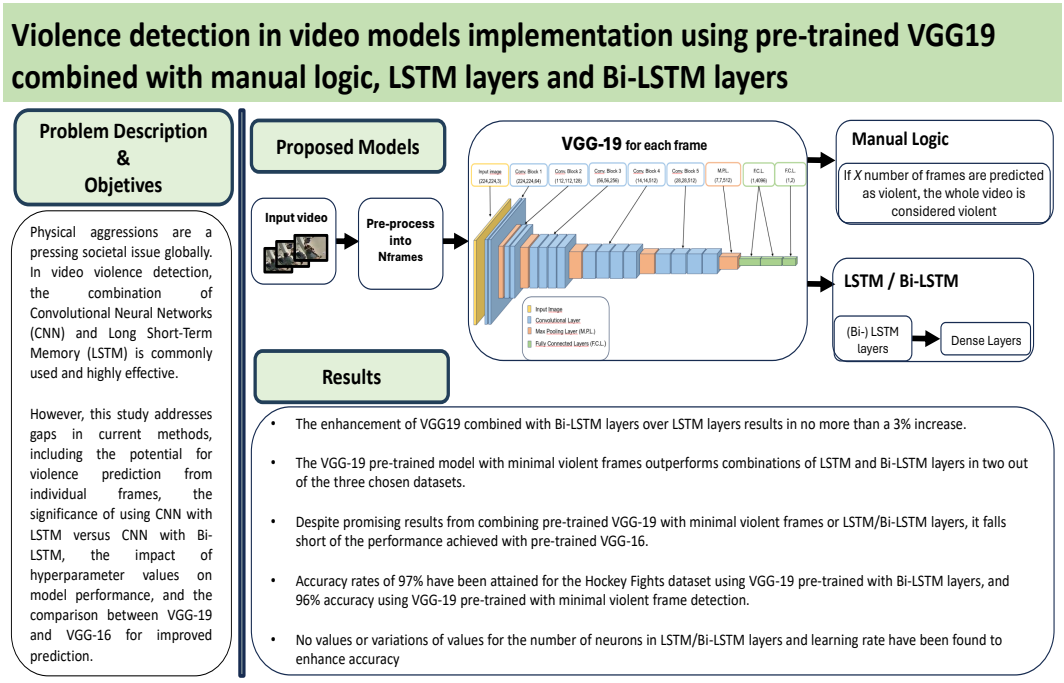
Graphical Abstract

Violence detection in video models implementation using pre-trained VGG19 combined with manual logic, LSTM layers and Bi-LSTM layers

Pablo Negre, Ricardo S. Alonso, Javier Prieto, Óscar García, Juan M. Corchado



Highlights

Violence detection in video models implementation using pre-trained VGG19 combined with manual logic, LSTM layers and Bi-LSTM layers

Pablo Negre, Ricardo S. Alonso, Javier Prieto, Óscar García, Juan M. Corchado

- The improvement of VGG19 combined with Bi-LSTM layers over LSTM layers is not more than a 4% increase.
- The VGG-19 pre-trained model with minimal violent frames achieves superior results compared to LSTM and Bi-LSTM layers combination in two of the three selected dataframes.
- Although the combination of pre-trained VGG-19 with minimal violent frames or with LSTM or Bi-LSTM layers yields promising results, it fails to surpass the performance achieved by utilizing pre-trained VGG-16.
- Accuracies of 97% have been achieved for the Hockey Fights dataset using VGG-19 pre-trained with Bi-LSTM layers and 96% accuracy using VGG-19 pre-trained with minimal violent frame detection.
- No values or values combinations of the number of neurons in LSTM/Bi-LSTM layers and learning rate have been found to increase the accuracy.

Violence detection in video models implementation using pre-trained VGG19 combined with manual logic, LSTM layers and Bi-LSTM layers

Pablo Negre[a], Ricardo S. Alonso[b,c], Javier Prieto[a], Óscar García[1], Juan M. Corchado[a]

[a]*BISITE Research Group, Universidad de Salamanca, Patio de Escuelas, 1, Salamanca, 37008, Castilla y León, Spain*
[b]*AIR Institute, Paseo de Belen 9A, Valladolid, 47011, Castilla y León, Spain*
[c]*UNIR (International University of La Rioja), Av. de la Paz, 137, Logroño, 37008, La Rioja, Spain*
[d]*Departamento de Informática, Universidade do Minho, Braga, 4704-553, Portugal*

Abstract

Video violence detection is crucial for societal safety, and although CNN and LSTM models are widely used for this purpose, significant gaps remain. This study investigates whether violence prediction can be achieved from individual frames without temporal analysis, evaluates the advantage of CNN with Bi-LSTM over CNN with LSTM, and examines the impact of hyperparameters, such as neuron count and model type (VGG-16 vs. VGG-19), on accuracy. Our results indicate that frame-by-frame analysis using pre-trained VGG-19 yields high accuracy (95% on the Hockey Fights dataset and 96% on Violent Flow) when setting a minimum threshold of violent frames. Additionally, Bi-LSTM layers provide a marginal improvement (up to 4%) over LSTM in certain datasets, while changes in hyperparameters, such as neuron count, do not consistently enhance accuracy. Notably, VGG-16 outperforms VGG-19 for video violence detection.

Email addresses: pablo.negre@usal.es (Pablo Negre), ralonso@air-institute.com (Ricardo S. Alonso), ricardoserafin.alonso@unir.net (Ricardo S. Alonso), javierp@usal.es (Javier Prieto), oscar.garcia.garcia@unir.net (Óscar García), corchado@usal.es (Juan M. Corchado)

URL: ORCiD (Pablo Negre), ORCiD (Ricardo S. Alonso), ORCiD (Javier Prieto), ORCiD (Óscar García), ORCiD (Juan M. Corchado)

Keywords: Violence detection, Physical aggression, Video Surveillance, Pre-trained VGG19, Long Short Term Memory (LSTM), Manual feature

1. Introduction

Physical aggressions are a significant issue in our society, impacting individuals globally. This issue affects multiple facets of life, impacting in the direct victims and their mental health [30], their families, communities, and the everyday progress of the country (people movement, tourism, shopping...) [27]. Regarding the causes that cause or encourage aggressive behaviours, those are caused by a person's difficulty on managing emotions, by the existence of disputes between individuals [35], as well as by the socio-economic and demographic features of societies [14].

Two of the groups that have been most affected by physical attacks have been women and minors. Regarding women, based on a World Health Organization (WHO) report, approximately one out of every three women globally has experienced either physical or sexual violence from an intimate partner or sexual violence from a non-partner during their lifetime [37]. An ActionAid study found that 79% of Indian women, 86% of Thai women, 89% of Brazilian women, and 75% of women in London face harassment or violence in public spaces [51]. Regarding children, work carried out by [21] and published in the American Academy of Pediatrics states that at least 50% or more of children in Asia, Africa and North America experienced violence in 2015, and that globally more than half of all children (1 billion children, aged 2-17) experienced such violence.

Overall, acts of physical aggression present a widespread issue on a global scale, impacting nearly every facet of citizens' lives and entire societies. Safeguarding the physical well-being of individuals worldwide ought to be a paramount concern and a right accessible to all.

2. State of the art

This Section explains the state of the art of video violence detection by means of artificial intelligence, and is divided into four sections. Section 2.1 outlines the solutions that have been applied to address violence in societies. Section 2.2 outlines the basic steps for AI-mediated violence detection. Section 2.3 outlines the basic steps for AI-mediated violence detection. Section

2.4 presents recent papers using the combination of Convolutional Nerual Networks (CNN) and Long Short-Term Memory (LSTM).

2.1. Approaches to addressing Violence

This Section explains the approaches that have been taken to address the problem of physical aggressions in societies. Multiple ways of dealing with violent acts in societies have been studied. Starting with less direct and long-term solutions, multiple studies seek to understand what situations or environments provoke people to commit violent acts, so that they can be remedied [30] [29] [16]. Other studies, with a more direct influence, analyse the relationship between the crime rate and the urban environment in which it occurs [49] [26] [53]. However, the last barrier against victims of violence is the real-time detection of violence in order to alert the authorities to mediate the conflict, as well as the collection of video or images that can later be used to identify the persons involved.

The implementation of video violence detection through artificial intelligence plays a great role in large-scale detection without the need for a large number of people to be monitoring video cameras or patrolling the streets [12]. This field is experiencing expansion [36], made feasible by the advancement of three primary pillars: heightened utilization of images and security cameras [2] [50], technological progress in big data platforms [6] [3], and the evolution of artificial intelligence algorithms [10] enabling image and video analysis. Violence detection in video through artificial intelligence belongs to the field of study of computer vision and is a subset of human action detection in video. Within the detection of human actions, this is considered as an anomalous action because it is an unusual or unexpected action in many social settings or context.

2.2. Basic violence detection steps by means of artificial intelligence

The basic steps that make up the training and real-time operation of a violence detection algorithm will be outlined in this Section. These are shown in Figure 1, which is divided into *Training* and *Real life situation*. The *Training* part contains the steps to train an artificial intelligence algorithm, which are: starting from a labelled dataframe (violence and non-violence videos) where the videos are usually cropped to contain the same number of frames and same image size (number of pixels in height and width); then preprocessing the data, converting the videos to the input data format to be passed to the violence detection algorithm.

There are many types of violence detection algorithms, but basically they are based on extracting features from the videos (spatial, temporal, etc.) and training an algorithm that learns from these features. There are algorithms that do this process in two stages (first features are extracted and another algorithm learns from them), others in which the same algorithm performs the extraction and learning of the extracted features, and others in which two algorithms extract and learn features (one after the other). Once this process is done, the extracted features are fed into a classifier that decides whether the scene is violent or not. This process requires testing multiple combinations of hyperparameters and evaluation of metrics until an optimal structure is found.

On the other hand, real-time violence detection using artificial intelligence consists of receiving a continuous image or video signal. The image must be pre-processed in the same way as the pre-processing was done in the training, so that the trained algorithm receives the same type of input. Since continuous analysis of real-time data is computationally and economically expensive, and especially if it involves image or video, there are computationally lighter techniques that can be used to analyse with the violence detection algorithm only those images or group of frames that may potentially contain violence; the extraction of characteristic frames is optional and not all state-of-the-art articles propose it. Finally, those images or frames that may contain violence are introduced to the trained algorithm that will classify the information as violent and non-violent.

2.3. Violence video algorithm types and datasets

This Section will present the most widely used datasets of violence videos, as well as the different types of violence detection algorithms that have been used in the state of the art.

Given that the detection of violence is an anomalous action and that a physical aggression can be very different (number of people involved, form of aggression, level of aggressiveness...), having an optimal amount of data for training the algorithm is crucial. In the state of the art there are numerous datasets that have emerged over the last few years, although a small group of them is used on a recurrent basis in order to be able to contrast results with other past articles. Some examples of these widely used datasets are: *Hockey Fights* [8] which contains videos about the National Hockey League (NHL), *Action Movies* [8] which includes scenes from action films, *Violent*

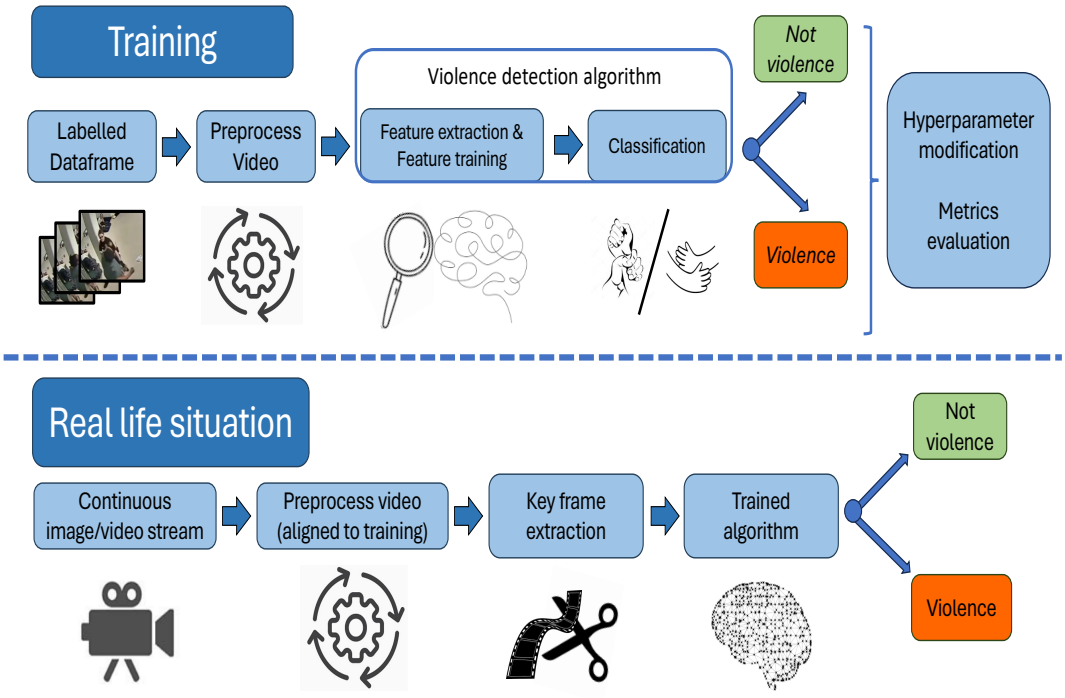


Figure 1: Basic steps in the implementation of a video violence detection algorithm.

Flow [20] or also referred to as *Violent Crowd*, containing YouTube videos depicting real crowded scenes with audio, *Real World Fight-2000 (RWF-2000)* [9] contains fights captured by surveillance cameras in real-world settings, and *Real Life Violence Situations (RLVS)* [39] comprises real-life violent scenarios captured from YouTube videos.

A wide variety of algorithms have been developed to detect violence, which are commonly categorized into the following types: *Traditional Methods*, which are those based on the use of manual feature extraction and the use of traditional Machine Learning algorithms, and *Deep Learning Methods*, which are those based on Deep Learning techniques [47], [46], [17], [31]. However, this division is too generic; therefore, we will now present some of the work carried out in video violence detection with artificial intelligence, grouped by the nature of the algorithms used. These groupings are as follows: CNN, LSTM, Manual feature, Skeleton based (Deep Learning or Manual), Transformer and Audio based (Deep Learning or Manual).

- **CNN**: works employing CNNs for violence detection focus on extracting spatial features from video frames to inform classification decisions. However, the use of CNN-3D and CNN-4D enables the extraction of both temporal and spatial patterns [36]). The CNN extracts features while training by modifying the values of its weights. There have also been several studies in which the use of pre-trained CNNs on large image datasets has been shown to improve accuracy by using fine-tunning techniques, i.e. taking advantage of the pre-trained weights of the network. Ahmed et al. [4] introduce a paper utilizing CNN-v4, which extracts positional and chronological data; the paper emphasizes that while other cutting-edge approaches in physical aggression detection with CNNs involve processing all video frames, this proves to be computationally demanding. Hence, the strategic selection of specific frames becomes crucial for more efficient computation. Jayasimhan et al. [25] suggest a hybrid model consisting of a 3D CNN succeeded by a 2D CNN. Not reliant on transfer learning or handcrafted extraction features, this architecture remains lightweight. Bi et al. [47] uses as key feature extraction method the number of selected relevant frames selected to exclude incorrect detections (like hugs or similar actions that could be mistaken for violence). As violence detection algorithm ResNet18 it's used as it has fewer parameters than other used pre-trained CNN.

- **LSTM**: works based on the use of LSTMs extract temporal features of violence (they can also be understood as temporal patterns). LSTMs, extracts features while training by modifying the values of its weights. However, to the best of our knowledge, no individual LSTM has been employed in the latest leading state of the art for violence detection. Combinations of CNNs and LSTMs have of course been used and, to a lesser extent, a combination of two RNNs has been used, as demonstrated by Ullah et al. [46] which combines the implementation of a ConvLSTM and GRU network structure.

- **Manual feature**: are those algorithms in which feature extraction, feature training, or both, are performed on a mathematical basis not relying on deep learning or machine learning. Jaiswal [24] utilizes *Local Binary Pattern (LBP)* and *Fuzzy Histogram of Optical Flow Orientations* for manual feature extraction. It adopts AdaBoost (Adap-

tive Boosting) as the feature training technique, a Machine Learning method. For classification purposes, it employs Ensemble RobustBoost aggregation, a technique rooted in decision trees.

- **Skeleton based (Deep Learning or Manual)**: those algorithms, aim to identify body positions of individuals in videos and infer the presence of violence based on these positions. These works employ techniques based on mathematics or deep learning. Srivastava et al. [40] introduce a method for violence detection utilizing a novel dataset composed of images captured by a drone at an elevated position from the ground. Their approach involves human figure identification, pinpointing key postural elements employing a CNN with dual inputs. Subsequently, an SVM classifier is employed to categorize physical aggresion detection based on these features that have been extracted and trained. Su et al. [42] extract skeleton-based features, creating a geometric map in the X, Y, Z dimensions where Z signifies the temporal aspect; this method effectively discerns "heads" to track their movements. By employing these positional geometric maps, they utilize SPIL (Skeleton Points Interaction Learning) to facilitate the classification training process.

- **Transformer**: this group is based on Transformer-based architectures to detect violence in video content. Akti et al. [5] employ the *ViT* algorithm, short for *Vision Transformer*, which melds transformer-based vision frameworks with self-attention mechanisms. This process initiates by segmenting the image into patches, extracting features from each patch while considering its spatial placement within the original image. Subsequently, these details traverse another layer that discerns temporal connections among the patches before culminating in a dedicated classification. Furthermore, the study introduces a dataset comprising images and videos sourced from the Internet. Ehsan et al. [13] focuses on extracting features from videos by identifying individuals, removing backgrounds using Yolo, and calculating image optical flow with the Farneback method. They utilize STAT, a generative adversarial network (GAN) that translates temporal motion features from video sequences into static images. This network comprises a generator that transforms motion features into images and a discriminator that evaluates the authenticity of these synthetic images compared to real

ones.

- **Audio based (Deep Learning or Manual)**: this category encompasses studies that utilize video audio for detecting violence. While these works might not exclusively depend on audio as the algorithm's input, its significance is emphasized within these works for its distinct focus. Mahalle & Rojatkar [28] indicate that the primary aim of employing audio feature extraction is to condense data dimensions by capturing the most crucial aspects from audio samples. By decreasing the number of dimensions of the feature vector, a concise set of features becomes effective in encapsulating the traits of audio samples. These extracted features are then used to assign labels to each instance. Subsequently, an Extreme Learning Machine (ELM) is introduced to the labeled audio data to be trained specifically for the identification of violent audio instances. Wu et al. [52] introduce XD-Violence, a substantial multi-scene dataset. Their violence detection model integrates a neural network with three simultaneous branches designed to analyze diverse relationships within video segments and combine different attributes. The comprehensive branch captures broad dependencies based on similarity precedence, the localized branch identifies positional correlations within the local area through proximity precedence, and the score branch adapts to evaluate the proximity of the anticipated score

All in all, there are many algorithms of different natures for generating violence detection algorithms, each of them with their approach. Additionally, merging two algorithms, regardless of whether they differ in nature, is the most common approach in current research [33]. This strategy aims to harness the strengths of both algorithm types. In the context of violence detection, the most frequent combination involves the use of CNN and LSTM.

2.4. Violence detection using CNN and LSTM combination

The combination of CNN and LSTM for violence detection is the most widely used architecture in the recent state of the art. It has shown very good performance, achieving outstanding results compared to other approaches in the violence detection sector. Its architecture is based on the fact that CNN extracts spatial features from the frames that make up the video. These extracted spatial features are fed to the LSTM in order to capture temporal

patterns from them [32] [48]. Within the combination of CNN and LSTM it is possible to separate between the use of non-pre-trained CNN and pre-trained CNN, where the use of pre-trained CNN and LSTM is higher.

Tables 1 and 2 set out detailed information on the papers collected in the systematic mapping study developed that use a combination of CNNs and LSTMs (or other types of RNNs such as GRUs) [33].

Table 1: Violence detection articles which use CNN and LSTM combination. Part 1.

Cite	**CNN**	**LSTM**	**Classification**	**Train P.T. Layers**	**LSTM Layers**	**LSTM Neurons**	**Dense Layers**	**Dense Neurons**
[48]	MobileNet V2	LSTM	F.C.L. (N.S.)	N	1	128	2	32,2
[19]	Own CNN	Bi-LSTM	F.C.L (N.S.)	No P.T.	N.S.	N.S.	N.S.	N.S.
[1]	PT VGG-16 (ImageNet)	LSTM	F.C.L (Sigmoid)	N	1	50	4	64,64,64,2
[31]	PT VGG-16 (ImageNet)	Bi-Conv-LSTM	F.C.L. (SoftMax)	N	3	256	4	1000,256,10,2
[23]	PT MNAS CNN	Conv-LSTM	Random Forest, SVM, K nearest neighbour	N	N.S.	N.S.	N F.C.L	N F.C.L
[45]	Two PT EfficienNet-B0 (ImageNet)	Bi-LSTM	F.C.L. (Sigmoid)	N	N.S.	N.S.	3	N.S.
[7]	Two PT VGG-16 (ImageNet) + Wide Dense Residual Blocks (WDRB)	LSTM	F.C.L. (SoftMax)	N	1	512	1	2
[18]	PT VGG-16 (ImageNet)	LSTM/Bi-LSTM	F.C.L (N.S.)	N	1	64	N.S.	N.S.
[47]	Darknet + Residual Optical Flow	M-LSTM	F.C.L (SoftMax)	No P.T.	N.S.	N.S.	N.S.	N.S.
[44]	PT VGG16 (INRA)	Bi-GRU	F.C.L. (SoftMax)	N	N.S.	N.S.	3	512,256,2
[22]	Own CNN	LSTM	F.C.L. (SoftMax)	No P.T.	1	128	3	256,16,2
[38]	PT Xception (ImageNet)	LSTM	F.C.L (SoftMax)	Y (4 last layers)	1	512	3	128,32,2
[43]	CNN	LSTM	F.C.L	No P.T.	N.S.	N.S.	N.S.	N.S.
[32]	PT VGG-19 (ImageNet) + extra CNN layers	Bi-LSTM	F.C.L (SoftMax)	Y (N.S.)	2	128,64	N.S.	N.S.
[41]	PT: VGG16/ VGG19/ InceptionV3...	LSTM	F.C.L (SoftMax)	N	1	512	3	1024,512,2
[11]	PT MobileNetV2 (ImageNet)	Bi-LSTM/ ConvLSTM	F.C.L. (SoftMax)	N	1/1	128/64	2	128/256

Table 2: Violence detection articles which use CNN and LSTM combination. Part 2.

Cite	P.T. Uses F.C.L.	Learning Rate	Hockey Fight	Action Movies	Violent Flow	RWF-2000	RLVS
[48]	No P.T.	N.S.	99.5	96.1	X	82	X
[19]	No P.T.	10^{-2}	99.27	100	98.64	X	X
[1]	Y	10^{-3}	99.1	X	X	X	X
[31]	N	10^{-4}	99.1	100	98.4	92.4	X
[23]	No F.C.L	N.S.	99	96	100	X	X
[45]	N	10^{-3}	99	X	93.75	X	96.74
[7]	N	10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}	98.8	98.99	97.1	X	X
[18]	Y	N.S.	97.6/98.8	X	92.2/95.1	X	X
[47]	No P.T.	10^{-4}	98	X	98.21	X	X
[44]	N	10^{-4}	98	X	95.5	X	90.25
[22]	No P.T.	10^{-4}	97	100	X	X	X
[38]	N.S.	N.S.	96.55	98.32	X	X	X
[43]	N	10^{-2}	94.9	92.2	77.31	X	X
[32]	Y	10^{-2}	91.29	X	90.47	90.47	X
[41]	Y	10^{-2}	X	X	X	X	X
[11]	N	N.S.	X	X	X	X	X

The columns of Table 1 are set out below. In case any of the columns have values separated by slashes "/", it means that the article has tested different combinations of architecture or hyperparameters. On the other hand, the value "N.S." means that the value is not clearly specified in the article. The column *CNN* contains the CNNs used in the papers for the extraction of the spatial characteristics of the video frames; in the case that these CNNs are pre-trained, it is expressed as "*PT CNN_Name (Trained_CNN_dataset)*". It can be seen that 11 of the 16 articles (68%) use pre-trained networks, so the use of transfer-learning is widely used. It is worth noting that of the articles selected in the table, to our knowledge none of them test how much the use of LSTMs improves the prediction of violence, as opposed to only the use of CNNs.

The column *LSTM* contains the type of RNN used, which is fed by the spatial features extracted by the CNN to extract the temporal features. A 50% uses LSTMs, 33% Bi-LSTMs, 11% Conv-LSTMs and only 5.5% GRU (which argues for its use in that it solves gradient fading more efficiently). LSTMs are a type of recurrent unit designed to overcome the problem of gradient fading in recurrent neural networks; it has a cell structure that allows for long-term information storage and retrieval, which makes it effective for modelling temporal sequences [48] [38] [7]. On the other hand, Bi-LSTMs are a variant of LSTMs that process the input sequence in two directions: forward and backward; they combine contextual information from both past and future time steps, which can help to capture bidirectional patterns in sequential data [19]. Finally, Conv-LSTMs combine the LSTM cell structure with the spatial processing capability of convolutional layers [23] [11].

Only one of the selected articles contains a comparison between LSTM and Bi-LSTM [18], where it is concluded that the combination of VGG-16 and Bi-LSTM outperforms the combination of VGG-16 and LSTM. When checking the results it is observed that in the Hockey Fights dataset the use of Bi-LSTM versus LSTM represents an increase of 97.6% compared to 98.8% (1.2% increase) and in the Violent Flow dataset the use of Bi-LSTM versus LSTM represents an increase of 92.2% compared to 95.1% (2.9% increase).

The column *Classification* contains the type of classifier used in the papers. In all cases, the *fully connected layers* (F.C.L), also called *dense layers*, are used, except in the work of Jahlan et al. [23] which opts to use several classical Machine Learning classifiers to test their accuracy; the *SoftMax* activation function is the most used for the last dense layer, although it is a multi-class activation function unlike the Sigmoid, for example. The col-

umn *Train P.T. Layers* indicates whether algorithms using pre-trained CNNs train some of the layers with the violence datasets. This is done to couple the pre-trained weights with the violence datasets, adjusting them to this new classification task. Only two of the selected papers do this, with the Sharma et. al paper being the only one that specifies that the weights are unfrozen for training with the violence data from the last 4 layers of the CNN [38].

In the following last four columns of Table 1 the number of layers and neurons which have been used in LSTM and L.C.F. have been compiled in order to understand if there is a certain combination which clearly obtains better results, however this is not the case. Column *LSTM layers* contains the number of LSTM layers that have been implemented; in several of the articles their number is not clearly specified, in 8 of them a single LSTM layer is used, and only in two of them 2 *mumtaz2022deep* and 3 *mugunga2021frame* LSTM layers are used (arguing that a higher number of layers allows more complex temporal patterns to be extracted). Column *LSTM Neurons* contains the number of neurons in the LSTM layers. While there are several articles that do not specify this, these are typical values: 64, 128 and 512. Column *Dense Layers* contains the number of dense layers acting as classifiers; typical values are the use of between 2 and 4 layers (including the output layer which must contain two neurons, given that the problem is divided between violence and non-violence), where a higher number of dense layers implies a higher learning capacity, but also a higher computational cost and a higher risk of overfitting [41] [31]. Column *Dense Neurons* contains the number of neuron that are implemented for the densely connected layers. Typical values are 512, 256 and 128.

We now turn to the data contained in Table 2 Column *P.T. Uses F.C.L.* which without acronyms stands for *Pre-trained uses Fully Connected Layer* indicates whether the algorithms using pre-trained CNNs maintain the fully connected layers when doing Transfer-Learning. As discussed above, transfer-learning leverages pre-trained neural networks on large datasets in order to take advantage of weights that have already been tuned for other purposes. After a series of convolutional layers, the pre-trained CNNs have a series of dense layers that act as a classifier with the last layer having as many neurons as classes the CNN can predict. This column therefore indicates whether the violence detection algorithm uses only the convolutional layers (whereby the column has value "N" or No) or whether it contains the convolutional layers and the dense layers excluding the last layer (whereby the column has value "Y"). If the column has the value *No P.T.* it means that the algorithm uses a

non-pre-trained CNN. In total, 6 of the 10 articles that use pre-trained CNNs do not include the dense layers for spatial feature extraction; the other 4 do.

The column *Learning Rate* contains the value of the learning rate implemented in the training process of the algorithm. The values used range from 10^{-2} to 10^{-4}. The columns *Hockey Fight*, *Action Movies*, *Violent Flow*, *RWF-2000* (Real World Fight - 2000 dataset) and *RLVS* (Real Live Violence Situations) contain the test accuracy obtained by the works performing the training and testing process with these datasets. These datasets have been selected because they are the 5 most used violence datasets in the recent state of the art [33] with the columns ordered from most used (*Hockey Dataset*) to least used (*RLVS*). It can be observed on the one hand that while most algorithms have trained and tested their proposed algorithm with the *Hockey Fight Dataset*, this is decreasing with the rest of the datasets, making drawing conclusions based on such results more complex. On the other hand, the *Hockey Fight Dataset* and *Action Moives Dataset*, which are the two most used datasets, do not contain scenes of violence in public settings, but are based on hockey matches and action movies where the focus and lighting of people is very good. This implies that the accuracies obtained by the state-of-the-art works are very high in those cases, but lower in datasets with real scenes, with security cameras in different angles and positions and variable lighting conditions such as the *RWF-2000* and *RLVS* datasets.

3. Objectives

This Section aims to establish the objectives of this work on the basis of the shortcomings analysed in the state of the art. As set out in Section 2, multiple approaches have been explored to address the existence of physical aggression in societies worldwide, with real-time detection of violence being the ultimate barrier to victim protection. The detection of violence using artificial intelligence is composed of multiple stages, with studies having been carried out in recent years in which single and multi-algorithm architectures of a very varied nature (CNN, RNN, Transformers...) have been implemented. Finally, the most widely used architecture in the recent state of the art, both individually and as a combination of algorithms, is the use of CNN and LSTM.

However, none of the articles reviewed that use CNN in combination with some form of RNN, as far as we are aware, test the improvement of the use of RNN over the use of CNN alone. In adittion, only one of the selected articles

contains a comparison between LSTM and Bi-LSTM [18]. While it is claimed that the use of Bi-LSTM improves over LSTM, the percentage improvement in the selected datasets does not exceed 4%. Furthermore, the analysis of the selected articles that use a combination of CNN with some form of RNN does not allow us to clearly extract architectures or hyperparameters that are clearly better than others. This is due to: the multiple steps that make up a violence detection algorithm, the few dataframes in common between papers to be able to compare their results and the use of dataframes that do not contain scenes of real aggressions, which results in obtaining accuracies that are too high to make evaluations between different architectures.

Based on these shortcomings in the state of the art, the following objectives are proposed for this work:

- Create a CNN-based violence detection model, and analyse its results with the same model combined with LSTM and Bi-LSTM layers. In this way, it will be possible to analyse the improvement that the combination of CNN and LSTM/Bi-LSTM brings, compared to the use of CNN alone.

- Creation of a violence detection model combining CNN and LSTM, as well as CNN and Bi-LSTM. This will allow the improvement of the use of Bi-LSTM over LSTM to be confirmed and quantified.

- Training of the models with a wide range of hyperparameters such as: Learning Rate and the number of neurons of the LSTM and Bi-LSTM layers as well as densely connected layers; this is expected to be able to analyse and observe whether a certain pattern of number of neurons obtains better results.

4. Proposed Violence Detection Algorithm Arquitecture

This Section consists of the explanation of the model architectures that have been decided to develop for this work, based on the Section 3 objectives. As discussed in Section 2, the use of pre-trained CNNs outperforms the use of untrained CNNs, where the VGG-16 CNN has been widely used. It has been decided to use pre-trained VGG-19 CNN, since its use is lower than VGG-16 and in the literature it obtains worse results, although in principle the use of a greater number of convolutional layers should provide a greater capacity to extract complex patterns. Like many other pretrained CNN-based algorithms

such as ResNet, MobileNet, or VGG-16, VGG-19 is trained on a large dataset of images such as ImageNet. These algorithms are thus used, as previously discussed, for spatial feature extraction rather than directly for frame-by-frame violence detection, which would be difficult to classify, as a frame within a violent scene might be considered non-violent without the temporal context it resides in. Nevertheless, this study will propose two methods to verify the effectiveness of VGG-19 compared to using VGG-19 in conjunction with RNN layers (specifically, LSTM and Bi-LSTM).:

- The first method consists of analyzing the number of frames that VGG-19 detects as violent and non-violent. It should be noted that, although the datasets used are efficiently trimmed, there may be a small number of frames in each video that may not be considered violent if the violent scene has not yet begun or has already ended.

- The second method consists of establishing a limit on the number of frames detected as violent, beyond which the video is considered violent, in order to jointly analyze a video after the prediction made by VGG-19 frame by frame. This method thus involves combining pre-trained VGG-19 with a manual method, to make it comparable to the other architectures in which the video is evaluated as violent or non-violent using LSTM layers and Bi-LSTM layers.

Regarding the proposed architectures based on the combined use of CNN and Bi-LSTM/LSTM layers, two architectures will be used: VGG-19 pre-trained together with LSTM layers and VGG-19 pretrained in conjunction with the use of Bi-LSTM layers.

Section 4.1 outlines the architecture of VGG-19. Section 4.2 will expose the arquitecture of the violence detection arquitecture using Pre-trained VGG-19 with minimum violent frame number. Section 4.3 will discuss the architecture created for violence detection using Pre-trained VGG-19 with Bi-LSTM/LSTM layers.

4.1. VGG-19 arquitecture

This Section will outline the architecture of VGG-19 [32] [41]. The difference between VGG-16 and VGG-19 is that VGG-16 has 13 convolutional layers, 5 pooling layers and 3 fully connected layers, while VGG-19 has 16 convolutional layers, 5 pooling layers and 3 fully connected layers. While a

larger number of convolutional layers means an increase in the number of parameters, which affects computational cost and training time, it also means a greater ability to understand complex structures and patterns.

In Figure 2 the structure of VGG-19 has been depicted. It can be seen how it expects to receive an array of shape (224,224,3), which indicates 224 pixels wide and high for each RGB colour channel. It then goes through a series of convolutional blocks, which are made up of consecutive convolutional layers; the convolutional layers use filters or kernels consisting of a series of weights that are applied on the array of pixels that is the image from which the network learns increasingly complex patterns. Between each convolutional block there are interspersed Max Poool Layers that reduce the dimension of the pixel matrix that is the image, selecting a maximum value from each region of the image. Once the image has passed through all the convolutional blocks, it is passed to three densely connected layers that act as a classifier. In Figure 2 it can be seen how the last fully connected layer generates an array with the form (1, 1, 1000), this is because when VGG-19 is trained on the image dataset *ImageNet*, it must classify between a total of 1000 classes. Once the output of the last layer is generated, the SoftMax function is applied to normalise the generated output, ensuring that the sum of probabilities is equal to 1.

Given the use of pre-trained VGG-19 in ImageNet for violence detection, it becomes necessary to alter the last fully connected layer, which originally contains 1000 neurons, to one with 2 neurons while retaining the weights of the pre-trained network. This adjustment is required due to violence detection being a binary problem (violence vs. non-violence).

4.2. Violence detection model using Pre-Trained VGG-19 with minimum violent frame number

As stated at the beginning of Section 4, in order to compare the accuracy of the prediction solely using pre-trained VGG-19, we will examine the number of frames in the test videos predicted as violent and non-violent. Since these results are analyzed frame by frame rather than the video as a whole, as done by the architectures combining VGG-19 with LSTM or VGG-19 with Bi-LSTM, presented in the upcoming sections, this section presents an architecture that predicts frame by frame whether the image is violent or not, and based on the number of frames predicted as violent, the entire video is considered violent or not. This architecture is illustrated in Figure 3.

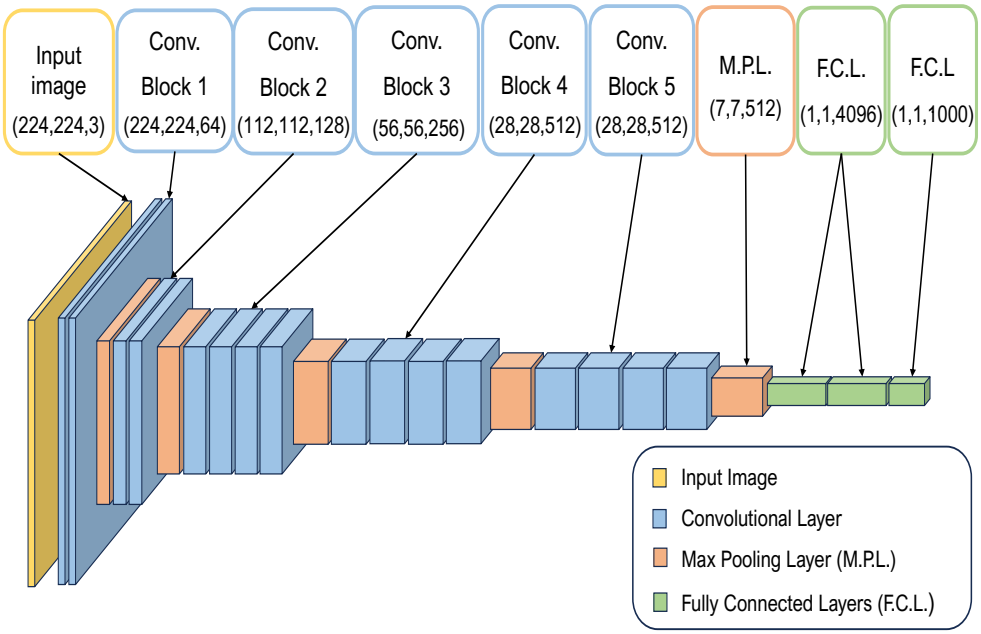


Figure 2: VGG-19 structure

Several values will be established as the minimum number of frames detected as violent to consider the video as violent, in order to observe how they vary for each dataframe. It is also worth noting that, to the best of our knowledge, in the recent state of the art, no violence detection model has been developed that combines a CNN with a Manual process, as proposed in this section. Thus, it represents an interesting model to investigate on its own.

4.3. Violence detection model arquitecture combining Pre-trained VGG-19 and Bi-LSTM/LSTM layers

This section outlines the proposed architecture for the combination of pre-trained VGG19 and Bi-LSTM/LSTM layers. This way of performing violence detection in combination using CNN and LSTM consists of extracting the spatial characteristics of the frames of a video on the one hand, and once all the spatial characteristics have been obtained, they are concatenated into a single element (an array) and introduced into the LSTM layers, which then

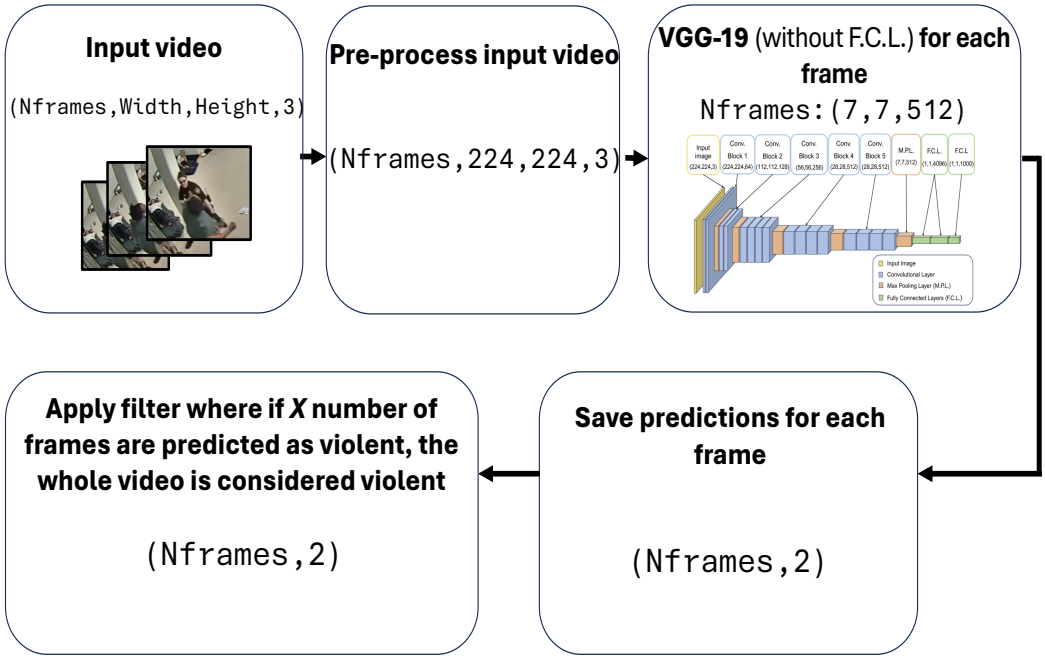


Figure 3: Violence detection model using Pre-Trained VGG-19 with minimum violent frame number

connect with the densely connected layers. This is represented in Figure **??**

First, the video is divided into N_{frames} and resized to the size (224, 224) which is supported by VGG-19. Then the frames are introduced to VGG-19 one by one. As the spatial features of each frame are calculated by VGG-19, they are grouped together and added to a single element as they leave the last convolutional layer in the format (7,7,512), resulting in a structure with the form: (N_{frames},7,7,512). Once the pooled features are grouped, a layer *Global Average Pooling 2D* is applied which reduces the dimensionality to the format (N_{frames}, 512) and is then introduced to two Bi-LSTM/LSTM layers and three densely connected layers with Sigmoid activation.

This architecture is processed in two distinct parts, rather than a single layered structure of consecutive neural networks. This means that if it is desired to train some layers of the pre-trained CNN to adjust it to the detection of violence, as is the case in this work, VGG-19 is trained on one side and the Bi-LSTM/LSTM layers together with the dense layers on the other. There is no evidence in the state of the art analysed whether this combination results

in lower accuracy because the CNN and (Bi-)LSTM are not trained together as a single architecture. The Bi-LSTM layered architecture, has been used in a conference paper presented as part of this line of research [34].

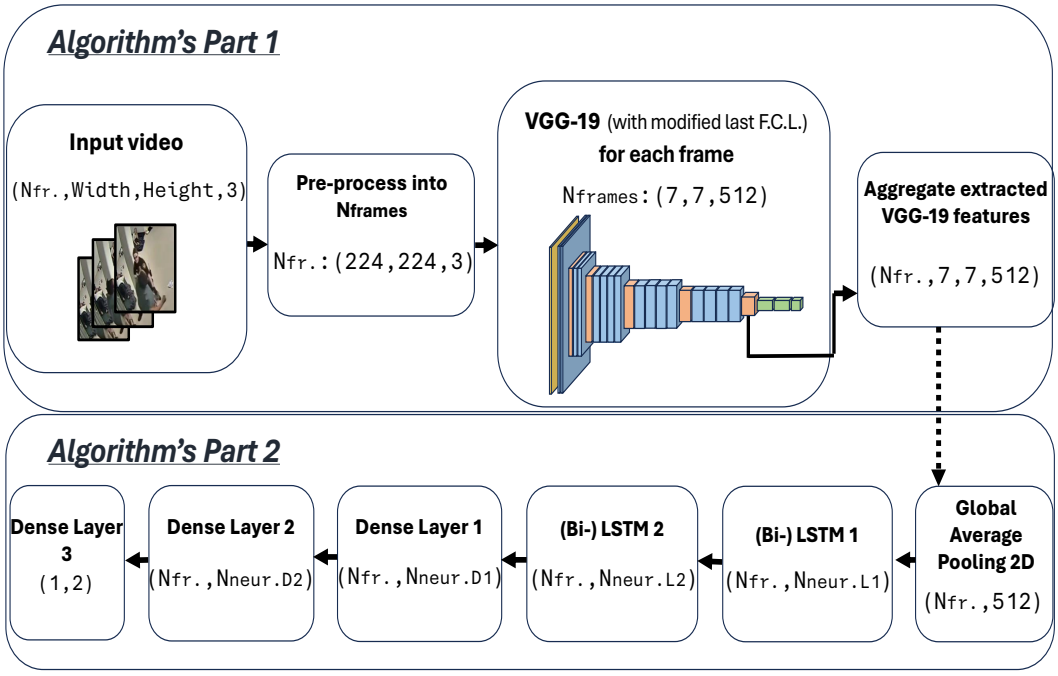


Figure 4: Violence detection arquitecture based on Spatial Features Concatenation

5. Results

This section is going to present the results obtained during the training and testing phase of the architecture presented in Section 4. All computations are done on a server running an Intel(R) Core(TM) i9-10940X CPU with 188 Gigabytes of RAM and an NVIDIA GeForce RTX 3090 GPU with 24 GigaBytes of memory.

5.1. Datasets

The datasets used to train and test the proposed models are explained in this Section. The most used datasets in the state of the art *Hockey Fights dataset* with 38 works using the dataset, the *Action movies dataset* with

27, the *Violent Flow dataset* with 24, the *Real World Fight 2000 dataset (RWF-2000)* [33].

It has been decided to use in this work: *Hockey Fights dataset*, *Violent Flow dataset* and *Real World Fight 2000 dataset (RWF-2000)*. This decision is due to the fact that, as discussed in Section 2.4, it is difficult in many cases to compare results of violence detection algorithms due to the use of different datasets for training and testing.

On one hand, *Hockey Fights dataset* [8] is chosen for being the most utilized dataset and contains action videos from field hockey games of the National Hockey League (NHL). Although the fights are real, the videos often contain close-ups, are well-lit, and do not depict a real-life situation where the assaulted person runs for help. Nevertheless, being able to compare results with a large number of articles is a compelling reason for its selection. Secondly, *Violent Flow dataset* [20] contains YouTube real-world crowded videos with audio. Lastly, the *Real World Fight 2000 dataset (RWF-2000)* contains scenes of violence recorded by security cameras. It is possibly the most comprehensive dataset due to its content and the quantity of 2000 videos.

It has been decided to skip the *Action movies dataset* in the list, which, although being the second most utilized dataset, contains scenes from action movies featuring close-ups and good lighting; ultimately, unreal scenes. With the chosen datasets, the aim is to compare the results obtained with many state-of-the-art articles and tackle three different types of violence: a well-lit and recorded environment such as a hockey game, crowded scenes, and real scenes recorded by security cameras.

Table 3 contains information on the selected datasets; all three datasets contain the same number of violent and non-violent videos. For all datasets, 70% of the videos will be used for training, 10% for validation and 20% for testing. *Violent Flow dataset* is the only one of the three selected dataframes that contains videos with different number of frames, therefore, those videos that have less frames than the median number of frames (107) will have the last frame inserted at the end of the video as many times as necessary to reach that value; while if the video is longer than the median duration, the frames at the end of the video will be trimmed. This is because the proposed architecture requires a constant duration of the videos.

Table 3: Selected datasets information

Name	**Year**	**Video Number**	**Median Frame Number**	**Frame rate (FPS)**
Hockey fights	2011	1000	50 frames	20-30
Violent Flow/Crowd	2012	246	107 frames	25
Real World Fight-2000 (RWF-2000)	2021	2000	150 frames	30

5.2. VGG-19 results

This Section is divided into two subsections. Section 5.2.1 will present the results obtained from training the last densely connected layer of pre-trained VGG-19 to classify frames from videos as violent or non-violent. Section 5.2.2 will present the results of the testing process of VGG-19, detecting frames from the testing datasets as violent or non-violent.

5.2.1. Pre-trained VGG-19 training

VGG-19 is imported with pre-trained weights, trained on ImageNet database, as shown in Figure 2. The last dense layer with 1000 neurons (since ImageNet has 1000 classes) is replaced by a layer with two neurons, so that VGG-19 switches to detecting violence and not violence (binary classification). To train VGG-19 with the train dataframes, it is decided to freeze all the layers except the last added dense layer with two neurons, so this will be the only layer to be trained. A Learning Rate of 10^{-3} and 50 epochs are used (this is considered sufficient for a single layer acting as a binary classifier). The optimiser used is *Adam* and the loss function *Binary cross entropy*.

To perform the training process all frames from the training videos will be resized to shape: (224, 224, 3), so each video has a size of (N_{frames}, 224, 224, 3). Since VGG-19 processes individual images (not videos), we put together (in an ordered way, to later be able to associate which video corresponds to which frame) in a single array all the frames of all the train videos of the dataset in question of size ($N_{train.videos} * N_{frames}$, 224, 224, 3). Lastly, an array of shape ($N_{train.videos} * N_{frames}$, 2) is created containing a value 0 in case the frames of the video they correspond to are non-violent, and 1 in case they are violent. It should be noted that although the videos in the violence datasets are normalised to a certain number of frames, to a certain height and width of pixels and cropped so that if the scene is violent, practically all

Table 4: Training VGG-19 results on selected violence datasets

Dataset	**Learning Rate**	**Epochs**	**Training Accuracy**	**Validation Accuracy**
Hockey Fights	0.001	50	1	0.99
RWF-2000			0.99	0.99
Violent Flow			1	0.99

the frames correspond to the violent act; it is possible that a few frames of some videos (at the beginning and/or at the end of the video) are frames that correspond before or after the violent action, adding images to the training process that are not within the violent action, but indicating that they are by belonging to a video labelled as violent.

Table 4 summarises the training information of VGG-19 with the three selected datasets, indicating in the columns *Training Accuracy* and *Validation Accuracy* the values obtained in the last epoch. In the case of the Hockey Fight dataset this accuracy and validation accuracy is obtained around epoch 5, in the RWF-2000 around 10 and in the Violent Flow around 3. This can be understood due to the directly proportional number of videos contained in the whole dataset. It also underlines the lack of the need to have trained with 50 epochs the last layer with two neurons that has been implemented in VGG-19.

5.2.2. Pre-trained VGG-19 testing

In this section, the pre-trained VGG-19 test process is explained. The test process consists of making predictions with VGG-19 using the 20% of the videos not used during the training process (neither as training, nor as validation), where there will be four possible situations: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FP). These four possibilities are the classic ones of any binary classification problem. True Positive indicates that the model predicts that the violent image is violent. True Negative indicates that the model predicts that the non-violent image is non-violent. False Positive indicates that the model predicts that the non-violent image is violent. False Negative indicates that the model predicts that the violent image is non-violent. The goal, of course, is to maximise both True Positive and True Negative, minimising False Positive and False Negative. A False Negative is worse than a False Positive, since if an aggression is occurring and the model does not detect it as such, the victim would not be

Table 5: Training Confusion Matrix for Trained Pre-trained VGG-19 on selected datasets

Dataset	Test Video Number	Frames per Video	Total Frame Number	TP	TN	FP	FN	TP (%)	TN (%)	FP (%)	FN (%)
Hockey Fight	800	40	32,000	20	23	2	5	40	46	4	10
RWF-2000	1600	150	240,000	143	149	51	57	35.75	37.25	12.75	14.25
Violent Flow	196	107	20,972	24	21	4	1	48	42	8	2

Table 6: Testing Pre-trained VGG-19 metrics on selected dataframes

Dataset	Test Video Number	Accuracy	Precision	Recall	F1 Score	Specificity	AUC
Hockey Fights	200	0.94	0.93	0.95	0.94	0.93	0.92
RWF-2000	400	0.64	0.65	0.63	0.64	0.66	0.71
Violent Flow	50	0.92	0.87	0.99	0.93	0.86	0.97

rescued.

Table 5 contains the confusion matrices for the datasets: Hockey Fights, RWF-2000 and Violent Flow; respectively. It can be observed that in the datasets Hockey Fights and Violent Flow very good results are obtained, being also lower the percentage of False Negatives (FN) than False Positives (FP); in the case of RWF-2000 there are 35.4% of false positives and false negatives, very different results from those obtained with the other two datasets.

Numerous metrics can be obtained from the confusion matrices, which are used to evaluate the results obtained depending on how the terms are grouped. The most widely used metric is the Accuracy [33], but there are also others such as: precision, recall, F1 Score and Specificity. Also widely used is the AUC metric, which is a measure of the discrimination capacity of a binary classification model, representing the probability that the model correctly classifies pairs of positive and negative observations by varying the decision thresholds. The results obtained for the three selected datasets are included in Table 6.

The results obtained are really promising for the Hockey Fights and Violent Flow datasets, taking into account that only VGG-19 is being used for frame-by-frame prediction, without taking into account the temporal relationship between them. With accuracy as the most used metric, Hockey Fights obtains 94%, RWF-2000 64% and Violent Flow 92%. RWF-2000 ob-

tains substantially lower results, either because of the difficulty of increasing the validation accuracy for such a large and diverse set of videos, or because of the content of the security camera videos themselves, such as: low luminosity in the case of night scenes, scenes far from the camera, variety in the type of scene (number of aggressors and form of aggression), etc.

5.3. VGG-19 with minimum violent frame number

This Section aims to test the effectiveness as a violence detection model of the architecture shown in Figure 3. In addition to being a method that combines CNN and a manual feature, which is an architecture that has not been used in the recent state of the art [33], it is of great interest to test the effectiveness of VGG-19 on its own against architectures in combination with (Bi-)LSTM layers. Although it was intended to use the set of validation videos used during the training of VGG-19 to observe which values for each dataset obtained the best results as the minimum number of frames detected as violent to consider the video violent, and then use those values with the test set, this will not be possible given that the validation accuracy obtained by VGG-19 gives a 99% accuracy for the three datasets. These results assume that the same results are obtained for any minimum number of frames, so the values will be applied directly to the test set; it follows that in a real case, this value should be adjusted manually as more experience is gained with the detected scenes.

The minimum frame values and the percentage of frames it represents with respect to the total number of frames of the videos of each dataframe are represented in Table 7. As can be seen, the values of 10, 20, 40, 80 and 120 frames have been chosen. Since the Hockey Fights videos contain 40 frames, these values will not be checked; as well as the value of 120 frames for Violent Flow dataset, since its processed videos contain 107 frames.

Table 8 contains the results for the minimum frame number with values 10, 20 and 40 on the Hockey Fights dataset; showing the accuracy and confusion matrices in absolute value for each case. It can be seen in green that the best accuracy obtained is with a low minimum frame number such as 10 frames.

Table 9 contains the results for the minimum frame number with values 10, 20, 40, 80 and 120 on the RWF-2000 dataset; showing the accuracy and confusion matrices in absolute value for each case. It can be seen in green that the best accuracy obtained is with a low minimum frame number such as 10 frames, as with the Hockey Fights dataset.

Table 7: Minimum frame number to consider a video as violent and the percentage of frames it represents with respect to the total number of frames of the videos of each dataframe

Minimum frame number to consider a video as violent	**Hockey Fights (%)**	**Violent Flow (%)**	**RWF-2000 (%)**
10	25	9.3	6.7
20	50	18.7	13.3
40	100	37.4	26.7
80	-	74.8	53.3
120	-	-	80.0

Table 8: Hockey Fights dataset test results of VGG19 with multiple violent frame number confusion matrix and accuracy

Minimum Frame Number	**Accuracy**	**TP**	**TN**	**FP**	**FN**	**TP%**	**TN%**	**FP%**	**FN%**
10	0.95	98	91	9	2	49.0	45.5	4.5	1.0
20	0.94	97	91	9	3	48.5	45.5	4.5	1.5
40	0.84	70	97	3	30	35.0	48.5	1.5	15.0

Table 9: RWF-2000 dataset test results of VGG19 with multiple violent frame number confusion matrix and accuracy

Minimum Frame Number	**Accuracy**	**TP**	**TN**	**FP**	**FN**	**TP%**	**TN%**	**FP%**	**FN%**
10	0.76	177	90	62	23	50.3	25.6	17.6	6.5
20	0.74	167	95	57	33	47.4	27.0	16.2	9.4
40	0.73	153	103	49	47	43.5	29.3	13.9	13.4
80	0.63	104	118	34	96	29.5	33.5	9.7	27.3
120	0.50	25	152	0	175	7.1	43.2	0.0	49.7

Table 10: Violent Flow dataset test results of VGG19 with multiple violent frame number confusion matrix and accuracy

Minimum Frame Number	Accuracy	TP	TN	FP	FN	TP%	TN%	FP%	FN%
10	0.84	25	17	8	0	50.0	34.0	16.0	0.0
20	0.88	25	19	6	0	50.0	38.0	12.0	0.0
40	0.94	25	22	3	0	50.0	44.0	6.0	0.0
80	0.96	25	23	2	0	50.0	46.0	4.0	0.0

Table 11: Metrics for the best VGG19 with Minimum frame number model for each of the selected datasets

Dataset	Minimum Frame Number	Accuracy	Precision	Recall	F1 Score	Specificity
Hockey Fights	10	0.95	0.92	0.98	0.95	0.91
RWF-2000	10	0.76	0.74	0.89	0.89	0.6
Violent Flow	80	0.96	0.93	1	0.96	0.92

Table 10 contains the results for the minimum frame number with values 10, 20, 40, 80 on the Violent Flow dataset; showing the accuracy and confusion matrices in absolute value for each case. It can be seen in green that the best accuracy obtained is with a low minimum frame number such as 80 frames, the maximum value of those selected with respect to the total number of frames per video for this dataset; contrary to the Hockey Fight and RWF-2000 videos.

Finally, the most used metrics in binary classification with the best accuracy of the minimum frame number obtained for each of the three selected dataframes are represented in Table 11.

Given the results obtained with the Violent Flow dataset, it cannot be stated with certainty that a lower number of frames for the classification of a violent video is better when deciding with the frame-by-frame analysis by VGG19. It can be stated that the accuracy results are very good in the Hockey Fights dataset and the Violent Flow dataset.

An important innovation of this study lies in achieving high accuracy solely by analyzing individual frames using the pre-trained VGG-19 model. This approach, which disregards temporal relationships between frames, yielded a competitive accuracy of 95% on the Hockey Fights dataset and 96% on the Violent Flow dataset. These results suggest that violence detection in video

may not always require complex temporal layers, such as LSTM or Bi-LSTM, particularly in settings with sufficient contextual visual cues within individual frames. This finding contributes to the state of the art by proposing that simpler frame-by-frame methods can achieve effective violence detection in certain scenarios.

5.4. Pre-Trained VGG-19 + LSTM results

In this Section it's exposed the results obtained during the training and testing phases of the architecture that combines the use of pre-trained VGG-19 and LSTM layers, as depicted in Figure 4. Section 5.4.1 will showcase the training phase, and Section 5.4.2 will cover the testing phase.

5.4.1. LSTM and Dense Layers training

At this moment only VGG-19 has been trained. The next step is to train part 2 of the architecture shown in Figure 4, which consists of 2 LSTM layers and 3 fully connected layers. For this, it is necessary to store the spatial features generated by trained VGG-19, obtaining an array with the form: $(N_{videos} * N_{frames}, 7, 7, 512)$. As exposed in Section 5.2.1, for the training of VGG-19 we put in the same array the succession of frames of all the training videos; however, now we want to recover the sense of which frame belongs to which video, so that the LSTM can extract temporal patterns from each one of them. To do so, we modify the array of extracted spatial features to obtain: $(N_{videos}, N_{frames}, 7, 7, 512)$, which is the output obtained after the fifth block of convolutional layers and after passing through a Max Pool Layer.

Finally, LSTMs expect to receive as input an array of the form ($BatchSize$, $TimeSteps$, $Features$), where *Batch Size* is the number of videos, *Time Steps* the number of frames of the videos and *Features* the features we want the LSTM to analyse; therefore we apply a layer *Global Average Pooling 2D* to obtain an array with the form $(N_{videos}, N_{frames}, 512)$ as input to the first LSTM layer. As training data an array with the shape $(N_{videos}, 2)$ must be generated where for each video if the array has shape [1, 0] it means that it is not violent and if it is [0, 1] it does contain violence. During the training, several hyperparameter options will be implemented. In this case multiple values are testes: Learning Rate, neurons of each of the LSTM layers and neurons of each of the fully connected layers.

Typical values used in the state of the art compiled in Table 1 values of: Learning rate of 10^{-2}, 10^{-3}, 10^{-4}, LSTM and fully connected layers neuron number of: 64, 128, 256, 512 and 1024. It is allowed in the combination

search that the first and second layer of LSTM and fully connected layers have different numbers of neurons, as it is desired to test how increasing, decreasing or keeping the same number of neurons affects the results obtained. The three datasets will be trained for a total of 150 epochs. This is three times higher than the training value of VGG-19 given the number of layers to be trained. As with VGG-19 The optimiser used is *Adam* and the loss function *Binary cross entropy*. Having such a high search range means that the number of possible combinations is very high. Keras-Tuner is used to manage this. A total of 200 possible combinations are implemented, which represents about 10% of the total of 1875 possible combinations.

Given that during training the goal is to maximise the Validation Accuracy metric, the three best models obtained by Keras-Tuner in its search for the best combination of hyperparameters are stored; this is because among models that have obtained a similar Validation Accuracy, not necessarily the best of them is the one that will best fit the test data. It should be noted that with the results obtained in training the LSTM and Bi-LSTM layers together with their densely connected layers, no clear pattern could be extracted from the number of neurons in the LSTM and Bi-LSTM layers, from the number of densely connected layers, nor from the relationship between the number of neurons in the layers and each other.

Table 12 contains the hyperparameters information of the 3 best models that extract the temporal features from the spatial features extracted by VGG-19 from the Max Pooling Layer after its fifth and last block of convolutional layers; by means of 2 LSTM layers and 3 densely connected layers, being the last one composed by two neurons as it is a binary classification. The column *Model* indicates as 1 the best model indicated by Keras-Tuner, with 2 the second best model and with 3 the third best model.

5.4.2. LSTM and Dense Layers testing

This section presents the results obtained during the testing of the 3 best models obtained for the architecture that, from the spatial features extracted by the pre-trained VGG-19 trained network, uses LSTM layers for the extraction of spatial features and densely connected layers for the classification between violence and non-violence. Since the Accuracy metric is the most widely used metric in the state of the art [33], it will be the one on which the choice of the best model among the three best ones stored by Keras-Tuner for each of the three selected datasets will be based. For each architecture, three tables will be presented: the first one showing the accuracys obtained

Table 12: Training results of the best three models with LSTM layers hyperparameters combinations for the three selected datasets

Dataset	Model	Validation Accuracy	Learning Rate	LSTM 1	LSTM 2	F.C.L 1	F.C.L 2
Hockey Fights	**1**	0.96	0.0001	256	128	64	64
	2	0.96	0.01	512	512	128	64
	3	0.96	0.001	64	1024	512	64
RWF-2000	**1**	0.87	0.001	1024	1024	256	512
	2	0.87	0.0001	512	64	128	128
	3	0.87	0.001	128	512	128	256
Violent Flow	**1**	0.96	0.01	64	512	64	128
	2	0.96	0.01	128	64	128	64
	3	0.96	0.0001	512	512	128	1024

Table 13: Testing results of the best three models with LSTM layers hyperparameters combinations for the three selected datasets

Dataset	Model	Test Accuracy	Learning Rate	LSTM 1	LSTM 2	F.C.L 1	F.C.L 2
Hockey Fights	**1**	0.94	0.0001	256	128	64	64
	2	0.96	0.01	512	512	128	64
	3	0.96	0.001	64	1024	512	64
RWF-2000	**1**	0.72	0.001	1024	1024	256	512
	2	0.71	0.0001	512	64	128	128
	3	0.71	0.001	128	512	128	256
Violent Flow	**1**	0.86	0.01	64	512	64	128
	2	0.86	0.01	128	64	128	64
	3	0.86	0.0001	512	512	128	1024

by the three models and their combination of hyperparameters, a second table showing the cofusion matrix of the best model chosen, and a third table containing the metrics of the best model for each selected dataset.

Table 13 contains the test results of the architecture where LSTM layers are used. In case there are two models with equal testing accuracy, the model with the highest selected by Keras-Tuner is selected. In the case of the Hockey Fights dataset, the best model is the second one, with a testing accuracy of 96%. For the RWF-2000 dataset, the best model is the second one, with a testing accuracy of 72%. Finally, in the case of the Violent Flow Dataset, the best model is the first model, with a testing accuracy of 86%.

Clearly, the best results are obtained with the Hockey Fights Dataset, then with the Violent Flow and finally with the RWF-2000. This descending order is in accordance with the length of the videos, as well as the diversity of video content and the quality and focus of the scene in the videos. The

Table 14: Confusion Matrix of the best model with LSTM layers hyperparameters combination for the three selected datasets

Dataset	Model	Test Video Number	TP	TN	FP	FN	TP (%)	TN (%)	FP (%)	FN (%)
Hockey Fight	**2**	200	96	96	4	4	48	48	2	2
RWF-2000	**1**	400	154	135	65	46	38.5	33.75	65	11.5
Violent Flow	**1**	50	20	23	2	5	40	46	4	10

Table 15: Testing metrics of the best model with LSTM layers hyperparameters combination for the three selected datasets

Dataset	Test Video Number	Model Number	Accuracy	Precision	Recall	F1 Score	Specificity	AUC
Hockey Fight	200	2	0.96	0.96	0.96	0.96	0.96	0.97
RWF-2000	400	1	0.72	0.70	0.77	0.74	0.68	0.79
Violent Flow	50	1	0.86	0.85	0.88	0.86	0.84	0.93

RWF-2000 has a noticeably higher number of videos than the number of videos, which can also make a convergence towards a high accuracy difficult.

We now turn to the confusion matrices for the best models shown in Table 14.

The metrics mostly calculated from the confusion matrices shown are also presented. Table 15 contains the evaluation metrics of the best model with LSTM layers for each of the three selected datasets.

5.5. Pre-Trained VGG-19 + Bi-LSTM results

This Section contains the results obtained during the training and testing of the violence detection model based on pre-trained VGG-19 combined with Bi-LSTM layers.

5.5.1. Bi-LSTM and Dense Layers training

The training is performed in exactly the same way as explained in the Section 5.4.1 but with the pre-trained VGG-19 architecture combined with Bi-LSTM layers. Table 16 contains the hyperparameter information of the 3 best models, this time using 2 Bi-LSTM layers instead of 2 LSTM layers. These results were obtained in our congress paper which is part of this line of research [34]. It can be seen that the three best stored models obtain the same Validation accuracy during the test process.

Table 16: Training results of the best three models with Bi-LSTM layers hyperparameters combinations for the three selected datasets [34]

Dataset	Model	Validation Accuracy	Learning Rate	Bi-LSTM 1	Bi-LSTM 2	F.C.L 1	F.C.L 2
Hockey Fights	**1**	0.97	0.01	64	64	64	512
	2	0.97	0.01	512	64	1024	512
	3	0.97	0.01	64	256	64	256
RWF-2000	**1**	0.87	0.01	512	64	256	64
	2	0.87	0.001	512	256	64	256
	3	0.87	0.001	128	128	64	128
Violent Flow	**1**	0.96	0.001	1024	512	64	256
	2	0.96	0.01	256	256	256	128
	3	0.96	0.01	128	512	128	256

Table 17: Testing results of the best three models with Bi-LSTM layers hyperparameters combinations for the three selected datasets

Dataset	Model	Test Accuracy	Learning Rate	Bi-LSTM 1	Bi-LSTM 2	F.C.L 1	F.C.L 2
Hockey Fights	**1**	0.93	0.01	64	64	64	512
	2	0.95	0.01	512	64	1024	512
	3	0.97	0.01	64	256	64	256
RWF-2000	**1**	0.73	0.001	1024	1024	256	512
	2	0.73	0.0001	512	64	128	128
	3	0.68	0.001	128	512	128	256
Violent Flow	**1**	0.90	0.01	64	512	64	128
	2	0.82	0.01	128	64	128	64
	3	0.88	0.0001	512	512	128	1024

5.5.2. Bi-LSTM and Dense Layers testing

The testing is performed in exactly the same way as explained in the Section 5.4.2 but with the pre-trained VGG-19 architecture combined with Bi-LSTM layers. The results are shown in Table 17, where it can be seen that for the Hockey Fights dataset the third model obtains the best results, unlike for RWF-2000 and Violent Flow which is the first model suggested by Keras-Tuner. Tables 18 and 19 contain the confusion matrix and metrics for the three best models with the VGG-19 architecture combined with Bi-LSTM layers.

5.6. Hyperparameter result analysis for LSTM and Bi-LSTM combinations

As exposed in Sections 5.4.1 and 5.5.1, models based on the combination of pre-trained VGG-19 with LSTM layers, and Bi-LSTM layers have been trained. In both cases, Keras-Tuner was used to obtain the best model combination out of 200 possible hyperparameter combinations. The aim was

Table 18: Confusion Matrix of the best model with Bi-LSTM layers hyperparameters combination for the three selected datasets

Dataset	Model	Test Video Number	TP	TN	FP	FN	TP (%)	TN (%)	FP (%)	FN (%)
Hockey Fight	**3**	200	20	23	2	5	40.00	46.00	4.00	10.00
RWF-2000	**1**	400	143	149	51	57	35.75	37.25	12.75	14.25
Violent Flow	**1**	50	24	21	4	1	48.00	42.00	8.00	2.00

Table 19: Testing metrics of the best model with Bi-LSTM layers hyperparameters combination for the three selected datasets

Dataset	Test Video Number	Model Number	Accuracy	Precision	Recall	F1 Score	Specificity	AUC
Hockey Fight	200	3	0.97	0.95	0.98	0.97	0.95	0.98
RWF-2000	400	1	0.73	0.74	0.72	0.73	0.75	0.79
Violent Flow	50	1	0.9	0.86	0.97	0.91	0.84	0.96

to observe if certain hyperparameter values resulted in a notable improvement in violence detection results.

For each model separately, the results of the hyperparameter combinations obtained for the three selected datasets were combined, the mean of the obtained validation accuracy results was calculated, and it was determined if any of the hyperparameter values showed a significant deviation from the rest of the values. However, no value stood out notably to affirm that its use clearly resulted in better accuracy. Furthermore, for each model, an analysis was conducted to determine if increasing, keeping the same number, or decreasing the number of neurons between the LSTM or Bi-LSTM layers and the dense layers resulted in a significant improvement in the results. However, none of the options stood out notably. Specific analysis of values for the first and second LSTM layers and dense layers was not performed, as even though approximately 10% of the total possible combinations were covered, there were not enough tested combinations to yield reliable results.

In conclusion, significant finding is the lack of a consistent improvement in accuracy across different configurations of hyperparameters, such as neuron count in LSTM/Bi-LSTM layers and dense layers. Despite testing various combinations, no clear pattern emerged that consistently enhanced performance. This observation highlights that extensive hyperparameter tuning

does not always translate to better accuracy, potentially simplifying future model optimization efforts for violence detection. This insight contributes a nuanced understanding to the state of the art, suggesting that high performance may be achievable with a streamlined approach to parameter selection.

5.7. Comparison of results between models

This Section aims to analyse the results obtained by the models developed in this paper with each other and with other state-of-the-art articles. Table 20 shows the results obtained using the datasets selected by the VGG-19 pre-trained network detecting violence from the videos frame by frame (Section 5.2), combined with a manual logic of minimum frames needed to consider the video violent (Section 5.3), combined with LSTM layers (Section 5.4) and combined with Bi-LSTM layers (Section 5.5). As mentioned above, the comparison of VGG-19 pre-training prediction results on a frame-by-frame basis must be done with care, as it is done on an image basis and not on a video basis. However, in the case of the Hockey Fights and Violent flow datasets, excellent results are obtained compared to those combined with LSTM and Bi-LSTM layers.

Regarding the model that together with VGG-19 pre-training uses a manual logic (although the minimum number of frames detected as violent to consider the video violent has been established directly on the training dataset) obtains better results than analysing the frames individually, in addition to having a prediction per video and not per frame. Surprisingly, this method obtains better results in the RWF-2000 dataset and Violent Flow dataset than the combinations with LSTM and Bi-LSTM layers, even exceeding them by 6% in the case of Violent Flow. All this raises the premise of the possibility of detecting violence only from images and not by analysing the temporal relationship between them, even though we know that an action intrinsically occurs over time and not in a punctual way, as could be the detection of an object.

On the other hand, it is confirmed that the structures using Bi-LSTM obtain better results than those using LSTM, however, it should be noted that this improvement is low. In the case of the Hockey Fights dataset and RWF-2000 it is only 1% accuracy, while in Violent Flow it is 4% (more remarkable). This modest enhancement suggests that while bidirectional architectures may add value in capturing temporal patterns, the advantages may not justify their added complexity in all datasets. This insight aligns with the study's objective of assessing the necessity of advanced RNN layers,

Table 20: Accuracy comparison of results between proposed models

	Pre-trained VGG-19			
	frames by frame	**+ Manual Feature**	**+ LSTM**	**+ Bi-LSTM**
Hockey Fights	94	95	96	97
RWF-2000	64	76	72	73
Violent Flow	92	96	86	90

thereby contributing to the ongoing debate regarding the cost-benefit ratio of bidirectional architectures in video-based violence detection.

Tables 21, 22 and 23 contain the accuracies of the models proposed in this paper and of the papers shown in Table 2, which are those recent papers that have used the combination of CNN and LSTM for video violence detection.

Using Hockey Fights dataset, all proposed models have outperformed the accuracy obtained in the work of Mumtaz et al. [32] which also uses pre-training VGG-19. Some of them outperform results from other works, but they do not outperform the accuracy of those using pre-trained VGG-16. With respect to the results using Violent Flow dataset, Mumtaz et. al obtains almost the same result for the combination of VGG-19 and Bi-LSTM. Our model of VGG-19 and Manual feature obtains better results in addition to one of the modes using pre-trained VGG-16. Finally, in the case of RWF-2000, none of the models achieves results close to those of the three papers that also use this dataset. It is by far the dataset with which the models proposed in this paper have the most difficulties.

6. Conclusions and Future Work

Physical assaults represent a notable concern in our society, affecting individuals worldwide and directly influencing victims and their psychological well-being [30] [27]. The identification of violence in real-time videos serves as the ultimate barrier in protecting victims, where artificial intelligence has demonstrated excellent results in this task [33].

Firstly, a table listing articles using a combination of CNN and LSTM for violence detection in the recent state of the art has been compiled and analyzed. From this table, several deficiencies in the state-of-the-art have been identified, which this work aims to address. Firstly, there is a lack of understanding about the real contribution of using RNN combined with

Table 21: Hockey Fights dataset state of art and proposed models accuracy

Cite	**CNN**	**Combination**	**Hockey Fights**
[48]	MobileNet V2	LSTM	99.5
[19]	Own CNN	Bi-LSTM	99.27
[31]	PT VGG-16 (ImageNet)	Bi-Conv-LSTM	99.1
[1]	PT VGG-16 (ImageNet)	LSTM	99.1
[23]	PT MNAS CNN	Conv-LSTM	99
[45]	Two PT EfficienNet-B0 (ImageNet)	Bi-LSTM	99
[18]	PT VGG-16 (ImageNet)	LSTM/Bi-LSTM	97.6/98.8
[7]	Two PT VGG-16 (ImageNet) + Wide Dense Residual Blocks (WDRB)	LSTM	98.8
[47]	Darknet + Residual Optical Flow	M-LSTM	98
[44]	PT VGG16 (INRA)	Bi-GRU	98
Proposed model	PT VGG19 (ImageNet)	Bi-LSTM	97
[22]	Own CNN	LSTM	97
[38]	PT Xception (ImageNet)	LSTM	96.55
Proposed model	PT VGG19 (ImageNet)	LSTM	96
Proposed model	PT VGG19 (ImageNet)	Manual Feature	95
[43]	CNN	LSTM	94.9
Proposed model	PT VGG19 (ImageNet) (frame by frame)	N	94
[32]	PT VGG-19 (ImageNet) + extra CNN layers	Bi-LSTM	91.29

Table 22: Violent Flow dataset state of art and proposed models accuracy

Cite	**CNN**	**LSTM**	**Violent Flow**
[18]	PT VGG-16 (ImageNet)	LSTM/Bi-LSTM	92.2/95.1
[23]	PT MNAS CNN	Conv-LSTM	100
[19]	Own CNN	Bi-LSTM	98.64
[31]	PT VGG-16 (ImageNet)	Bi-Conv-LSTM	98.4
[47]	Darknet + Residual Optical Flow	M-LSTM	98.21
[7]	Two PT VGG-16 (ImageNet) + Wide Dense Residual Blocks (WDRB)	LSTM	97.1
Propos model	PT VGG19 (ImageNet)	Manual Feature	96
[44]	PT VGG16 (INRA)	Bi-GRU	95.5
[45]	Two PT EfficienNet-B0 (ImageNet)	Bi-LSTM	93.75
Propos model	PT VGG19 (ImageNet) (frame by frame)	N	92
[32]	PT VGG-19 (ImageNet) + extra CNN layers	Bi-LSTM	90.47
Propos model	PT VGG19 (ImageNet)	Bi-LSTM	90
Propos model	PT VGG19 (ImageNet)	LSTM	86
[43]	CNN	LSTM	77.31

Table 23: RWF-2000 dataset state of art and proposed models accuracy

Cite	**CNN**	**LSTM**	**RWF-2000**
[31]	PT VGG-16 (ImageNet)	Bi-Conv-LSTM	92.4
[32]	PT VGG-19 (ImageNet) + extra CNN layers	Bi-LSTM	90.47
[48]	MobileNet V2	LSTM	82
Propos model	PT VGG19 (ImageNet)	Manual Feature	76
Propos model	PT VGG19 (ImageNet)	Bi-LSTM	73
Propos model	PT VGG19 (ImageNet)	LSTM	72
Propos model	PT VGG19 (ImageNet) (frame by frame)	N	64

CNN, compared to using only CNN. Secondly, the actual improvement of using LSTM over Bi-LSTM remains unclear, with only one recent study having made this comparison, showing a modest 4% improvement. Thirdly, there is no clear relationship between hyperparameter values or combination of these values and an increase in model accuracy. Fourthly, while it has been demonstrated that VGG-16 achieves excellent results in violence detection, the few recent studies using VGG-19 have not achieved better results, although theoretically a greater number of convolutional layers should lead to a better understanding of the scene in question.

Therefore, in this work, several models based on the well-known VGG-19 network with pre-trained weights from the ImageNet dataset have been developed. VGG-19 is used to extract spatial features frame by frame from violence dataset videos. The first model involves only a frame-by-frame analysis of detections made with pre-trained VGG-19, although it is assumed that temporal relationships between frames are lost, providing information on the effectiveness of action analysis through images. The second model incorporates VGG-19 with manual logic implementation, such that a video is not considered violent unless a certain number of frames are detected as violent by pre-trained VGG-19. The third and fourth models involve the combination of pre-trained VGG-19 with the use of LSTM and Bi-LSTM layers, respectively.

Upon comparing the results of the architecture based on pre-trained VGG-19 with manual feature use to architectures based on pre-trained VGG-19 with LSTM and Bi-LSTM layers, it has been observed that CNNs play a significant role in prediction compared to the improvement brought about by the use of LSTM and Bi-LSTM layers. This raises the question of how necessary the temporal relationship between frames is for violence detection in video, although it is clear that an action occurs over time and is not instantaneous. It also raises the question of whether combining CNNs and LSTMs in parallel rather than concatenated would yield better results, as the input to LSTM and Bi-LSTM layers would be the original video rather than the spatial features extracted by pre-trained VGG-19.

As results, it has been observed that even when establishing a wide range of hyperparameter values with learning rate and number of neurons in LSTM/Bi-LSTM layers and dense layers, there are no specific values that clearly increase the accuracy obtained. Neither does the increase, decrease, or same number of neurons notably affect the accuracy obtained.

The violence detection results obtained for the Hockey fights and Violent Flow dataset are promising, while those for the RWF-2000 dataset are less satisfactory due to its more varied and complex scenes. VGG-19 demonstrates strong frame-by-frame predictions with 94% and 92% accuracy for the Hockey fights and Violent Flow dataset, respectively. Predictions using a minimum number of detected violent frames surpass those combined with LSTM and Bi-LSTM on the RWF-2000 and Violent Flow dataset, achieving 76% and 96% accuracy, respectively, and are comparable within a range of 1-2% in the Hockey fights dataset with 95%. Furthermore, the use of pre-trained VGG-19 with LSTM and Bi-LSTM layers achieves 96% and 97% accuracy, respectively, for the Hockey fights dataset, 86% and 90% for the Violent Flow dataset, and 72% and 73% for the RWF-2000 dataset, indicating a marginal 1% improvement with the use of Bi-LSTM in two out of the three datasets.

The proposed models improve the results obtained on the Hockey fights and Violent Flow datasets compared to one of the recent state-of-the-art studies using pre-trained VGG-19. Moreover, they outperform a study employing pre-trained VGG-16 on the Violent Flow dataset. Notably, despite VGG-19 containing more convolutional layers, which theoretically should enhance feature extraction capabilities, the simpler VGG-16 model has proven to be more effective in violence detection tasks in videos. This suggests that increased model complexity does not always translate into better performance. The

findings encourage further investigation into model selection, proposing that less complex architectures may yield better results under certain conditions, thereby challenging some assumptions in the current literature. Additionally, the results obtained with the proposed models surpass those of other state-of-the-art works employing different CNN and RNN layers.

As future work, it is proposed to apply trustworthy artificial intelligence [15] models to the models developed, as these are models that do not provide information on the decisions made. In particular, we propose the application of explainability models such as GradCAM, designed specifically for CNN, which applies a heat map to an image, allowing us to understand which part of the image is relevant for decision making.

Acknowledgements

This research is part of the International Chair Project on Trustworthy Artificial Intelligence and Demographic Challenge within the National Strategy for Artificial Intelligence (ENIA), in the framework of the European Recovery, Transformation and Resilience Plan. Referencia: TSI-100933-2023-0001. This project is funded by the Secretary of State for Digitalization and Artificial Intelligence and by the European Union(Next Generation).

References

[1] Aarthy, K., Nithya, A.A., 2022. Crowd violence detection in videos using deep learning architecture, in: 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), IEEE. pp. 1–6.

[2] Afra, S., Alhajj, R., 2020. Early warning system: From face recognition by surveillance cameras to social media analysis to detecting suspicious people. Physica A: Statistical Mechanics and its Applications 540, 123151.

[3] Ageed, Z., Zeebaree, S., 2021. A comprehensive survey of big data mining approaches in cloud systems 1, 29–38.

[4] Ahmed, M., Ramzan, M., Khan, H.U., Iqbal, S., Khan, M.A., Choi, J.I., Nam, Y., Kadry, S., 2021. Real-time violent action recognition using key frames extraction and deep learning .

[5] Aktı, Ş., Ofli, F., Imran, M., Ekenel, H.K., 2022. Fight detection from still images in the wild, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 550–559.

[6] Alonso, R.S., Sittón-Candanedo, I., Casado-Vara, R., Prieto, J., Corchado, J.M., 2020. Deep reinforcement learning for the management of software-defined networks and network function virtualization in an edge-iot architecture. Sustainability 12, 5706.

[7] Asad, M., Yang, J., He, J., Shamsolmoali, P., He, X., 2021. Multi-frame feature-fusion-based model for violence detection. The Visual Computer 37, 1415–1431.

[8] Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R., 2011. Violence detection in video using computer vision techniques, in: Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14, Springer. pp. 332–339.

[9] Cheng, M., Cai, K., Li, M., 2021. Rwf-2000: An open large scale video database for violence detection, in: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 4183–4190.

[10] Collins, C., Dennehy, D., Conboy, K., Mikalef, P., 2021. Artificial intelligence in information systems research: A systematic literature review and research agenda. International Journal of Information Management 60, 102383.

[11] Contardo, P., Tomassini, S., Falcionelli, N., Dragoni, A.F., Sernani, P., 2023. Combining a mobile deep neural network and a recurrent layer for violence detection in videos .

[12] Ding, D., Ma, Z., Chen, D., Chen, Q., Liu, Z., Zhu, F., 2021. Advances in video compression system using deep neural network: A review and case studies. Proceedings of the IEEE 109, 1494–1520.

[13] Ehsan, T.Z., Nahvi, M., Mohtavipour, S.M., 2023. An accurate violence detection framework using unsupervised spatial–temporal action translation network. The Visual Computer , 1–21.

[14] Enaifoghe, A., Dlelana, M., Durokifa, A.A., Dlamini, N.P., 2021. The prevalence of gender-based violence against women in south africa: A call for action. African Journal of Gender, Society & Development 10, 117.

[15] European Commission, High-Level Expert Group on AI, 2019. Ethics guidelines for trustworthy ai. https://ec.europa.eu/digital-strategy/news-redirect/65479. Accessed: 2024-10-30.

[16] Fekih-Romdhane, F., Malaeb, D., Sarray El Dine, A., Obeid, S., Hallit, S., 2022. The relationship between smartphone addiction and aggression among lebanese adolescents: the indirect effect of cognitive function. BMC pediatrics 22, 735.

[17] Freire-Obregón, D., Barra, P., Castrillón-Santana, M., Marsico, M.D., 2022. Inflated 3d convnet context analysis for violence detection. Machine Vision and Applications 33, 1–13.

[18] Gupta, H., Ali, S.T., 2022. Violence detection using deep learning techniques, in: 2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI), IEEE. pp. 121–124.

[19] Halder, R., Chatterjee, R., 2020. Cnn-bilstm model for violence detection in smart surveillance. SN Computer science 1, 201.

[20] Hassner, T., Itcher, Y., Kliper-Gross, O., 2012. Violent flows: Real-time detection of violent crowd behavior, in: 2012 IEEE computer society conference on computer vision and pattern recognition workshops, IEEE. pp. 1–6.

[21] Hillis, S., Mercy, J., Amobi, A., Kress, H., 2016. Global prevalence of past-year violence against children: a systematic review and minimum estimates. Pediatrics 137.

[22] Islam, M.S., Hasan, M.M., Abdullah, S., Akbar, J.U.M., Arafat, N., Murad, S.A., 2021. A deep spatio-temporal network for vision-based sexual harassment detection, in: 2021 Emerging Technology in Computing, Communication and Electronics (ETCCE), IEEE. pp. 1–6.

[23] Jahlan, H.M.B., Elrefaei, L.A., 2021. Mobile neural architecture search network and convolutional long short-term memory-based deep features toward detecting violence from video. Arabian Journal for Science and Engineering 46, 8549–8563.

[24] Jaiswal, S.G., Mohod, S.W., . Classification of violent videos using ensemble boosting machine learning approach with low level features .

[25] Jayasimhan, A., Pabitha, P., 2022. A hybrid model using 2d and 3d convolutional neural networks for violence detection in a video dataset, in: 2022 3rd International Conference on Communication, Computing and Industry 4.0 (C2I4), IEEE. pp. 1–5.

[26] Jing, F., Liu, L., Zhou, S., Song, J., Wang, L., Zhou, H., Wang, Y., Ma, R., 2021. Assessing the impact of street-view greenery on fear of neighborhood crime in guangzhou, china. International journal of environmental research and public health 18, 311.

[27] Long, D., Liu, L., Xu, M., Feng, J., Chen, J., He, L., 2021. Ambient population and surveillance cameras: The guardianship role in street robbers' crime location choice. Cities 115, 103223.

[28] Mahalle, M.D., Rojatkar, D.V., 2021. Audio based violent scene detection using extreme learning machine algorithm, in: 2021 6th international conference for convergence in technology (I2CT), IEEE. pp. 1–8.

[29] Martínez-González, M.B., Turizo-Palencia, Y., Arenas-Rivera, C., Acuña-Rodríguez, M., Gómez-López, Y., Clemente-Suárez, V.J., 2021. Gender, anxiety, and legitimation of violence in adolescents facing simulated physical aggression at school. Brain Sciences 11.

[30] Muarifah, A., Mashar, R., Hashim, I.H.M., Rofiah, N.H., Oktaviani, F., 2022. Aggression in adolescents: The role of mother-child attachment and self-esteem. Behavioral Sciences 12.

[31] Mugunga, I., Dong, J., Rigall, E., Guo, S., Madessa, A.H., Nawaz, H.S., 2021. A frame-based feature model for violence detection from surveillance cameras using convlstm network, in: 2021 6th International Conference on Image, Vision and Computing (ICIVC), IEEE. pp. 55–60.

[32] Mumtaz, N., Ejaz, N., Aladhadh, S., Habib, S., Lee, M.Y., 2022. Deep multi-scale features fusion for effective violence detection and control charts visualization. Sensors 22, 9383.

[33] Negre, P., Alonso, R.S., Prieto, J., Dang, C.N., Corchado, J.M., 2024a. Systematic mapping study on violence detection in video by means of trustworthy artificial intelligence. Available at SSRN 4757631 .

[34] Negre, P., Alonso, R.S., Prieto, J., Novais, P., Corchado, J.M., 2024b. Violence detection in video models implementation using pre-trained vgg19 combined with manual logic, lstm layers and bi-lstm layers, in: DCAI 2024. In press.

[35] Nurisma, S.Z., Astuti, B., 2023. Peace sociodrama: A strategy to reduce junior high school aggression. International Journal of Social Service and Research 3, 1319–1324.

[36] Omarov, B., Narynov, S., Zhumanov, Z., Gumar, A., Khassanova, M., 2022. State-of-the-art violence detection techniques in video surveillance security systems: a systematic review. PeerJ Computer Science 8, e920.

[37] de la Salud, O.M., 2021. Violence against women.

[38] Sharma, S., Sudharsan, B., Naraharisetti, S., Trehan, V., Jayavel, K., 2021. A fully integrated violence detection system using cnn and lstm. International Journal of Electrical & Computer Engineering (2088-8708) 11.

[39] Soliman, M.M., Kamal, M.H., El-Massih Nashed, M.A., Mostafa, Y.M., Chawky, B.S., Khattab, D., 2019. Violence recognition from videos using deep learning techniques, in: 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 80–85.

[40] Srivastava, A., Badal, T., Garg, A., Vidyarthi, A., Singh, R., 2021. Recognizing human violent action using drone surveillance within real-time proximity. Journal of Real-Time Image Processing 18, 1851–1863.

[41] Srivastava, A., Badal, T., Saxena, P., Vidyarthi, A., Singh, R., 2022. Uav surveillance for violence detection and individual identification. Automated Software Engineering 29, 28.

[42] Su, Y., Lin, G., Zhu, J., Wu, Q., 2020. Human interaction learning on 3d skeleton point clouds for video violence recognition, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, Springer. pp. 74–90.

[43] Talha, K.R., Bandapadya, K., Khan, M.M., 2022. Violence detection using computer vision approaches, in: 2022 IEEE World AI IoT Congress (AIIoT), IEEE. pp. 544–550.

[44] Traoré, A., Akhloufi, M.A., 2020a. 2d bidirectional gated recurrent unit convolutional neural networks for end-to-end violence detection in videos, in: International Conference on Image Analysis and Recognition, Springer. pp. 152–160.

[45] Traoré, A., Akhloufi, M.A., 2020b. Violence detection in videos using deep recurrent and convolutional neural networks, in: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE. pp. 154–159.

[46] Ullah, F.U.M., Muhammad, K., Haq, I.U., Khan, N., Heidari, A.A., Baik, S.W., de Albuquerque, V.H.C., 2021. Ai-assisted edge vision for violence detection in iot-based industrial surveillance networks. IEEE Transactions on Industrial Informatics 18, 5359–5370.

[47] Ullah, F.U.M., Obaidat, M.S., Muhammad, K., Ullah, A., Baik, S.W., Cuzzolin, F., Rodrigues, J.J., de Albuquerque, V.H.C., 2022. An intelligent system for complex violence pattern analysis and detection. International Journal of Intelligent Systems 37, 10400–10422.

[48] Vijeikis, R., Raudonis, V., Dervinis, G., 2022. Efficient violence detection in surveillance. Sensors 22, 2216.

[49] Vomfell, L., Härdle, W.K., Lessmann, S., 2018. Improving crime count forecasts using twitter and taxi data. Decision Support Systems 113, 73–85.

[50] Vosta, S., Yow, K.C., 2022. A cnn-rnn combined structure for real-world violence detection in surveillance cameras. Applied Sciences 12.

[51] Wilkinson, S., 2016. Meet the heroic campaigners making cities safe for women. Global Street Harassment-Making Street Safer: Action Aid .

[52] Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z., 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16, Springer. pp. 322–339.

[53] Yue, H., Xie, H., Liu, L., Chen, J., 2022. Detecting people on the street and the streetscape physical environment from baidu street view images and their effects on community-level street crime in a chinese city. ISPRS International Journal of Geo-Information 11, 151.