging the robustness to false positives of the violence detection techniques, testing architectures which already showed a good classification performance. In fact, we tested our models also on the Hockey Fight and Crowd Violence datasets, building a comparison over the existing literature. Finally, by comparing the proposed models against the performance got by end-to-end models based on pre-trained 2D CNNs, we also provide an extended baseline of results of well-established networks on the AIRTLab dataset, as well as on the Hockey Fight and Crowd Violence datasets.

The rest of this paper is organized as follows. Section II lists related research papers, highlighting similarities and differences with the presented research. Section III describes the used deep learning techniques, providing the necessary background, detailing the dataset structure, and explaining the architectures of the three proposed models. Section IV presents and discusses the experimental results and the main findings, describing the tests on the proposed dataset as well as the results got on other datasets. Finally, Section V draws the conclusions of this research.

II. RELATED WORKS

Concerning the violence detection techniques, in recent years deep learning models showed their potential on the listed datasets, achieving top level performance in terms of classification accuracy. Among these techniques, 3D CNNs and ConvLSTMs have been proven effective in learning the spatio-temporal information contained in videos [20]. In this regard, Table 1 lists the accuracy of some deep learning-based violence detection techniques on the Hockey Fight and Crowd Violence datasets. Among the datasets used to compare violence detection techniques, the Hockey Fight Dataset [17], the Movie Fight Dataset [17], and the Crowd Violence Dataset [3] had been widely adopted. The Hockey Fight Dataset includes 1000 clips equally divided into ''fight'' and ''no-fight''. Each clip has between 41 and 50 frames (as also reported in [21]), originally at a resolution of 720 x 576 pixels, even if the 320 x 240 pixels version as proposed in [22] is commonly used. The Movie Fight Dataset includes 200 clips extracted from movies, 100 labeled as ''fight'' and 100 as ''no-fight''. Similarly to the Hockey Fight Dataset, each clip is composed of 50 frames at 720 x 576 and 720 x 480 pixels. The Crowd Violence Dataset is composed of 246 clips downloaded from Youtube (123 violent, 123 non-violent), at a resolution of 320 x 240 clips and with an average length of 3.6 seconds. All these datasets includes low resolution videos; the Hockey Fight and Crowd Violence datasets include clips recorded in very specific environments (such as hockey arenas and football stadiums);

[1] Source code of the experiments: https://github.com/airtlab/violence-detection-tests-on-the-airtlab-dataset

[2] AIRTLab dataset: https://github.com/airtlab/A-Dataset-for-Automatic-Violence-Detection-in-Videos

TABLE 1. Accuracy of deep learning-based violence detection techniques on the Hockey Fight and Crowd Violence datasets. The last row reports the results of our previous work [11], based on the combination of the pre-trained C3D network with an SVM classifier.

Authors	Architecture	Hockey Fight	Crowd Violence
Ding et al. (2014) [26]	3D CNN	91.0%	-
Song et al. (2019) [21]	3D CNN	99.6%	94.3%
Li et al. (2019) [24]	3D CNN	98.3%	97.2%
Ullah et al. (2019) [8]	C3D + Fully connected layers	96.0%	98.0%
Sudhakaran and Lanz. (2017) [23]	2D CNN + ConvLSTM	97.1%	94.5%
Hanson et al (2018) [27]	2D CNN + ConvLSTM	98.1%	96.3%
Accattoli et al. (2020) [11]	C3D + SVM	98.5%	99.2%

the Movie Fight Dataset contains few frames (10000) and most of the recent studies achieved 100% accuracy on it (see, for example, [23] and [24]). Recently, Cheng *et al.* [25] proposed a dataset to overcome these issues, the RWF-2000, with 2000 clips of surveillance camera videos collected from youtube, at various resolutions. Similarly to the RWF-2000, the dataset that we propose overcomes the limitations of the traditional datasets, offering 350 clips of various length (mean 5.36 seconds), in Full HD (1920 x 1080 pixels) resolution. However, differently from the other datasets, the AIRTLab dataset includes in the non-violent clips actions such as hugging, giving high fives and clapping, exulting, and gesticulating which might result into false positives, due to fast movements and similarity with some violent behaviours. Therefore, the AIRTLab dataset is designed to test violence detection techniques robustness against false positives. In this regard, 350 clips might seem few compared to other tasks of computer vision and to the 1000 clips of the Hockey Fight. However, we must highlight that the average clip length (5.63 seconds) is higher than the Hockey Fight (around 2 seconds) and Crowd Violence (3.6 seconds), including more frames than these two datasets. Specifically, Ding *et al.* [26] proposed to use a 9-layer 3D CNN for violence detection: processing 40 frames at a time, with a resolution of 60 x 90 pixels, three 3D convolutional layers alternated with two pooling layers, two fully connected layers and a softmax layer for classification achieved a 91% accuracy on the Hockey Fight Dataset. More recently, Song *et al.* [21] achieved 99.6% accuracy on the Hockey Fight Dataset, and 94.3% on the Crowd Violence, training from scratch a 3D CNN reproducing the C3D architecture and improving the sampling method. Similarly, Li *et al.* [24], with a 10-layer 3D CNN alternating dense and transitional layers after a convolutional layer, achieved 98.3% accuracy on the Hockey Fight, and 97.2% on the Crowd Violence. However, transfer learning approaches based on 3D CNNs achieved even better results by using models pre-trained with different classification tasks. Ullah *et al.* [8] implemented a pre-trained model based on C3D until the second fully connected layer (''fc7'') with a two-classes softmax layer to perform the classification, achieving good performance in both the Hockey Fight (96% accuracy) and Crowd Violence (98%) datasets. Similarly, in our previous work [11], we used the pre-trained C3D until the first fully connected layer (''fc6'') as a feature extractor, and a SVM classifier for the violence detection tasks, achieving excellent performance on both the Hockey Fight (98.5% accuracy) and Crow Violence (99.2%) datasets. Also the use of the ConvLSTM architecture in violence detection models achieved promising results. For example, Sudhakaran and Lanz [23] proposed to aggregate the spatial information extracted from the frames by 2D CNNs with a ConvLSTM, to extract the temporal information. With such architecture, they achieved 97.1% accuracy on the Hockey Fight dataset, and 94.5% on the Crowd Violence dataset. A similar approach was proposed by Hanson *et al.* [27] who combined the VGG13 CNN [13] with a ConvLSTM layer to achieve 98.1% accuracy on the Hockey Fight dataset, and 96.3% on the Crowd Violence dataset.

III. MATERIALS AND METHODS

As pointed out in the Introduction and the Related Works sections, deep learning-based architectures and, specifically, 3D CNNs and ConvLSTMs, are capable of modeling the spatio-temporal features of videos, and demonstrated their accuracy on violence detection. For this reason, we based the classifiers proposed in this paper on such neural network architectures, comparing:

- end-to-end networks, i.e. a unique model to execute the classification, with a model composed of a 3D CNN and an SVM;
- the training of a model from scratch with two networks based on a transfer learning approach, i.e. the use of models already trained on a large dataset to execute a different classification task.

Figure 1 shows the workflow followed for the study presented in this paper. We propose three deep learning models (two are based on transfer learning and one is trained from scratch) to perform violence detection in videos. We introduce a dataset, the AIRTLab dataset, and present the performance of the proposed models on such dataset. To build a comparison over the existing literature, we test their performance on the Hockey Fight and Crowd Violence datasets, traditionally used in literature to benchmark violence detection techniques. Finally, to prove the significance of the proposed classifiers, we compare our models to the performance got by well-established pre-trained 2D CNNs such as VGG16, VGG19, and ResNet50, adapted to be applied to videos (which are 3D, being composed of multiple frames).

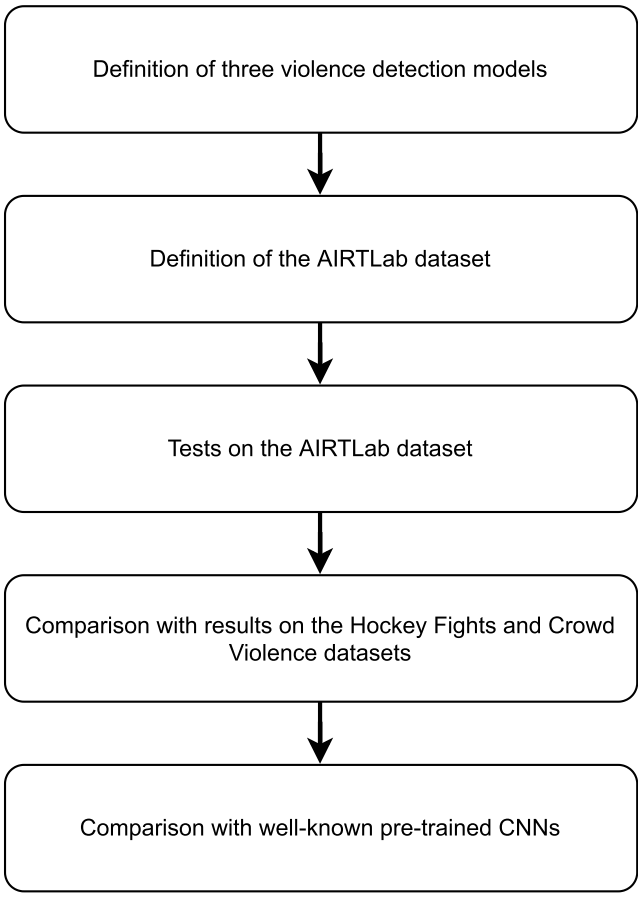


FIGURE 1. The workflow of the study proposed in this paper.

To this end, we provide some background notions about the 3D CNN and the ConvLSTM architectures (III-A), we present the AIRTLab dataset, used to test the proposed classifiers (III-B), and we describe the classifier architectures (III-C).

A. BACKGROUND: 3D CNN AND ConvLSTM

As highlighted in the seminal work of LeCun and Bengio [28], in a 2D CNN each unit of a layer receives inputs from a set of units located in a small neighborhood (the local receptive field) in the previous layer, by convoluting with a set of kernels composed of shared weights. Ji *et al.* [29] extended this concept by proposing to use 3D CNNs. The 3D convolution is obtained using a 3D kernel on the cube formed by stacking more adjacent frames together. In this way, the resulting feature map represents the temporal information available in sample data, in addition to the spatial information usually modeled by a 2D CNN.

In this work, we use an existing 3D CNN, C3D [18], which is trained on the Sports-1M dataset [30] to recognize sport categories in videos. Since it has been proven useful to extract spatio-temporal features from videos, we use C3D as a feature extractor, using the weights until the first fully connected layer (''fc6''), in a transfer learning fashion.

Whilst C3D is used in two of the three proposed models, the last model is based on the ConvLSTM that has also been proven useful to represent spatio-temporal features. Specifically, we use the formulation of Shi *et al.* [19], who extended the LSTM architecture [31] by adding convolutional structures to state transitions. A LSTM hidden unit is composed by a self-recurrent cell, called memory cell, whose input/output is regulated by three multiplicative gates, i.e. the input gate, the output gate, and the forget gate [32].

As Shi *et al.* pointed out, the LSTM architecture is adequate to extract temporal features, but includes too much redundancy for spatial features. In this regard, they proposed to add convolutional structures in the transitions between the input gate and the memory cell, and in the self-recurrency of the memory cell, regulated by the forget gate.

B. THE AIRTLab DATASET

To evaluate the three proposed deep learning models, we developed a dataset, called the AIRTLab dataset, to specifically test the robustness of violence detection techniques against false positives in non-violent clips with rapid moves (such as hugs, claps, high-fives, etc.). The dataset is publicly available as a GitHub repository.

The dataset is composed of 350 clips which are MP4 video files (H.264 codec) of a mean length of 5.63 seconds, with the shortest video lasting 2 seconds and the longest 14 seconds. For all the clips, the resolution is 1920 $\times$ 1080 pixels and the frame rate 30 fps. The dataset is split into two main directories, ''non-violent'' and ''violent'', labeling the included clips as showing non-violent behaviours and violent behaviours respectively. The directories are split into two subdirectories, ''cam1'' and ''cam2'':

- ''non-violent/cam1'' includes 60 clips representing non-violent behaviours;
- ''non-violent/cam2'' includes 60 clips with the same non-violent behaviours in ''non-violent/cam1'', but recorded from a different point of view;
- ''violent/cam1'' includes 115 clips representing violent behaviours;
- ''violent/cam2'' includes 115 clips with the same violent behaviours in ''violent/cam1'', but recorded from a different point of view.

All the clips were recorded in the same room, with natural lighting conditions, placing two cameras into two different spots (the top left corner in front of the room door, and the top right corner on the door side).

The clips were performed by a group of non-professional actors, varying from 2 to 4 per clip. For the violent clips, the actors were asked to simulate actions frequent in brawls, such as kicks, punches, slapping, clubbing (beating with a cane), stabbing, and gun shots. For the non-violent clips, the actors were asked to simulate actions which can result in false positives by violence detection techniques due to the speed of movements or the similarity with violent actions. Specifically, the non-violent clips include actions such as hugging, giving high fives and clapping, exulting,

TABLE 2. Layers of C3D used as a feature extractor in two of the proposed models. We used C3D until the first fully connected (i.e. dense) layer, called by its autors "fc6" [18].

Layer	Architecture	Output Shape	Params #
Conv3D	64 filters, 3x3x3 (stride 1), ReLu	(16, 112, 112, 64)	5248
MaxPooling3D	1x2x2	(16, 56, 56, 64)	0
Conv3D	128 filters, 3x3x3 (stride 1), ReLu	(16, 56, 56, 128)	221312
MaxPooling3D	2x2x2	(8, 28, 28, 128)	0
Conv3D	256 filters, 3x3x3 (stride 1), ReLu	(8, 28, 28, 256)	884992
Conv3D	256 filters, 3x3x3 (stride 1), ReLu	(8, 28, 28, 256)	1769728
MaxPooling3D	2x2x2	(4, 14, 14, 256)	0
Conv3D	512 filters, 3x3x3 (stride 1), ReLu	(4, 14, 14, 512)	3539456
Conv3D	512 filters, 3x3x3 (stride 1), ReLu	(4, 14, 14, 512)	7078400
MaxPooling3D	2x2x2	(2, 7, 7, 512)	0
Conv3D	512 filters, 3x3x3 (stride 1), ReLu	(2, 7, 7, 512)	7078400
Conv3D	512 filters, 3x3x3 (stride 1), ReLu	(2, 7, 7, 512)	7078400
ZeroPadding3D	(0, 0), (0, 1), (0, 1)	(2, 8, 8, 512)	0
MaxPooling3D	2x2x2	(1, 4, 4, 512)	0
Flatten	-	(8192)	0
Dense	4096 units, ReLu	(4096)	33558528

and gesticulating. All the actions in both violent and non-violent clips have been manually annotated. The complete dataset specification is available in a dedicated open-access data paper [12].

In terms of average clip length and total number of frames, the proposed dataset is bigger than the datasets generally used for the comparison of violence detection techniques, such as the Hockey Fight and Crowd Violence datasets. However, it seems relative small when compared with other datasets used in computer vision and video classification, such as the Sports-1M. Whilst scraping more videos from the internet would be possible, requiring bundersome manual checks and annotations, the proposed dataset is tailored to specifically test the robustness against false positives. Therefore, to achieve such objective, the dataset needs to be designed on purpose.

C. PROPOSED MODELS

In this paper we propose three different deep learning-based models to classify violence detection in videos and we test their accuracy on the AIRTLab datasets. Specifically:

1) the first model consists of C3D, as a feature extractor, and a linear SVM to classify clips into violent and non-violent;
2) the second model also uses C3D as a feature extractor, but the classification is done by two extra fully connected layers, building an end-to-end model;
3) the third model is trained from scratch and is based on the ConvLSTM architecture.

In two out of the three proposed models, the C3D network trained on the Sports-1M dataset is used as a feature extractor, in a transfer learning fashion. In fact, transfer learning can achieve better generalization than a dedicated training from scratch and prevent overfitting [33], [34]. In the original definition of Tran *et al.* [18], C3D uses $3 \times 3 \times 3$ kernels (with stride equal to 1) in a total of eight convolution layers alternated with five pooling layers, followed by two fully connected layers and a softmax output layer to compute the probability distribution over the sport categories. All the neurons in the convolution layers use the rectified linear activation function (ReLu). Table 2 lists the layers of C3D used in this work: we used all the original convolution and pooling layers and we take the output of the first fully connected layer (called "fc6" by the C3D authors) as a feature descriptor of the original input, removing the second fully connected layer and the final softmax layer. C3D takes as input sequences of 16 frames at a resolution of 112×112 pixels. Hence, we kept this input format for all the violence detection models proposed in this paper.

Figure 2 depicts the schematic of the proposed models. The clips given as input to the three models are divided into 16-frames chunks and resized at the resolution of 112×112 pixels, to be compliant with the C3D input.

In the first of the three proposed models (Figure 2a), the 4096 feature descriptor given as output by the first fully connected layer of C3D is fed into an SVM, with linear kernel and $C = 1$, in order to classify the 16 frames sequence as violent or not. In fact, in our previous work [11], we already demonstrated the capability of this model on the Hockey Fight and Crow Violence datasets, obtaining a 98.51% and a 99.29% accuracy respectively.

Table 3 shows the architecture of the second proposed model. Differently from the previous model, we built an end-to-end architecture, extending the portion of C3D used as a feature extractor with additional layers (Figure 2b). Specifically, we added a dropout layer, with a rate of 0.5, to prevent overfitting [35]. Subsequently we added a fully connected layer with 512 neurons using the rectified linear activation function. After another 0.5 dropout, the final layer composed by a neuron with the sigmoid activation performs the actual classification of the 16-frames clips into violent or not. Also in this model, the used C3D layers are trained on the Sports-1M dataset. Instead, the added layers were trained from scratch on the available training data, as explained in Section IV.

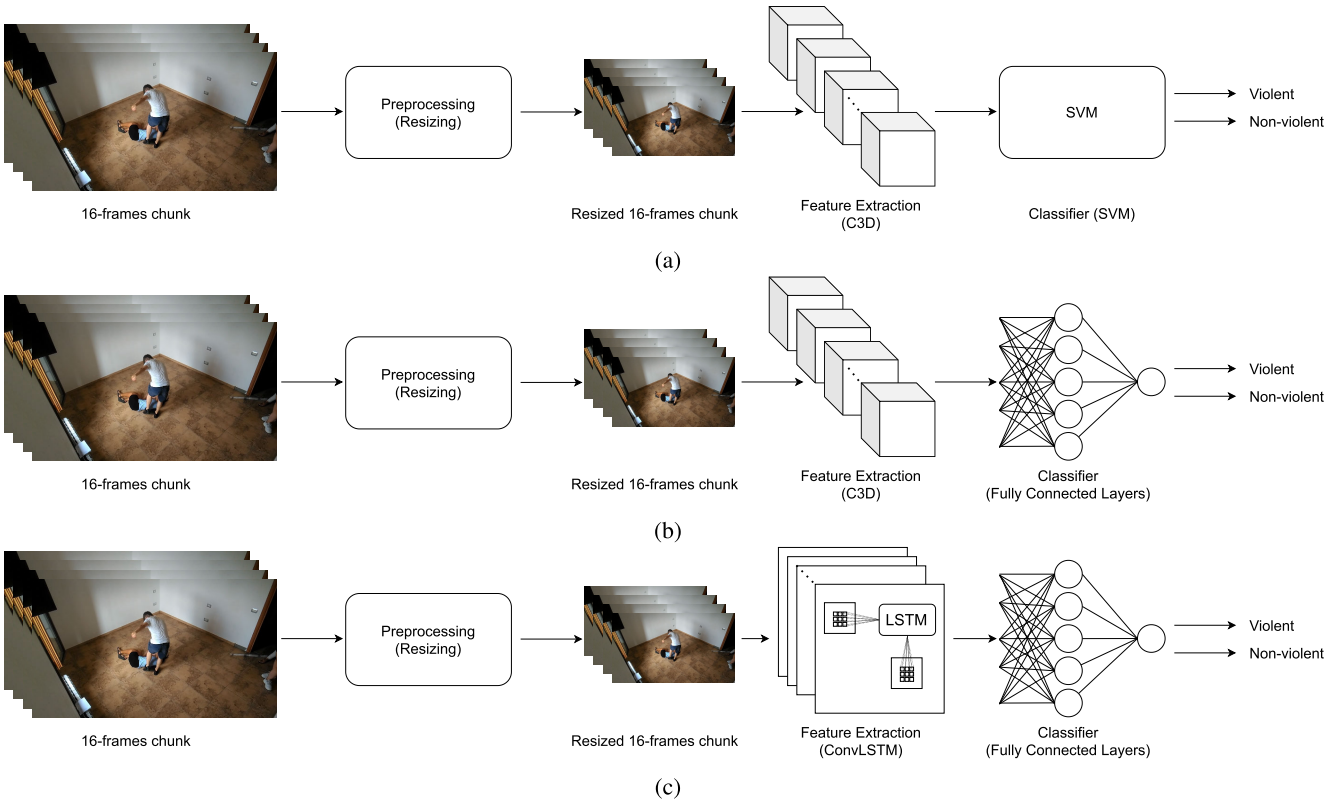


FIGURE 2. The schematic of the three models proposed in this paper. All the models process sequences composed of 16 frames (16-frames chunks) resized to 112 x 112 pixels. The first proposed model (a) uses the pre-trained C3D network as a feature extractor and an SVM classifier to label the chunks as violent or not. The second model (b) also uses C3D as a feature extractor, and the classifier is made of fully connected layers. The third model (c) uses a ConvLSTM layer trained from scratch, with fully connected layers for the final classification.

TABLE 3. The second proposed model. It is an end-to-end model which adds two fully connected layers to C3D (until "fc6"). C3D is not trained again, therefore the total number of trained parameters is 2,098,177 which are the weights of the final fully connected layers.

Layer	Architecture	Output Shape	Params #
C3D until "fc6"	(see Table 2)	(4096)	61214464
Dropout	0.5 rate	(4096)	0
Dense	512 units, ReLu	(512)	2097664
Dropout	0.5 rate	(512)	0
Dense	1 unit, Sigmoid	(1)	513

TABLE 4. The third proposed model. It is an end-to-end model based on the ConvLSTM architecture. It is trained from scratch and the total number of trained parameters is 198,401,537.

Layer	Architecture	Output Shape	Params #
ConvLSTM2D	64 filters, 3x3	(110, 110, 64)	154624
Dropout	0.5 rate	(110, 110, 64)	0
Flatten	-	(774400)	0
Dense	256 units, ReLu	(256)	198246656
Dropout	0.5 rate	(256)	0
Dense	1 unit, Sigmoid	(1)	257

The third proposed model (Figure 2c) is based on the ConvLSTM architecture, and it is trained end-to-end from scratch. The layers are listed in Table 4. The first layer is a ConvLSTM composed of 64 3 $\times$ 3 filters, with a total of 154,624 trainable parameters. After a 0.5 dropout to prevent overfitting, we flatten the ConvLSTM output and add a fully connected layer with 256 neurons, using the rectified linear activation function. Finally, after another 0.5 dropout, the final classification into violent or not is computed by a neuron with the sigmoid activation function. To allow a comparison with the models based on C3D, the input of the ConvLSTM-based network is also composed of sequences of 16 video frames at a resolution of 112 $\times$ 112 pixels.

To prove the significance of the proposed models, we carry out a comparison with the performance of well-established 2D CNNs, namely VGG16, VGG19, ResNet50V2, Xception, and NASNet Mobile, pre-trained on the ImageNet database [36]. Figure 3 describes the schematic of the models based on such 2D CNNs. In order to be applied to videos and process the related spatio-temporal information, the 2D CNNs are time distributed over the 16 frames composing an input chunk and combined with a recurrent layer, whereas two fully connected layers implement the final classification. ConvLSTM and Bidirectional-LSTM (Bi-LSTM) were both tested as the recurrent layer. Specifically, Table 5 includes the layers composing the models with the pre-trained 2D CNNs and the ConvLSTM layer as the recurrent module. The ConvLSTM layer was composed of 64 3 $\times$ 3 filters, followed by a fully connected layer with 256

TABLE 5. The model based on pre-trained 2D CNNs and ConvLSTM. The ConvLSTM and two fully connected layers were added to well-defined 2D CNNs (VGG16, VGG19, ResNet50V2, Xception, NASNet Mobile), pre-trained on ImageNet. The 2D CNNs were time distributed in order to be applied to a 3D input, i.e. the videos of the datasets. Note that the number of parameters of the ConvLSTM layer depends on the previous 2D CNN architecture.

Layer	Architecture	Output Shape	Params #
Time Distr. 2D CNN	-	-	-
ConvLSTM2D	64 filters, 3x3	(5, 5, 64)	-
Flatten	-	(1600)	0
Dense	256 units, ReLu	(256)	409856
Dropout	0.5 rate	(256)	0
Dense	1 unit, Sigmoid	(1)	257

TABLE 6. The model based on pre-trained 2D CNNs and Bi-LSTM. The Bi-LSTM and two fully connected layers were added to well-defined 2D CNNs (VGG16, VGG19, ResNet50V2, Xception, NASNet Mobile), pre-trained on ImageNet. The 2D CNNs were time distributed in order to be applied to a 3D input, i.e. the videos of the datasets. Note that the output shape of the time distributed flatten layer and the number of parameters of the Bi-LSTM depend on the previous 2D CNN architecture.

Layer	Architecture	Output Shape	Params #
Time Distr. 2D CNN	-	-	-
Time Distr. Flatten	-	-	0
Bi-LSTM	128 units	(256)	-
Dropout	0.5 rate	(256)	0
Dense	128 units, ReLu	(128)	32896
Dropout	0.5 rate	(128)	0
Dense	1 units, Sigmoid	(1)	129

ReLu neurons, a 0.5 dropout and a fully connected neuron with sigmoid activation to perform the final classification. Table 6 describes the layers composing the models with the pre-trained 2D CNNs and the Bi-LSTM layer as the recurrent module. The Bi-LSTM was composed of 128 hidden units, followed by a 0.5 dropout, a fully connected layer with 128 ReLu neurons, another 0.5 dropout and a fully connected sigmoid neuron for the final classification.

With the models based on the pre-trained 2D CNNs, the input frames were resized to 224 × 224 pixels instead of 112 × 112. In fact, most of the tested 2D CNNs uses 224 × 224 as the default input dimension; moreover, an input size of 112 × 112 led to a significantly lower accuracy with the pre-trained 2D CNNs.

IV. EXPERIMENTAL EVALUATION

We evaluated the proposed deep learning models by collecting the classification results over the AIRTLab dataset, in addition to the tests on the Hockey Fight and Crowd Violence datasets. The objective is twofold: on one hand, we want to compare the accuracy of our models in identifying violent scenes; on the other hand, we want to create a benchmark with baseline metrics on the proposed dataset, and validate its design intended to check the technique robustness against false positives. To this end, in the following subsections we present the experimental setup (IV-A) and the results of the evaluation (IV-B). Of course, the obtained results present some limitations, as explained in Subsection IV-C.

TABLE 7. Number of training epochs in each split (S1-S5) of the AIRTLab dataset for the two end-to-end model i.e. C3D with two fully connected layers (C3D + FC) and the ConvLSTM-based architecture.

	S1	S2	S3	S4	S5	Mean
C3D + FC	19	26	21	30	21	23.40 ± 4.03
ConvLSTM	10	8	6	15	8	9.40 ± 3.07

A. EXPERIMENTAL SETUP AND EVALUATION METRICS

We tested the three proposed models on the AIRTLab dataset, by applying a stratified shuffle split cross-validation scheme. To this end, we repeated a randomized 80-20 split 5 times, using the 80% of the data as the training set, and the 20% as the test set, preserving the percentage of samples from each class, in each split. The data splits were the same for all the tested models, to implement a fair comparison. Given that the inputs for the models are sequences composed of 16 frames and the clips in the dataset include a total of 3537 of such sequences, 2829 samples (i.e. 16 frames chunks) were used for training, and 708 for testing, in each split. The 12.5% of the training data, i.e. the 10% of the entire dataset, was used as validation data for the training of the two end-to-end neural networks based on C3D and ConvLSTM. In addition, to compare the proposed models with the literature on violence detection presented in Section II, we used the same 80-20 stratified shuffle split cross-validation to test on the Hockey Fight and Crowd Violence datasets. Finally, we compared the results of the proposed models on the AIRTLab, Hockey Fight and Crowd Violence datasets, with those obtained by the models based on pre-trained 2D CNNs. For the tests, we used the same 80-20 stratified shuffle split cross-validation used to measure the performance of the proposed models.

For the two end-to-end models, we used the Adam optimizer to minimize the Binary Cross-Entropy loss function during the training of the neural networks. The number of training epochs varied for each split, as we early stopped the training after 5 epochs without an improvement on the minimum validation loss, restoring the weights corresponding to the best validation loss. Table 7 shows the number of training epochs in each split, for each neural network, on the AIRTLab dataset. The mean number of training epochs was 23.4 (± 4.03) for the model based on C3D and the fully connected layers, and 9.4 (± 3.07) for the ConvLSTM-based networks. The batch size was 32 samples for the C3D-based model, and 8 for the ConvLSTM-based model.

As highlighted in the Introduction section, the Jupyter notebooks with the described experiments are available in a GitHub public repository, in order to guarantee the reproducibility of the tests. The tests ran on Google Colab with the GPU runtime, using Keras 2.4.3, TensorFlow 2.4.1, and scikit-learn 0.22.2.post1.

Labeling as 0 (negative) the 16-frames chunks of the non-violent clips and as 1 (positive) the chunks of the violent clips, we computed the following metrics over the test set in each split of the stratified shuffle split cross validation scheme:

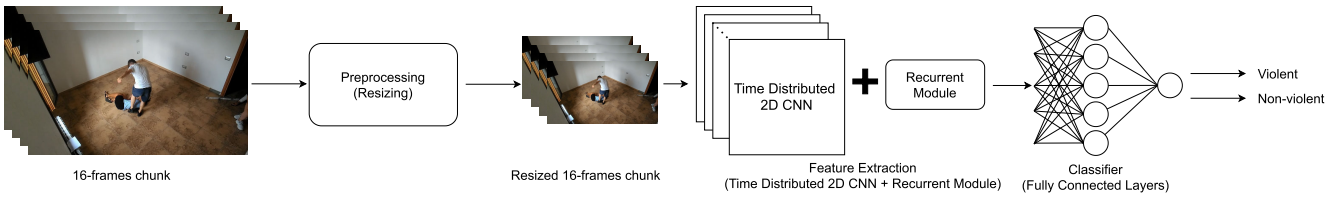


FIGURE 3. The schematic representation of the 2D CNN-based models, developed to compare the proposed models against the performance of well-established pre-trained 2D CNNs, such as VGG16, VGG19, and ResNet50. To apply the 2D CNNs to videos, they were time-distributed on the 16-frames chunks used as input and combined to recurrent layers (ConvLSTM and Bi-LSTM).

- sensitivity (True Positive Rate – TPR), i.e. the portion of positives that are correctly identified (over all the available positives);
- specificity (True Negative Rate – TNR), i.e. the portion of negatives that are correctly identified (over all the available negatives);
- accuracy, i.e. the portion of samples that are correctly identified (over all the available samples);
- F_1 score, i.e. the armonic mean of precision (the ratio between the positives correctly identified and all the identified positives) and sensitivity.

These metrics can be formulated in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) according to the following equations:

$$sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$specificity = \frac{TN}{TN + FP} \tag{2}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$F_1\ score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4}$$

Moreover, in each split, we computed the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC), showing the TPR against the False Positive Rate (FPR = 1 - TNR) when the classification threshold varies, to understand the diagnostic capability of each model. Finally, for each end-to-end model (i.e. all the models but the C3D + SVM model), we also report the value of the Binary Cross-Entropy loss function computed on the test set.

B. RESULTS AND DISCUSSION

We discuss in this subsection the metrics got by the three proposed models on the AIRTLab dataset as well as on the Hockey Fight and Crowd Violence datasets (to compare against the existing literature). Finally, in the last part of this Subsection, we present the results got by the pre-trained 2D CNNs on the datasets, in order to highlight the significance of the three proposed models.

1) TESTS ON THE AIRTLab DATASET

For each of the proposed models we report the results obtained on each split of the AIRTLab dataset in Tables 8, 9, and 10, in addition to the average computed for all the metrics (Table 11). The results on each split allow to understand to which extent the model is stable and independent from a particular split of the data.

TABLE 8. The results of the model composed of C3D and the SVM, computed for each split of the stratified shuffle-split cross validation scheme, on the AIRTLab dataset.

	Split 1	Split 2	Split 3	Split 4	Split 5
Sensitivity	**97.90%**	97.06%	**97.90%**	95.80%	96.64%
Specificity	93.53%	92.24%	**96.12%**	95.69%	93.10%
Accuracy	96.47%	95.48%	**97.32%**	95.76%	95.48%
F_1 score	97.39%	96.65%	**98.00%**	96.82%	96.64%
AUC	99.44%	98.89%	**99.46%**	99.45%	99.15%

TABLE 9. The results of the model composed of C3D and the fully connected layers for classification, computed for each split of the stratified shuffle-split cross validation scheme, on the AIRTLab dataset.

	Split 1	Split 2	Split 3	Split 4	Split 5
Loss	0.1471	**0.0996**	0.1135	0.1358	0.1113
Sensitivity	97.90%	98.32%	97.27%	**98.74%**	96.85%
Specificity	89.66%	92.24%	92.24%	87.93%	**93.53%**
Accuracy	95.20%	**96.33%**	95.62%	95.20%	95.76%
F_1 score	96.48%	**97.30%**	96.76%	96.51%	96.85%
AUC	98.32%	**99.21%**	99.05%	98.98%	99.08%

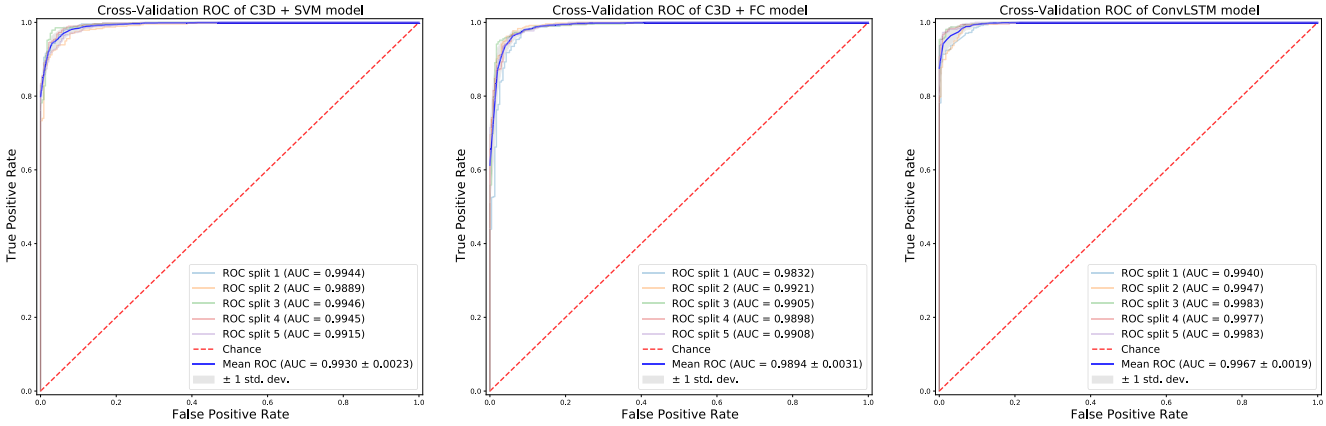
TABLE 10. The results of the model based on the ConvLSTM architecture, computed for each split of the stratified shuffle-split cross validation scheme, on the AIRTLab dataset.

	Split 1	Split 2	Split 3	Split 4	Split 5
Loss	0.1041	0.1004	**0.0511**	0.0759	0.0576
Sensitivity	97.48%	**99.58%**	98.53%	97.90%	97.90%
Specificity	91.38%	90.09%	97.41%	**97.84%**	97.41%
Accuracy	95.48%	96.47%	**98.16%**	97.88%	97.74%
F_1 score	96.67%	97.43%	**98.63%**	98.42%	98.31%
AUC	99.40%	99.47%	**99.83%**	99.77%	**99.83%**

Table 8 shows the metrics computed on the AIRTLab dataset for the model composed of C3D and the SVM classifier, in each of the splits of the stratified shuffle split cross-validation scheme. The specificity is lower than the sensitivity in each split, ranging from 92.24% in the second split to 96.12% in the third split. These results confirm that most of the errors are in the non-violent class, with the classifier giving some false positives in output. For example, in the second split, there are 18 false positives, with 214 out of 232 non-violent 16-frames chunks correctly classified as non-violent. Instead, in the same split, 462 over 476 violent

TABLE 11. The mean values of the metrics computed on the AIRTLab dataset over the five splits of the stratified shuffle split, for each of the proposed models.

	Loss	Sensitivity	Specificity	Accuracy	F_1 score	AUC
C3D + SVM	-	$97.06 \pm 0.80\%$	$94.14 \pm 1.51\%$	$96.10 \pm 0.71\%$	$97.10 \pm 0.53\%$	$99.30 \pm 0.23\%$
C3D + FC	0.1215 ± 0.0174	$97.82 \pm 0.69\%$	$91.12 \pm 2.03\%$	$95.62 \pm 0.42\%$	$96.78 \pm 0.30\%$	$98.94 \pm 0.31\%$
ConvLSTM	$\mathbf{0.0779 \pm 0.0215}$	$\mathbf{98.28 \pm 0.73\%}$	$\mathbf{94.83 \pm 3.37\%}$	$\mathbf{97.15 \pm 1.02\%}$	$\mathbf{97.89 \pm 0.74\%}$	$\mathbf{99.67 \pm 0.19\%}$



(a) (b) (c)

FIGURE 4. ROC curve and AUC for the C3D + SVM (a), the C3D + FC (b) and the ConvLSTM (c) models, on the AIRTLab dataset.

chunks are correctly labeled as violent, with only 14 false negatives. Of course, these results might be partially affected by the fact that the two classes are unbalanced in the AIRTLab dataset. The accuracy is in line with our previous work, getting a 97.32% in the best split.

Table 9 lists the results on the AIRTLab dataset got by the model composed of C3D and two fully connected layers. The trend in the metrics is similar to the model which uses C3D and the SVM classifier. However, with C3D and the fully connected layers, the difference between the sensitivity and the specificity is higher than the previous model. For example, in the fourth split, which has the highest difference (98.74% sensitivity, 87.93% specificity), 204 non-violent chunks are correctly classified, while there are 28 false positives. Instead, 470 violent chunks out of 476 are correctly classified. A significant difference between sensitivity and specificity is also visible in the other splits of the cross-validation scheme. In fact, the AUC of this model is slightly lower than the model based on C3D and the SVM classifier. Moreover, split 1 and split 4 present the highest number of false positives and the lowest accuracy as confirmed by the loss value, greater than the other splits.

Table 10 shows the results achieved by the ConvLSTM-based model on the AIRTLab dataset. The difference between sensitivity and specificity depends much more on the data split than the previous two models, showing that the ConvLSTM might have too many parameters given the amount of training data. In fact, in the first two splits, this difference is significant, being around 6% and 9%. Instead, in the third, fourth and fifth splits, sensitivity and specificity are much more closer, and the model seems robust to false positives and false negatives as well. The loss values on the test set are slightly lower than those obtained by the model composed of C3D and the fully connected layers, highlighting that the ConvLSTM tends to adapt to the dataset (in fact, the model is trained from scratch).

The results on the AIRTLab dataset are summarized in Table 11, which includes a comparison of the three models, showing the mean and the population standard deviation of the sensitivity, specificity, accuracy, F_1 score, and AUC, in addition to the loss for the two end-to-end models. In terms of accuracy, F_1 score, and AUC the model based on the ConvLSTM is slightly better than the others, with 97.15%, 97.89%, and 99.67% respectively. However, it is worth noting that this is the only model trained from scratch on the AIRTLab dataset and, therefore, it might overfit on the dataset, even if the validation split and the early stopping should limit overfitting. This is confirmed also by the lowest loss on the test set. The other two models exhibit a slightly lower accuracy, but, being based on a transfer learning methodology, their results can be interpreted as more general. The model composed of C3D and SVM is the one with the lowest difference between the sensitivity (97.06%) and specificity (94.14%), showing stable results both with violent and non-violent videos. The model composed of C3D and the fully connected layers exhibits the lowest metrics (only the sensitivity is slightly better than the SVM classifier). Nevertheless, the three models show similar diagnostic capability on identifying the violent videos, with very similar ROC curves and AUC, as highlighted in Fig. 4.

TABLE 12. The mean values of the metrics computed on the Hockey Fight dataset over the five splits of the stratified shuffle split, for each of the proposed models.

	Loss	Sensitivity	Specificity	Accuracy	F_1 score	AUC
C3D + SVM	-	**97.82 ± 0.80%**	**97.90 ± 1.24%**	**97.86 ± 0.56%**	**97.87 ± 0.55%**	**99.62 ± 0.30%**
C3D + FC	**0.1276 ± 0.0662**	96.93 ± 1.75%	96.40 ± 0.97%	96.67 ± 1.15%	96.69 ± 1.16%	99.27 ± 0.40%
ConvLSTM	0.1492 ± 0.0839	96.44 ± 1.19%	96.70 ± 1.54%	96.57 ± 0.79%	96.58 ± 0.78%	99.31 ± 0.32%

TABLE 13. The mean values of the metrics computed on the Crowd Violence dataset over the five splits of the stratified shuffle split, for each of the proposed models.

	Loss	Sensitivity	Specificity	Accuracy	F_1 score	AUC
C3D + SVM	-	**100.00 ± 0.00%**	**99.07 ± 1.18%**	**99.60 ± 0.50%**	**99.66 ± 0.43%**	**100.00 ± 0.01%**
C3D + FC	**0.0356 ± 0.0323**	99.59 ± 0.55%	98.32 ± 0.70%	99.05 ± 0.54%	99.18 ± 0.46%	99.94 ± 0.11%
ConvLSTM	0.3535 ± 0.0726	95.21 ± 1.37%	69.16 ± 15.15%	84.19 ± 5.95%	87.63 ± 3.89%	94.43 ± 2.15%

The sensitivity is greater than the specificity for all the three models, validating the purpose of the proposed dataset. In fact, the non-violent videos contain fast movements and contacts between the subject with the objective of testing the robustness of violence detection techniques, while preserving the capability of identifying violent scenes.

2) TESTS ON THE HOCKEY FIGHT AND CROWD VIOLENCE DATASETS

To compare the proposed models to those available in literature and described in Section II, we carried out the tests also on the Hockey Fight and Crowd Violence datasets. To this end, Table 12 lists the mean results got by the three models on the Hockey Fight dataset, while Table 13 includes the results on the Crowd Violence dataset. The model based on C3D and SVM confirms the good performance showed in our previous work: the accuracy is around 98% on the Hockey Fight dataset and above 99% on the Crowd Violence. The end-to-end model composed of C3D and two fully connected layers scores similar results: the accuracy is almost 97% on the Hockey Fight, and 99% on the Crowd Violence. In fact, the Crowd Violence is the smallest tested dataset: it includes 1265 16-frames chunks, while the Hockey Fight and the AIRTLab include 2007 and 3537 chunks respectively. Therefore, despite the 80-20 stratified shuffle split strategy, the models might overfit on the Crowd Violence. The ConvLSTM-based model behaves similarly to the other models on the Hockey Fight dataset, getting a 96.57% mean accuracy. However, on the Crowd Violence, the accuracy of the ConvLSTM-based models drops to 84.19%. Such model has more than 198 million of parameters, and might be too complex to converge to a relatively small dataset as the Crowd Violence, as also highlighted by the loss value computed on the test set, which is significantly greater than the loss of the model composed of C3D and fully connected layers. Moreover, most of the errors of the ConvLSTM are false positives: on the Crowd Violence, the mean specificity is 69.16% while the mean sensitivity is 95.21%. In fact, the videos of the Crowd Violence are quite similar on the two classes and, due to the smaller number of samples, the ConvLSTM struggles to converge.

Therefore, on the Hockey Fight and Crowd Violence datasets, the model based on C3D and SVM gets the best performance. Moreover, the two C3D-based models demonstrated able to perform the classification on different datasets, confirming the generalization capability of transfer learning methodologies. The detailed results on the Hockey Fight and Crowd Violence are available in Appendix A and Appendix B respectively, including the metrics on each split of the datasets.

The experimental results collected on the Hockey Fight and Crowd Violence datasets allow comparing our models with the related research works described in Section II. To this end, we reported their accuracy in Table 1. While the trained-from-scratch 3D CNNs proposed by Song *et al.* [21] and Li *et al.* [24] achieved a very good accuracy on the Hockey Fight dataset (99.6% and 98.3% respectively), they got lower results on the Crowd Violence dataset (94.3% and 97.2%). Instead, our model based on transfer learning with C3D combined with the SVM gets an excellent accuracy performance on both datasets, with 97.9% on the Hockey Fight and 99.6% on the Crowd Violence, confirming the results of our previous work [11]. The other proposed model based on C3D also has good results on both datasets, with 96.7% on the Hockey Fight and 99% on the Crowd Violence. However, the 3D CNN model proposed by Li *et al.* has fewer parameters than the C3D model we used and, therefore, it requires lower computational resources to detect violence in videos. Similarly to our work, Ullah *et al.* [8] proposed to use transfer learning with C3D, but they use the output of the second fully connected layer (“fc7”) as a feature descriptor instead of the first (“fc6”) as we did in our work. In fact, their accuracy (96% on Hockey Fight and 98% on Crowd Violence) is slightly lower than the one got by our models based on C3D. Furthermore, our model based on C3D and SVM is more balanced on both datasets than multi-stream networks, such as the one proposed by Sudhakaran and Lanz [23] which uses pre-trained 2D CNNs (AlexNet [37]) to process the videos frame by frame and than a ConvLSTM trained from scratch to detect violence in the sequence of frames. In fact, they scored 97.1% accuracy on the Hockey Fight dataset, and 94.5% on the Crowd Violence dataset.

TABLE 14. The mean values of the metrics computed on the AIRTLab dataset over the five splits of the stratified shuffle split, for the models composed of pre-trained 2D CNNs and the Bi-LSTM for the feature extraction.

	Loss	Sensitivity	Specificity	Accuracy	F_1 score	AUC
VGG16 + Bi-LSTM	**0.1314 ± 0.0142**	96.93 ± 0.90%	**90.78 ± 2.34%**	**94.92 ± 0.51%**	**96.25 ± 0.36%**	**98.91 ± 0.19%**
VGG19 + Bi-LSTM	0.3554 ± 0.0910	94.03 ± 3.30%	71.63 ± 15.52%	86.69 ± 6.07%	90.57 ± 3.97%	92.12 ± 3.44%
ResNet50V2 + Bi-LSTM	0.6331 ± 0.0006	**100.00 ± 0.00%**	0.00 ± 0.00%	67.23 ± 0.00%	80.41 ± 0.00%	51.07 ± 1.67%
Xception + Bi-LSTM	0.6298 ± 0.0034	**100.00 ± 0.00%**	0.00 ± 0.00%	67.23 ± 0.00%	80.41 ± 0.00%	55.47 ± 4.11%
NASNet + Bi-LSTM	0.3776 ± 0.0461	91.85 ± 3.27%	64.48 ± 17.73%	82.88 ± 3.92%	87.93 ± 2.11%	90.41 ± 2.42%

TABLE 15. The mean values of the metrics computed on the AIRTLab dataset over the five splits of the stratified shuffle split, for the models composed of pre-trained 2D CNNs and the ConvLSTM for the feature extraction.

	Loss	Sensitivity	Specificity	Accuracy	F_1 score	AUC
VGG16 + ConvLSTM	0.1169 ± 0.0201	97.48 ± 1.09%	91.81 ± 3.17%	**95.62 ± 0.56%**	**96.77 ± 3.88%**	99.11 ± 0.24%
VGG19 + ConvLSTM	**0.1105 ± 0.0221**	96.63 ± 1.39%	**93.01 ± 2.01%**	95.45 ± 0.85%	96.62 ± 2.01%	**99.14 ± 0.27%**
ResNet50V2 + ConvLSTM	0.6331 ± 0.0006	**100.00 ± 0.00%**	0.00 ± 0.00%	67.23 ± 0.00%	80.41 ± 0.00%	51.07 ± 1.67%
Xception + ConvLSTM	0.2450 ± 0.0344	95.13 ± 2.27%	80.17 ± 2.68%	90.23 ± 1.92%	92.89 ± 1.44%	95.61 ± 1.10%
NASNet + ConvLSTM	0.2972 ± 0.0192	91.93 ± 3.62%	79.05 ± 7.63%	87.71 ± 1.17%	90.95 ± 0.91%	94.77 ± 0.48%

TABLE 16. The mean values of the metrics computed on the Hockey Fight dataset over the five splits of the stratified shuffle split, for the models composed of pre-trained 2D CNNs and the Bi-LSTM for the feature extraction.

	Loss	Sensitivity	Specificity	Accuracy	F_1 score	AUC
VGG16 + Bi-LSTM	**0.1389 ± 0.0280**	**96.63 ± 0.73%**	94.30 ± 2.06%	**95.47 ± 1.17%**	**95.55 ± 1.11%**	**98.75 ± 0.49%**
VGG19 + Bi-LSTM	0.1538 ± 0.0398	95.05 ± 1.66%	**94.90 ± 3.10%**	94.98 ± 1.80%	95.02 ± 1.73%	98.42 ± 0.74%
ResNet50V2 + Bi-LSTM	0.6906 ± 0.0061	46.93 ± 45.07%	59.40 ± 48.51%	53.13 ± 6.27%	36.95 ± 30.40%	53.45 ± 7.12%
Xception + Bi-LSTM	0.3767 ± 0.0456	86.14 ± 7.61%	87.50 ± 5.27%	86.82 ± 3.08%	86.65 ± 3.66%	92.51 ± 1.75%
NASNet + Bi-LSTM	0.3336 ± 0.0355	86.93 ± 3.71%	87.60 ± 3.87%	87.26 ± 1.69%	87.27 ± 1.77%	93.33 ± 1.26%

Moreover, the multi-stream network based on the pre-trained VGG13 and the ConvLSTM proposed by Hanson *et al.* [27] scores similarly to our models on the Hockey Fight (98.1% accuracy), but it is still far from our results on the Crowd Violence dataset (96.3%).

Therefore, our models confirm the effectiveness of transfer learning-based architectures for violence detection, even when compared to the existing literature.

3) COMPARISON WITH MODELS BASED ON PRE-TRAINED 2D CNNs

To demonstrate the effectiveness of the three models proposed in this paper, we compare their performance with the metrics obtained by five well-established 2D CNNs, pre-trained on ImageNet and combined with a recurrent layer for feature extraction. The pre-trained 2D CNNs are VGG16, VGG19, ResNet50V2, Xception, and NASNet Mobile. As explained in Subsection III-C, we built five models by combining these 2D CNNs with a Bi-LSTM layer (and two fully connected layers for the final classification) and other five models by combining the 2D CNNs with a ConvLSTM layer (and two fully connected layers for the final classification). Therefore, in addition to our three models, we tested other ten models on the AIRTLab, Hockey Fight, and Crowd Violence datasets.

Table 14 includes the mean values of the metrics computed for the models composed of the 2D CNNs and the Bi-LSTM layer, tested on the AIRTLab dataset; instead, Table 15 lists the metrics for the models composed of the 2D CNNs and the ConvLSTM layer. Among the models based on the 2D CNNs, those using VGG16 get the best accuracy: VGG16 and the ConvLSTM has a mean accuracy of 95.62% (± 0.56%) whereas VGG16 and the Bi-LSTM gets 94.92% (± 0.51%). However, none of the 2D CNN-based models performs better than the models proposed in this paper, on the AIRTLab dataset. In fact, both the C3D plus SVM model and the ConvLSTM-based model get better accuracy, F_1 score, and AUC than all the 2D CNN-based models. The model composed of C3D and fully connected layers also has the same accuracy as VGG16 and the ConvLSTM.

Among the other 2D CNN-based models, the one using ResNet50V2 completely fails the training on the AIRTLab dataset, as highlighted by the high loss value (0.63). In fact, the model wrongly classifies all the negative samples, labeling them as positive, as the specificity is equal to 0 in all the splits of the stratified shuffle split cross-validation scheme. The specificity is lower than the sensitivity for all the 2D CNNs, meaning that most of the errors are false positives.

Table 16 and Table 17 list the results of the 2D CNNs with the Bi-LSTM layer and the 2D CNNs with the ConvLSTM layer on the Hockey Fight Dataset. Among the 2D CNNs, VGG16 with the ConvLSTM layer gets the best accuracy (97.31% ± 0.40%), followed by VGG19 with the ConvLSTM layer (96.36% ± 1.39%) and VGG16 with the Bi-LSTM layer (95.47% ± 1.17%). However, as happened on the AIRTLab dataset, the model composed of C3D and the SVM classifier has the highest accuracy. The model composed of C3D and fully connected layers and the ConvLSTM-based model get a slightly lower accuracy than VGG16 plus the ConvLSTM layer, but they perform better than any other tested 2D CNN).

TABLE 17. The mean values of the metrics computed on the Hockey Fight dataset over the five splits of the stratified shuffle split, for the models composed of pre-trained 2D CNNs and the ConvLSTM for the feature extraction.

	Loss	Sensitivity	Specificity	Accuracy	F_1 score	AUC
VGG16 + ConvLSTM	**0.0965 ± 0.0290**	**98.12 ± 1.58%**	**96.51 ± 1.14%**	**97.31 ± 0.40%**	**97.34 ± 0.42%**	**99.63 ± 0.15%**
VGG19 + ConvLSTM	0.1129 ± 0.0337	97.82 ± 1.61%	94.90 ± 2.20%	96.36 ± 1.39%	96.44 ± 1.35%	99.50 ± 0.32%
ResNet50V2 + ConvLSTM	0.5925 ± 0.0348	76.93 ± 14.95%	78.90 ± 10.33%	77.91 ± 3.55%	77.11 ± 5.96%	86.78 ± 1.22%
Xception + ConvLSTM	0.2491 ± 0.0244	92.47 ± 2.45%	92.70 ± 1.21%	92.59 ± 1.02%	92.60 ± 1.11%	96.01 ± 0.80%
NASNet + ConvLSTM	0.2787 ± 0.0615	91.58 ± 1.63%	91.00 ± 3.00%	91.29 ± 1.63%	91.33 ± 1.74%	96.21 ± 0.77%

TABLE 18. The mean values of the metrics computed on the Crowd Violence dataset over the five splits of the stratified shuffle split, for the models composed of pre-trained 2D CNNs and the Bi-LSTM for the feature extraction.

	Loss	Sensitivity	Specificity	Accuracy	F_1 score	AUC
VGG16 + Bi-LSTM	**0.0703 ± 0.0185**	99.58 ± 0.82%	94.39 ± 3.30%	**97.39 ± 1.02%**	**97.79 ± 0.84%**	**99.84 ± 0.10%**
VGG19 + Bi-LSTM	0.0817 ± 0.0193	97.53 ± 0.93%	**96.07 ± 1.91%**	96.92 ± 1.13%	97.34 ± 0.97%	99.61 ± 0.19%
ResNet50V2 + Bi-LSTM	0.6817 ± 0.0006	**100.00 ± 0.00%**	0.00 ± 0.00%	57.71 ± 0.00%	73.18 ± 0.00%	51.27 ± 2.02%
Xception + Bi-LSTM	0.5838 ± 0.0384	82.47 ± 6.32%	58.88 ± 12.23%	72.49 ± 3.76%	77.56 ± 2.87%	78.36 ± 2.66%
NASNet + Bi-LSTM	0.4427 ± 0.0489	89.32 ± 4.32%	66.17 ± 8.39%	79.53 ± 3.72%	83.45 ± 2.92%	86.72 ± 3.42%

TABLE 19. The mean values of the metrics computed on the Crowd Violence dataset over the five splits of the stratified shuffle split, for the models composed of pre-trained 2D CNNs and the ConvLSTM for the feature extraction.

	Loss	Sensitivity	Specificity	Accuracy	F_1 score	AUC
VGG16 + ConvLSTM	0.0781 ± 0.0265	97.26 ± 1.44%	96.64 ± 3.95%	97.00 ± 0.96%	97.41 ± 0.76%	99.80 ± 0.16%
VGG19 + ConvLSTM	**0.0434 ± 0.0214**	**98.63 ± 1.06%**	**98.88 ± 1.50%**	**98.74 ± 0.81%**	**98.90 ± 0.70%**	**99.93 ± 0.05%**
ResNet50V2 + ConvLSTM	0.5634 ± 0.0714	92.33 ± 6.36%	44.85 ± 26.35%	72.25 ± 8.44%	79.69 ± 4.20%	76.93 ± 14.14%
Xception + ConvLSTM	0.3691 ± 0.0742	89.45 ± 5.81%	82.05 ± 3.26%	86.32 ± 3.77%	88.23 ± 3.45%	93.19 ± 2.96%
NASNet + ConvLSTM	0.3961 ± 0.0640	90.27 ± 3.61%	82.05 ± 4.24%	86.80 ± 1.57%	88.73 ± 1.46%	93.99 ± 1.70%

ResNet50V2 gets bad results also on the Hockey Fight dataset, with an accuracy similar to a random classifier when combined with the Bi-LSTM, increasing up to 77.91% when combined with the ConvLSTM. The models based on VGG16 and VGG19 score a specificity lower the sensitivity (with more false positives than false negatives). Instead, the models based on Xception and NASNet Mobile have similar sensitivity and specificity on the Hockey Fight, even if they are in general worse classifiers, in terms of accuracy, F_1 score, and AUC than the VGG-based models.

Table 18 and Table 19 list the results of the 2D CNN models with the Bi-LSTM and the ConvLSTM on the Crowd Violence dataset. The models based on VGG16 and VGG19 get a significantly better accuracy than the other 2D CNNs. The best model is the one based on VGG19 and the ConvLSTM layer, with an accuracy equal to 98.74% (± 0.81%). However, our models based on C3D behaved, as classifiers, better than all the 2D CNNs, getting a better accuracy, F_1 score, and AUC on the Crowd Violence dataset. Only the ConvLSTM-based model struggled to train on such dataset, with an accuracy similar to the one obtained by Xception and NASNet Mobile with the ConvLSTM layer.

The model based on ResNet50V2 failed to learn the classification task also on the Crowd Violence dataset, getting the highest loss value and failing to correctly identifying the negative samples.

To summarize, among the pre-trained 2D CNNs tested in this paper, VGG16 and VGG19 get the best results on all the datasets, specifically when combined with a ConvLSTM layer. Xception and NASNet Mobile got significantly lower results, while ResNet50V2 had a very poor performance. In general, the 2D CNNs with the ConvLSTM obtain slightly better results than the 2D CNNs with the Bi-LSTM layer. Comparing the 2D CNN-based models with the models available in the literature and listed in Table 1, the effectiveness of transfer learning in the task of violence detection is confirmed. For example, the model combining VGG19 and ConvLSTM scores slightly better than the model proposed by Ullah *et al.* [8]. However, the models originally proposed in this paper get better results on the AIRTLab dataset, where the worst proposed model (C3D and the fully connected layers) gets the same accuracy of the best 2D CNN-based models; the two C3D-based proposed models get better performance than the 2D CNNs on the Crowd Violence; finally, the model based on C3D and the SVM classifier has the best accuracy on the Hockey Fight dataset; the ConvLSTM-based model, and the model composed of C3D and the fully connected layers get better accuracy than nine 2D CNNs-based models out of ten.

C. LIMITATIONS

The results of the research described in this paper are promising, but include some limitations. Concerning the proposed dataset, the videos were recorded by non-professional actors and, therefore, do not include real violence. For this reason, whilst the metrics computed for the proposed deep learning-based models are promising, the results cannot be considered general. Nevertheless, the proposed models were validated on the real videos of the Hockey Fight and Crowd Violence

datasets, in addition to being rooted in action recognition and violence detection literature.

Concerning the presented results, we built our model on the results of our previous work as well on the research related to violence detection, as explained in the Related Works section. However, a systematic study on alternative hyperparameters and models as well as a comparison on more datasets should be performed to get more general results, and therefore fully validate our methods.

Moreover, we tested our model on sequences composed of 16 frames taken from short video clips (the average length of a clip from the AIRTLab dataset is 5.6 seconds). In fact, most of the literature is based on tests with short videos. However, the accuracy on full length, real videos should be evaluated before going into production. Evaluating short sequences of frames taken from long videos might result in too many false positives, interfering in practical uses of the proposed techniques. Therefore, in order to maximize the accuracy on full length videos, results on the sub-sequences of frames should be merged together. In this regard, a simple strategy might be to label a part of a long video as positive only when a fixed number of consecutive 16-frames sub-sequences are labeled as positive.

V. CONCLUSION

We presented the architecture of three deep learning-based models for violence detection in videos: we tested them on the clips of the novel AIRTLab dataset, specifically designed to check the robustness against false positives, as well as on the Hockey Fight and Crowd Violence datasets, traditionally used in literature to benchmark violence detection techniques. The experiments presented in this paper allow drawing two main conclusions:

- the proposed transfer learning-based models (C3D combined with an SVM classifier and C3D combined with new fully connected layers) get stable accuracy results on all the three tested datasets, being better, in many cases, than the related works tested on the Hockey Fight and Crowd Violence. This suggests to persist with transfer learning-based models for the task of violence detection;
- our models based on 3D CNNs perform better than well-known 2D CNNs pre-trained on ImageNet and combined with a recurrent module to extract the spatio-temporal features of the videos in the datasets, suggesting continuing the research about 3D architectures for violence detection.

Moreover, all the proposed models demonstrated more capable of identifying violent videos than non-violent, given that most of the errors are false positives. Whilst this behaviour is partially affected by the fact that the samples from the two classes are unbalanced, it also validates the design of the AIRTLab dataset in checking the robustness against false positives.

In addition to dealing with the limitations of the described research, future works will address a deeper comparison between transfer learning-based models and trained from scratch models for violence detection, both on the AIRTLab dataset and on the other datasets available in scientific literature.

TABLE 20. Number of training epochs in each split (S1-S5) for the two end-to-end models i.e. C3D with two fully connected layers (C3D + FC) and the ConvLSTM-based architecture on the Hockey Fight dataset.

	S1	S2	S3	S4	S5	Mean
C3D + FC	7	16	24	9	10	13.20 ± 6.18
ConvLSTM	12	12	13	18	11	13.20 ± 2.48

TABLE 21. The results of the model composed of C3D and the SVM, computed for each split of the stratified shuffle-split cross validation scheme on the Hockey Fight dataset.

	Split 1	Split 2	Split 3	Split 4	Split 5
Sensitivity	97.03%	97.03%	97.52%	98.51%	**99.01%**
Specificity	97.50%	**99.50%**	97.50%	99.00%	96.00%
Accuracy	97.26%	98.26%	97.51%	**98.76%**	97.51%
F_1 score	97.27%	98.25%	97.52%	**98.76%**	97.56%
AUC	99.01%	**99.82%**	99.62%	99.80%	99.78%

APPENDIX A
RESULTS ON THE HOCKEY FIGHT DATASET

To test the proposed models on the clips of the Hockey Fight dataset, we followed the same experimental protocol applied on the AIRTLab dataset. Therefore, we applied a stratified shuffle split cross-validation scheme, randomizing a 80-20 split 5 times, with 80% of the data serving as the training set and 20% of the data serving as the test set. With the two end-to-end models, 12.5% of the training data was used as the validation set. The 1000 videos of the Hockey Fight dataset include 2007 16-frames chunks in total.

Table 20 lists the number of training epochs in each split of the data, for each end-to-end model. While the mean number of epochs is 13.2 for both models, the one based on C3D varied the number of training epochs in each split more significantly (standard deviation 6.18) than the ConvLSTM model (standard deviation 2.48). The batch size was 32 for the C3D model and 8 for the ConvLSTM model.

Table 21 includes the results obtained by the model composed of C3D and the SVM classifier in each split of the Hockey Fight dataset. The classifier is independent from the specific data split, as the results are similar across all the splits. With this dataset, the specificity is usually similar or higher than the sensitivity (except in split number 5). In fact, differently from the AIRTLab dataset, the Hockey Fight dataset is not tailored to specifically challenge the violence detection in labeling as non-violent rapid moves and behaviors which might resemble violent. However, in split 5, 8 out of 10 classification errors are false positives, with the 8 non-violent 16-frames chunks labeled as violent.

Table 22 shows the metrics computed for the C3D-based end-to-end model on the splits of the Hockey Fight dataset. The results are similar to those obtained with the model

TABLE 22. The results of the model composed of C3D and the fully connected layers for classification, computed for each split of the stratified shuffle-split cross validation scheme on the Hockey Fight dataset.

	Split 1	Split 2	Split 3	Split 4	Split 5
Loss	0.2556	0.0972	**0.0675**	0.1217	0.0961
Sensitivity	95.54%	98.02%	**99.50%**	97.03%	94.55%
Specificity	95.00%	**97.50%**	97.00%	95.50%	97.00%
Accuracy	95.27%	97.76%	**98.26%**	96.27%	95.77%
F_1 score	95.31%	97.78%	**98.29%**	96.31%	95.74%
AUC	98.59%	99.29%	**99.72%**	99.09%	99.60%

TABLE 23. The results of the model based on the ConvLSTM architecture, computed for each split of the stratified shuffle-split cross validation scheme on the Hockey Fight dataset.

	Split 1	Split 2	Split 3	Split 4	Split 5
Loss	0.3141	**0.0886**	0.0944	0.1177	0.1311
Sensitivity	95.54%	95.54%	97.03%	**98.51%**	95.54%
Specificity	96.00%	**99.00%**	98.00%	95.00%	95.50%
Accuracy	95.77%	97.26%	**97.51%**	96.77%	95.52%
F_1 score	95.78%	97.23%	**97.51%**	96.84%	95.54%
AUC	98.68%	**99.52%**	99.43%	99.42%	99.03%

based on C3D and SVM. Sensitivity and specificity are similar across all the splits, even if, in most of the splits, the specificity is slightly lower, highlighting a greater number of false positives.

Table 23 lists the results of the end-to-end model based on the ConvLSTM architecture, on the splits of the Hockey Fight dataset. The ConvLSTM-based network obtains a slightly lower sensitivity than the previous two models. Such model might be too complex to train from scratch on the 1405 16-frames chunks used as training data, being unable to correctly identify the violent samples. As the previous two models, the ConvLSTM architecture does not highlight a significant difference between false positives and false negatives, in terms of classification errors.

APPENDIX B
RESULTS ON THE CROWD VIOLENCE DATASET

To test the proposed models on the clips of the Crowd Violence dataset, we followed the same experimental protocol applied on the AIRTLab and Hockey Fight datasets: we applied a stratified shuffle split cross-validation scheme, randomizing a 80-20 split 5 times, with 80% of the data serving as the training set and 20% of the data serving as the test set. With the two end-to-end models, 12.5% of the training data was used as the validation set. The Crowd Violence dataset includes a total of 1265 16-frames chunks.

Table 24 includes the number of training epochs in each split of the data, for each end-to-end model. The model based on C3D exhibits the lowest mean number of training epochs of all the datasets: 10.6 ($\pm$ 3.2). In fact, the Crowd Violence dataset has the lowest number of samples, which results in a faster convergence of the neural network on the training set. The mean number of training epochs for the ConvLSTM-based model is 9.8 ($\pm$ 2.23). The batch size was 32 for the C3D model and 8 for the ConvLSTM model.

TABLE 24. Number of training epochs in each split (S1-S5) for the two end-to-end models i.e. C3D with two fully connected layers (C3D + FC) and the ConvLSTM-based architecture on the Crowd Violence dataset.

	S1	S2	S3	S4	S5	Mean
C3D + FC	8	8	8	15	14	10.60 ± 3.20
ConvLSTM	8	8	13	12	8	9.80 ± 2.23

TABLE 25. The results of the model composed of C3D and the SVM, computed for each split of the stratified shuffle-split cross validation scheme on the Crowd Violence dataset.

	Split 1	Split 2	Split 3	Split 4	Split 5
Sensitivity	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
Specificity	**100.00%**	98.13%	**100.00%**	**100.00%**	97.20%
Accuracy	**100.00%**	99.21%	**100.00%**	**100.00%**	98.81%
F_1 score	**100.00%**	99.32%	**100.00%**	**100.00%**	98.98%
AUC	**100.00%**	99.98%	**100.00%**	**100.00%**	**100.00%**

TABLE 26. The results of the model composed of C3D and the fully connected layers for classification, computed for each split of the stratified shuffle-split cross validation scheme on the Crowd Violence dataset.

	Split 1	Split 2	Split 3	Split 4	Split 5
Loss	0.0909	0.0161	0.0535	0.0107	**0.0070**
Sensitivity	99.32%	**100.00%**	98.63%	**100.00%**	**100.00%**
Specificity	97.20%	98.13%	98.13%	**99.07%**	**99.07%**
Accuracy	98.42%	99.21%	98.42%	**99.60%**	**99.60%**
F_1 score	98.64%	99.32%	98.63%	**99.66%**	**99.66%**
AUC	99.72%	99.98%	99.86%	99.99%	**100.00%**

TABLE 27. The results of the model based on the ConvLSTM architecture, computed for each split of the stratified shuffle-split cross validation scheme on the Crowd Violence dataset.

	Split 1	Split 2	Split 3	Split 4	Split 5
Loss	0.4845	**0.2876**	0.3780	0.3195	0.2977
Sensitivity	**97.26%**	94.52%	95.21%	93.15%	95.89%
Specificity	41.12%	81.31%	67.29%	72.90%	**83.18%**
Accuracy	73.52%	88.93%	83.40%	84.58%	**90.51%**
F_1 score	80.91%	90.79%	86.88%	87.46%	**92.11%**
AUC	90.47%	**96.78%**	93.91%	94.85%	95.74%

Table 25 shows the results obtained by the model composed of C3D and SVM on the Crowd Violence dataset. Due to the low number of samples, the architecture performs extremely well in terms of accuracy. All the few classification errors are false positives. Specifically, 2 chunks in split 2 and 3 chunks in split 5 were wrongly identified as violent.

Table 26 lists the results got on the Crowd Violence dataset by the end-to-end model based on C3D. In terms of accuracy, the model behaves similarly to the one based on C3D and SVM, with few classification errors. The loss is also less than 0.1 on the test set in all the splits, demonstrating the model's capability of classifying on the Crowd Violence dataset. Split 1 and split 3 exhibit the lowest accuracy 98.42% with 4 classification errors out of 353 testing samples. Specifically, in split 1, there are 3 false positives and 1 false negative; instead, in split 3, both the numbers of false negatives and positives is equal to 2.

Table 27 includes the results of the ConvLSTM-based end-to-end model on the Crowd Violence dataset. While the two C3D-based models performed extremely well, to the point

that they seem to overfit the data, the ConvLSTM-based model struggles in terms of classification accuracy. In fact, the model is too complex (198 million of parameters) for the low number of samples of the Crowd Violence dataset, as also the loss on the test set is very high compared to the values obtained by the model based on C3D and the fully connected layers. In addition, the similarity between violent and non-violent clips, as well as the low resolution, might have a role as the model tends to label samples as violent. This is evident in split 1: 63 out of the 107 negative samples are misclassified, while only 8 out of 146 positive samples are misclassified.

ACKNOWLEDGMENT

The presented research has been part of the Memorandum of Understanding between the Università Politecnica delle Marche, Centro "CARMELO" and the Ministero dell'Interno, Dipartimento di Pubblica Sicurezza, Direzione Centrale Anticrimine della Polizia di Stato.

REFERENCES

[1] Z. Xu, C. Hu, and L. Mei, "Video structured description technology based intelligence analysis of surveillance videos for public security applications," *Multimedia Tools Appl.*, vol. 75, no. 19, pp. 12155–12172, 2016.
[2] A. Castillo, S. Tabik, F. Pérez, R. Olmos, and F. Herrera, "Brightness guided preprocessing for automatic cold steel weapon detection in surveillance videos with deep learning," *Neurocomputing*, vol. 330, pp. 151–161, Feb. 2019.
[3] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 1–6.
[4] L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao, "Violent video detection based on MoSIFT feature and sparse coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3538–3542.
[5] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented VIolent flows," *Image Vis. Comput.*, vols. 48–49, pp. 37–41, Apr. 2016.
[6] Z. Meng, J. Yuan, and Z. Li, "Trajectory-pooled deep convolutional networks for violence detection in videos," in *Computer Vision Systems*, M. Liu, H. Chen, and M. Vincze, Eds. Cham, Switzerland: Springer, 2017, pp. 437–447.
[7] S. D. Jackson, E. Fenil, M. Gunasekaran, G. Vivekananda, T. Thanjaivadivel, S. Jeeva, and A. Ahilan, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM," *Comput. Netw.*, vol. 151, pp. 191–200, Mar. 2019.
[8] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence detection using spatiotemporal features with 3D convolutional neural network," *Sensors*, vol. 19, no. 11, p. 2472, May 2019.
[9] L. Zhang, G. Zhu, P. Shen, J. Song, S. A. Shah, and M. Bennamoun, "Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3120–3128.
[10] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human action recognition by learning Spatio-temporal features with deep neural networks," *IEEE Access*, vol. 6, pp. 17913–17922, 2018.
[11] S. Accattoli, P. Sernani, N. Falcionelli, D. N. Mekuria, and A. F. Dragoni, "Violence detection in videos by combining 3D convolutional neural networks and support vector machines," *Appl. Artif. Intell.*, vol. 34, no. 4, pp. 329–344, Mar. 2020.
[12] M. Bianculli, N. Falcionelli, P. Sernani, S. Tomassini, P. Contardo, M. Lombardi, and A. F. Dragoni, "A dataset for automatic violence detection in videos," *Data Brief*, vol. 33, Dec. 2020, Art. no. 106587.
[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[14] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision–(ECCV)*. Cham, Switzerland: Springer, 2016, pp. 630–645.
[15] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.
[16] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8697–8710.
[17] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Computer Analysis of Images and Patterns*, P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano, and W. Kropatsch, Eds. Berlin, Germany: Springer, 2011, pp. 332–339.
[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.
[19] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, vol. 1, 2015, pp. 802–810.
[20] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail, M. Ahmed, M. Ilyas, and A. Mahmood, "A review on state-of-the-art violence detection techniques," *IEEE Access*, vol. 7, pp. 107560–107575, 2019.
[21] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A novel violent video detection scheme based on modified 3D convolutional neural networks," *IEEE Access*, vol. 7, pp. 39172–39179, 2019.
[22] I. Serrano Gracia, O. Deniz Suarez, G. Bueno Garcia, and T.-K. Kim, "Fast fight detection," *PLoS ONE*, vol. 10, no. 4, Apr. 2015, Art. no. e0120448.
[23] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.
[24] J. Li, X. Jiang, T. Sun, and K. Xu, "Efficient violence detection using 3D convolutional neural networks," in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.
[25] M. Cheng, K. Cai, and M. Li, "RWF-2000: An open large scale video database for violence detection," 2019, *arXiv:1911.05913*.
[26] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, "Violence detection in video by using 3D convolutional neural networks," in *Advances in Visual Computing*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, R. McMahan, J. Jerald, H. Zhang, S. M. Drucker, C. Kambhamettu, M. El Choubassi, Z. Deng, and M. Carlson, Eds. Cham, Switzerland: Springer, 2014, pp. 551–558.
[27] A. Hanson, P. Koutilya, S. Krishnagopal, and L. Davis, "Bidirectional convolutional LSTM for the detection of violence in videos," in *Computer Vision–(ECCV) Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham, Switzerland: Springer, 2019, pp. 280–295.
[28] Y. LeCun and Y. Bengio, *Convolutional Networks for Images, Speech, and Time Series*. Cambridge, MA, USA: MIT Press, 1998, pp. 255–258.
[29] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
[30] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.
[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[32] A. Graves, N. Jaitly, and A.-R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 273–278.
[33] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, vol. 1, 2014, pp. 568–576.
[34] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: Unsupervised learning using temporal order verification," in *Computer Vision–(ECCV)*. Cham, Switzerland: Springer, 2016, pp. 527–544.
[35] S. Wager, S. Wang, and P. Liang, "Dropout training as adaptive regularization," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, vol. 1, 2013, pp. 351–359.
[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.

PAOLO SERNANI received the Ph.D. degree in information engineering from Università Politecnica delle Marche, Ancona, Italy, in March 2016, defending a thesis titled "Design and virtualization of intelligent systems for the management of assistive environments," dealing with the modeling and engineering of smart homes and ambient intelligence environments as multi-agent systems.

He is currently a Postdoctoral Research Fellow at the Artificial Intelligence and Real-Time Systems Laboratory. Università Politecnica delle Marche. His main research interests include deep learning for image and video analysis, multi-agent systems, expert systems, and decision support systems.

NICOLA FALCIONELLI received the Ph.D. degree from the Università Politecnica delle Marche Doctoral School, in April 2020, successfully defended a thesis with title "From symbolic artificial intelligence to neural networks universality with event-based modeling."

He is currently a Postdoctoral Research Fellow at the Artificial Intelligence and Real-Time Systems Laboratory, Università Politecnica delle Marche, Ancona, Italy. His research interests include the spectrum of artificial intelligence, from symbolic approaches to statistical learning techniques, including common-sense reasoning (event calculus, fact caching, and indexing), neural networks (deep learning, spiking neural networks, and hardware implementations), and multi-agent systems.

SELENE TOMASSINI (Student Member, IEEE) received the B.S. and M.S. degrees in biomedical engineering from Università Politecnica delle Marche, Ancona, Italy, in 2016 and 2018, respectively, where she is currently pursuing the Ph.D. degree in information engineering with the Artificial Intelligence and Real-Time Systems Laboratory, Department of Information Engineering, with a scholarship financed by the CARIVERONA Foundation for the project "Using 3D convolutional neural networks for the recognition of lung cancer histotypes directly from CT scans."

From 2015 to 2016 and in 2018, she was an Undergraduate Intern at the Cardiovascular Bioengineering Laboratory, Department of Information Engineering, Università Politecnica delle Marche, focusing her research interest on biomedical signal processing (mainly, electrocardiography and phonocardiography). In 2019, she was a Postgraduate Intern at the Cardiovascular Bioengineering Laboratory, working on phonocardiographic signal processing techniques. Her current main research interest includes machine and deep learning architectures for biomedical data analysis.

PAOLO CONTARDO received the M.Sc. degree in industrial mechanical engineering from Università Politecnica delle Marche, Ancona, Italy, in 2019, with a M.Sc. thesis titled "Dactyloscopy 2.0: New perspectives on Fingerprint identification," and he is currently pursuing the Ph.D. degree at the Information Engineering Department.

In November 2020, he joined the Artificial Intelligence and Real-Time Systems Laboratory, Università Politecnica delle Marche, as a Ph.D. Student for the projects "Dactyloscopy 2.0," "Fotosegnalamento 2.0," and "Violence detection in videos" which fall within the topics covered with the understanding agreement between the UNIVPM CARMELO inter-departmental research center (Center for Advanced Research on Measurements for Engineering and Life Optimization) and the Ministero dell'Interno, Dipartimento della Pubblica Sicurezza, and Direzione Centrale Anticrimine della Polizia di Stato. His current main research interests include machine learning and deep learning techniques for the analysis of biometrics data.

ALDO FRANCO DRAGONI received the Laurea degree in electronics engineering from the University of Ancona, Italy, discussing a thesis in artificial intelligence about "Plan Recognition from Visual Information."

He is currently in charge as an Associate Professor at the Università Politecnica delle Marche, where he teaches "fundamentals of computer sciences," "artificial intelligence," and "dedicated operating systems." Moreover, he is the Head of the Artificial Intelligence and Real-Time Systems Laboratory, Università Politecnica delle Marche, focusing on artificial intelligence problems to be solved within precise time constraints (deadlines). In addition, he opened a new application area for artificial intelligence, called "NetMedicine," which means every "intelligent" health-related activity which is carried on through the Internet. His scientific interests include several aspects of artificial intelligence, from classic knowledge-based approaches to more advanced hybrid systems that integrate symbolic reasoning with neural networks.