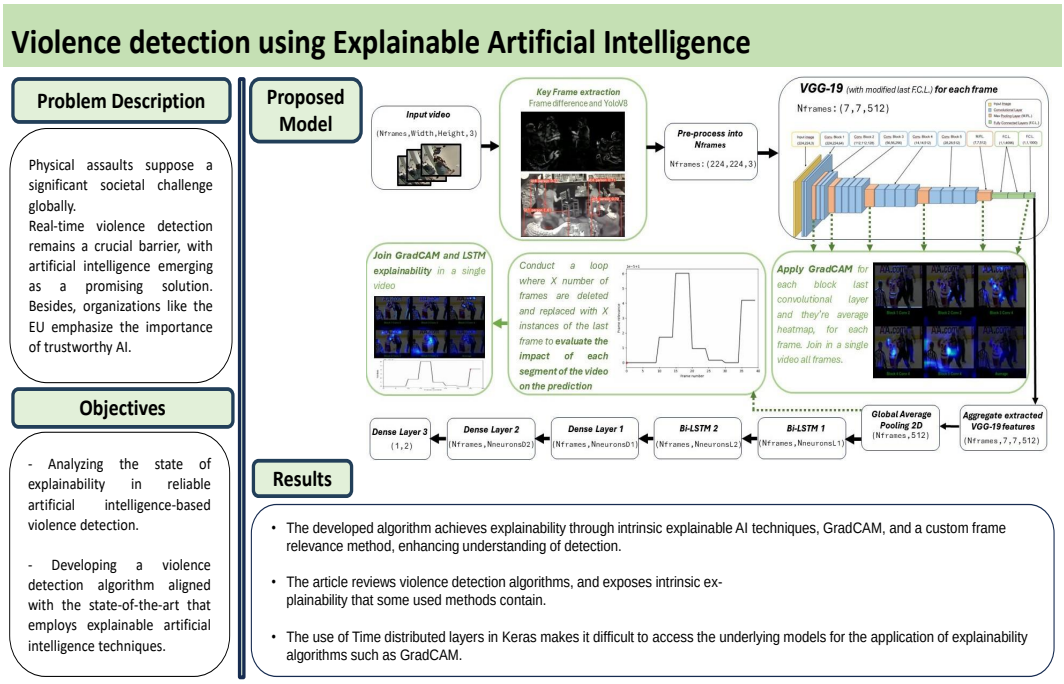
Graphical Abstract

eXplainable Artificial Intelligence combining CNN and RNN layers for violence detection in video

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



Highlights

eXplainable Artificial Intelligence combining CNN and RNN layers for violence detection in video

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

- The first algorithm for detecting violence in videos using explainable artificial intelligence has been developed.
- The algorithm achieves explainability through intrinsic explainable AI techniques, GradCAM, and a custom frame relevance method, enhancing understanding of detection.
- The article reviews violence detection algorithms, and the intrinsic explainability that some methods contain.
- The use of Time distributed layers in Keras makes it difficult to access the underlying models for the application of explainability algorithms such as GradCAM.

eXplainable Artificial Intelligence combining CNN and RNN layers for violence detection in video

Pablo Negre[a], Javier Prieto[a], Ricardo S. Alonso[b,c], Pablo Chamoso[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*

Abstract

Physical assaults suppose a significant societal challenge globally, causing both physical and mental harm while disrupting social order. Despite various approaches, real-time violence detection remains a crucial barrier, with artificial intelligence emerging as a promising solution. Besides, organizations like the EU emphasize the importance of trustworthy AI. First the work exposes a comprehensive state of the art on video violence detection focusing on the use of the combination of Convolutional Neural Network and Long Short Term Memory, on the use of trustworthy artificial intelligence (in particular of explainable artificial intelligence), as well as the combination of video violence detection using explainable artificial intelligence. The main objective of the work has been the development of a violence detection algorithm in video using explainable artificial intelligence. The explainable arquitecture is based on a previous study model, which uses pre-trained VGG19 combined with Bi-LSTM layers. The developed explainable architecture uses YoloV8 and Frame difference as the Key frame extraction method. GradCAM has been used over the last five convolutional layers of each VGG19 block across

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

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

all video frames to identify areas highlighted by VGG19 in violence detection. Additionally, a method has been devised to determine which frames of the video are most relevant in violence detection based on iteratively eliminating frames from the video, making successive predictions and comparing their results with the prediction of the original video.

Keywords: Violence detection, Explainable Artificial Intelligence (XAI), Pretrained Convolutional Neural Network (PT CNN), Long Short Term Memory (LSTM), Video Surveillance

1. Introduction

Physical violence is a big problem in our world, affecting people everywhere. This issue hurts not only the people directly involved but also their families, communities, and the overall progress of society (like people moving around, tourism, and shopping) [1] [2]. The reasons behind aggressive behavior often stem from difficulties in handling emotions, conflicts between individuals, as well as the social and economic conditions of societies [3] [4].

Two groups hit hardest by physical violence are women and children. According to a report from the World Health Organization (WHO), about one in every three women globally has experienced physical or sexual violence from an intimate partner or non-partner in their lifetime [5]. A study by ActionAid [6] found that high percentages of women in India, Thailand, Brazil, and London face harassment or violence in public spaces. Regarding children, research published in the American Academy of Pediatrics revealed that more than half of children in Asia, Africa, and North America experienced violence in 2015, with over a billion children worldwide affected. Additionally, a study by the European Union Agency for Fundamental Rights (FRA) [7] found that more than one in four Europeans experienced harassment, and 22 million were physically assaulted in a year.

Overall, physical violence is a widespread problem globally, affecting nearly every aspect of people's lives and entire societies. Ensuring the safety of individuals worldwide should be a top priority and a fundamental right for everyone.

Many ways to address violent acts in societies have been explored. Some studies focus on understanding what situations or environments trigger people to commit violence so they can be improved [1] [8] [9]. Other studies look at how crime rates relate to urban environments where they happen [10] [11]

[12]. The last line of defense for victims of violence is detecting violence in real-time to alert authorities and gather evidence like videos or images to identify those involved. Video violence detection using artificial intelligence is crucial for large-scale detection without needing many people to monitor cameras or patrol streets.

This field is growing rapidly, thanks to advances in three main areas: increased use of images and security cameras, technological advancements in big data platforms, and improvements in artificial intelligence algorithms for image and video analysis [13] [14]. Violence detection in videos using artificial intelligence falls under computer vision, specifically in the subset of human action detection. Within this, it's seen as an anomalous action because it's unusual or unexpected in many social settings or contexts.

Most artificial intelligence algorithms, especially those based on deep learning, are hard to understand, often called black-boxes because we can't see what happens from input to output [15]. Simpler algorithms like decision trees are easier to understand but less accurate [16]. Creating trustworthy AI is a key goal. Big organizations like the European Commission are working on defining standards and legislation. The European Union has outlined trustworthy AI with three pillars [17]: lawful, ethical and robust. Within the ethic pillar, there is the explainability of algorithms, which means that their decision processes must be understandable by users; explainable artificial intelligence (XAI) has gained relevance in recent years, where new techniques have appeared to provide understanding of algorithms [18].

Therefore, the aim of this work is to develop a violence detection algorithm that combines a pre-trained Convolutional Neural Networks, in this case VGG19 pre-trained in Imagenet, together with Bi-LSTM layers applying explainable artificial intelligence techniques, As far as is known, it will be the first violence detection algorithm to apply explainability. Intrinsically explainable methods such as key frame selection, GradCAM will be applied to find out which area of the video frames are key for violence detection and a proprietary method will be developed for the quantification of how important each video frame is for violence detection.

2. State of the Art

This Section is divided into three Sections. Section 2.1 explains the state of the art of AI-based violence detection. Section 2.2 presents the state of the art of explainable artificial intelligence (XAI). Finally, Section 2.3

presents the state of the art of violence detection using explainable artificial intelligence.

2.1. Violence detection in video by means of Artificial Intelligence

This Section presents the basic steps of video violence detection, the types of algorithms and datasets used, with special attention to the use of the combination of CNN and LSTM.

The basic steps in violence detection in video using AI are exposed in Figure 1. Real-time violence detection with AI involves analyzing a continuous stream of images or videos. Since analyzing real-time data constantly can be costly, especially for videos, lighter computational techniques are often used; these techniques focus on analyzing only potentially violent images or frames, although not all advanced methods include extracting characteristic frames. After preprocessing, potentially violent images or frames are inputted into a trained algorithm to classify them as violent or non-violent. Violence detection algorithms typically extract features from videos (like spatial or temporal information) and then train an algorithm to recognize violence based on these features. Some algorithms have two-stage processes, where features are extracted first and then another algorithm learns from them. Others combine feature extraction and learning into a single step, or use two separate algorithms for feature extraction and learning. Once features are extracted, they're fed into a classifier to determine if the scene is violent. This process involves testing various combinations of settings and evaluating metrics until the best structure is found.

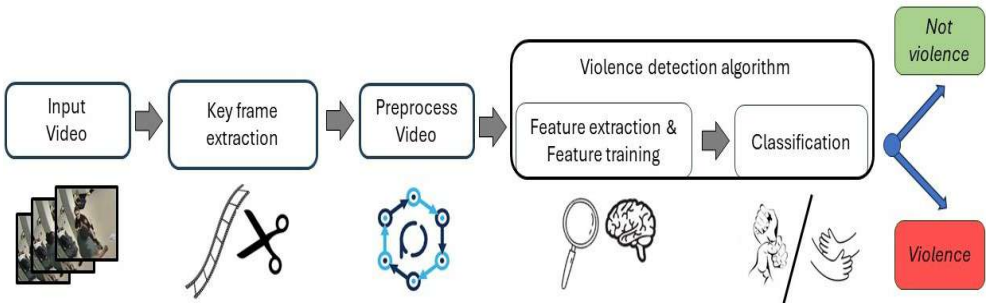


Figure 1: Violence detection basic steps

There exists a wide array of algorithms for identifying violence. Traditionally, they've been categorized into: *Traditional Methods*, which rely on

manual feature extraction and traditional Machine Learning algorithms; and *Deep Learning Methods*, which utilize Deep Learning techniques [19], [20], [21], [22]. However, this classification is overly broad. The preferred classification used in this work is by algorithm type, of which the following have been used: CNN [23] [24], LSTM [20], Manual feature (those algorithms in which feature extraction, feature training, or both, are performed on a mathematical basis) [25] [26], Skeleton-based (Deep Learning or Manual) [27] [28], Transformer [29] [30], and Audio-based (Deep Learning or Manual) [31] [32].

Additionally, combining two algorithms, regardless of their nature, is the most common approach in current research. This is done to leverage the advantages of both types of algorithms. The most popular combination for violence detection is using CNN and LSTM [33]. The papers published between July 2021 and July 2023 that have used CNN in combination with some form of RNN are compiled in Table 1.

Table 1: Violence detection articles which use CNN and RNN combination

Cite	**CNN**	**LSTM**	**CNN-LSTM Connection**	**Hockey Fight**	**Action Movies**	**Violent Flow**	**RWF-2000**
[34]	MobileNet V2	LSTM	T.D.L	99.5	96.1	X	82
[35]	Own CNN	Bi-LSTM	Concatenation, Flatten	99.27	100	98.64	X
[36]	PT VGG-16 (ImageNet)	LSTM	Concatenation	99.1	X	X	X
[22]	PT VGG-16 (ImageNet)	Bi-Conv-LSTM	No aggregation (Conv-LSTM)	99.1	100	98.4	92.4
[37]	PT MNAS CNN	Conv-LSTM	No aggregation (Conv-LSTM)	99	96	100	X
[38]	Two PT EfficienNet-B0 (ImageNet)	Bi-LSTM	Concatenation	99	X	93.75	X
[39]	Two PT VGG-16 (ImageNet) + Wide Dense Residual Blocks (WDRB)	LSTM	Concatenation	98.8	98.99	97.1	X
[40]	PT VGG-16 (ImageNet)	LSTM/Bi-LSTM	Concatenation	97.6/98.8	X	92.2/95.1	X
[19]	Darknet + Residual Optical Flow	M-LSTM	Concatenation	98	X	98.21	X
[**?**]	PT VGG16 (INRA)	Bi-GRU	Concatenation	98	X	95.5	X
[41]	Own CNN	LSTM	Concatenation	97	100	X	X
[42]	PT Xception (ImageNet)	LSTM	T.D.L, Flatten	96.55	98.32	X	X
[43]	CNN	LSTM	Concatenation	94.9	92.2	77.31	X
[44]	PT VGG-19 (ImageNet) + extra CNN layers	Bi-LSTM	Concatenation, Flatten	91.29	X	90.47	90.47
[45]	PT: VGG16/ VGG19/ InceptionV3/ DenseNet201/ ResNet101/ MobileNet/ NASNetLarge/ VGG16+VGG19/ ResNet50+ResNet152V2/ InceptionV3/ResNet101V2	LSTM	Concatenation	X	X	X	X
[46]	PT MobileNetV2 (ImageNet)	Bi-LSTM/ ConvLSTM	T.D.L, Flatten	X	X	X	X

Column *CNN-LSTM Connection* indicates how the extracted spatial features are transmitted from the CNN to the LSTM, although none of the articles makes spatial mention of it when developing the model architecture. This step is relevant for the application of explainable artificial intelligence algorithms. The pre-trained CNNs used in the selected papers are trained on image datasets, so their structure is designed to receive an image, extract spatial information from it and, after densely connected layers, give a result about the probability of belonging to a specific class. In general, CNNs (pre-trained or not) have usually been used for the analysis of images, not video (since these have temporal information), so a way must be found to make the CNN finish calculating all the spatial features contained in the frames of a video to pass them later as input to the LSTM. The three forms used in the works shown in Table 1 are outlined below and will be addressed again in Section 4:

- **Conv-LSTM**: the convolutional-LSTM structures allow that, while the CNNs are extracting spatial features from the video frames, simultaneously the Conv-LSTM is analysing for each time jump the temporal relations [46] [37]. Therefore, this architecture has its own solution.
- **Spatial feature concatenation**: another option is to have the two models separately, on the one hand the CNN and on the other hand the LSTM. In this way, when the CNN finishes analysing all the frames of the videos, the spatial features extracted from all the frames are grouped in an ordered way in a single element (for example, an array). This implies that the two algorithms (CNN and LSTM) are trained separately. In case the CNN is pre-trained, it may not be trained again with the violence data. On the other hand, the densely connected layers would be trained together with the LSTM. However, this process implies that the training process is not carried out jointly, although there is no mention in the cited articles that this might affect the accuracy of the whole process.
- **Time Distributed Layer**: Keras Time Distributed Layers allow the execution of one layer of a neural network, several layers or even a complete model for each time sequence (for each frame, for example). This means that it is possible to introduce a pre-processed video to a CNN, that the CNN processes each of the frames or time sequences and that until they are all finished and encapsulated in the same element,

the output is not transmitted to the next layer (an LSTM, in this case). This is a big change with respect to the *Spatial feature concatenation* approach, as it allows having a single model that can be trained jointly, in addition to the fact that the whole process from the input of the pre-processed video to the bi-class output of whether the scene is considered violent or not, is done by calling a single model.

All in all, video violence detection has common general steps, although numerous types of algorithms have been used to tackle the task; the combination of CNN and LSTM is the most widely used, with three ways of linking the two types of algorithms having been used in recent state-of-the-art work.

2.2. Explainable artificial intelligence in computer vision

This section will discuss the need for the development of trustworthy artificial intelligence, the development of its definition and standardisation, how to categorise explainable artificial intelligence(XAI) algorithms and a brief description of the most commonly used explainability algorithms.

The vast majority of artificial intelligence algorithms, especially those based on deep learning, make it very difficult to understand the results obtained, making the term black-boxes popular, referring to not understanding what happens from the input to the output of the algorithm. Relatively simple algorithms such as decision trees based on traditional machine learning are more understandable, however they tend to have lower accuracy [16]. Therefore, the generation of artificial intelligence algorithms that, despite their complexity, would be able to guarantee certain bases or guarantees regarding their good results, is a very interesting objective to achieve; in essence, *trustworthy* artificial intelligence. Big organisations have tried in recent years to create definitions, standards and legislation around this concept, including the International Organization for Standardization (ISO) [15] and the European Commission. In particular, the European Union published a report defining trustworthy artificial intelligence and the pillars on which it is based [17]:

- Lawful: Ensuring adherence to applicable laws and regulations.
- Ethical: Prioritizing principles and values that uphold moral considerations and standards.

- Robust: Technically and socially resilient to prevent inadvertent harm caused by well-intentioned AI systems.

These three pillars are defined as necessary for trustworthy artificial intelligence, not being possible the lack of one of them. Within ethical artificial intelligence, it is stated that it must be explainable, that is, the process of generation and the result must be comprehensible in some way. The development of explainable artificial intelligence algorithms has been increasing in recent years, generating algorithms with different approaches in order to achieve algorithms whose processes and results are more understandable. The following are the classification forms of explainable artificial intelligence (XAI)-based algorithms

- **Global or Local**. Global XAI methods try to understand the functioning of the model as a whole, trying to give an overview of how certain features affect its decisions. On the other hand, local methods try to give specific explanations to individual elements (an image, a text...).

- **Intrinsic or Post-hoc**. Intrinsically explainable methods are those that by their very nature make their way of processing the results comprehensible. For example, decision trees are inherently explainable models since each node in the tree represents a specific feature that is split according to which value is taken. Post-hoc methods, on the other hand, are applied once a trained model is available. A clear example of Post-hoc methods are neural networks, since even knowing their internal architecture, the decision making and generation of the result is, if not completely opaque, very difficult to understand.

- **Agnostic or Specific**. Agnostic methods are independent of the type of model on which they are applied, i.e. they are not designed for a certain type of architecture. On the other hand, specific methods are those designed specifically for one type of algorithm.

There are a large number of models that have different approaches to the classifications explained. Three of the best known and most widely used models are: LIME, SHAP and Grad-CAM.

LIME is model agnostic, local and is applied post-hoc. That is, regardless of the underlying architecture, it derives explainability from the individual

instances that the model generates. To generate explainability LIME tries to generate perturbed instances of the original input data; based on how the outcome of the data varies according to the small perturbations introduced, trying to fit a linear model (or a simple decision tree) to the prediction made [47]. Despite the good results obtained, LIME can present some drawbacks such as: the variation in the interpretation it performs according to the perturbations generated and the choice of the interpretable model chosen, since different models can give rise to different explanations [48]. LIME has been utilized for multiple machine learning models, including classification and regression, providing local explanations on model decisions by highlighting the importance of features used in prediction [16].
As well as LIME, SHAP is model agnostic, local and is applied post-hoc. SHAP uses game theory to assign fair values to each feature, decomposing their contribution in an additive way [49]. It considers all possible combinations of features, generating consistent and local explanations for model predictions. As a trade-off, SHAP can be computationally expensive for large datasets or complex models [50]. In summary, both use perturbed data instances to understand and explain the behaviour of artificial intelligence models; in addition to being both XAI agnostic, local and post-hoc algorithms widely used for different types of algorithms in the literature.
On the other hand, Grad-CAM is local, post-hoc and specific. This is an important difference from SHAP and LIME, as GradCAM is specifically designed for Convolutional Neural Networks (CNN). It is a visualisation technique that highlights important regions of an image for class-specific prediction, providing activation maps visualising which parts of the image most influenced the network's decision [16]. GradCAM obtains the gradients of the target class with respect to the desired convolutional layer activations. These gradients are globally averaged over each channel, generating importance weights. By combining the information weighted by these weights, GradCAM produces an activation map that highlights crucial regions for target class prediction in an image [50].

Both SHAP, LIME and GradCAM have been applied to image classification with CNN architectures. Although each of them has a different approach, as discussed above, in the case of image classification they generate heatmaps indicating which part of the image is most relevant for decision making. These results should be taken as approximations to the reasoning of the model and not as exact representations. On the other hand, while these algorithms are able to quantify how relevant an element is in a classi-

fier (be it text, structured data or image), they are not able to quantify the relevance of one element with respect to another over time. Explainability in algorithms that contain a temporal relationship of their elements has been little addressed, although it would be of great interest to be able to know in a predicted time series, which element is more characteristic to make a prediction, or, in the case of video violence detection, which frames have been more relevant in the prediction. Only one explainable model that quantifies the importance of events (over time) and not of classes (as do SHAP, LIME and GradCAM) has been found. This method is called TimeSHAP [51] and has so far been little used in the state of the art. TimeSHAP is a technique that extends the SHAP framework to assess the importance of features in time series models.

2.3. Violence detection in video by means of Explainable artificial intelligence

As it has been seen, there are a wide variety of explainable artificial intelligence algorithms that can also be applied to the objective of this work, which is the development of a violence detection algorithm using explainable artificial intelligence. It is therefore worth asking whether any violence detection papers using explainable artificial intelligence algorithms have been published to date. To the best of our knowledge, no articles have been published that take this approach, although there are articles that use techniques in some of the processes of violence detection that could provide explainability, even if this is not its objective.

As discussed in Section 2.1, the key frame extraction process aims to provide computational lightness when selecting frames with potential violence. There are articles that use people detection algorithms for this purpose, which in the case of a violent scene, could provide explanations as to why the scene is considered potentially violent in the first place, as well as the location of the people involved in the scene (provided that the scene is not crowded) [52], [53], [54]. On the other hand, the use of certain kinds of inputs can help to understand the scene from another point of view. For example, there are articles that use as inputs of the algorithm *Frame difference* [37] [55] or *Separate motion energy picture* [56] [57] which perform calculations based on the subtraction of values between pixels and different motion components; which highlights certain areas where violence may potentially exist. Most violence detection algorithms are generally "black-boxes"; a term used to describe that it is not understood what leads an algorithm to obtain certain results. Within this group are CNNs, LSTMs, Transformers, etc. However,

the Skeleton-based type algorithms are based on the detection of the scene according to the position and movement of the body of the people involved; often emphasising these positions with dots and lines [28].

However, the reality is that no articles have been found that deal with the detection of violence through the use of explainable artificial intelligence, although methods have been used that can provide certain intrinsic explainability, even if they were not initially used for that purpose.

3. Proposed Violence Detection Algorithm Arquitecture

This section will be divided into two parts. First, the explanation of the violence detection model will be presented. Then, the architecture of the model shown by applying explainable artificial intelligence techniques will be described.

3.1. Violence detection in video algorithm arquitecture

This section will outline the architecture of the proposed video violence detection model. The architecture consists of the combination of pre-trained VGG19 in ImageNet combined with Bi-LSTM layers, where VGG19 will extract spatial features that will be introduced into the Bi-LSTM layers for temporal feature extraction. VGG19 is a Convolutional Neural Network (CNN) designed for image classification. Since its last dense layer has 1000 neurons, as ImageNet dataset has a thousand different classes, this last layer has been removed and replaced by a two-neuron layer, so that after training this last new layer, VGG19 classifies images of violence and non-violence. While the spatial features obtained by the convolutional layers of VGG19 will be extracted for violence detection, VGG19 does need to be able to classify images of violence when applying explainability algorithms.

First, an architecture was developed using a Time Distributed Layer to encapsulate pre-trained VGG19, which can be seen in Figure 2. This architecture has the advantage that the training of the model can be performed as a single algorithm since the use of the Time Distributed Layers allows that until the extraction of the features of all the frames of the video is finished, the process does not continue towards the use of the Bi-LSTM. Although there is no evidence that doing this process as a single block has benefits, it facilitates the training process by having a single model and process to train. In addition, as mentioned above, adding VGG-19 within a Time Distributed

Layer that is part of a more complex algorithm (all in a relatively simple way) makes the model creation process much easier.

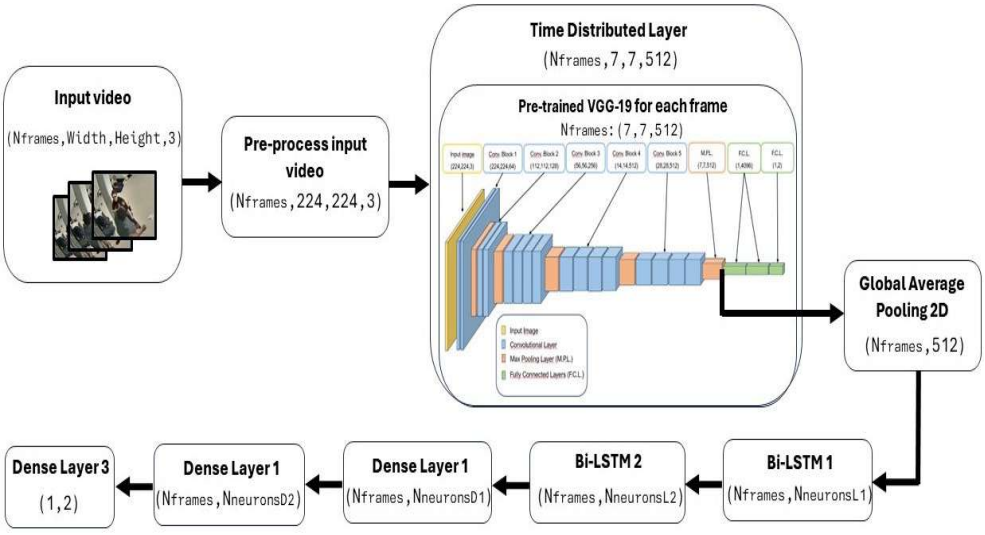


Figure 2: Violence detection arquitecture using Time Distributed Layers

Despite the advantages, problems have been encountered when implementing explainable artificial intelligence algorithms with this architecture. To apply GradCAM (as a widely used XAI algorithm) a submodel must be created from the trained model to obtain two outputs: the classification result and the result of the convolutional layer to be analysed. However, Keras and Tensorflow make it really complex to access the layers of a model that is in turn inside another model, as is the case of VGG-19 in the architecture designed. During this work, version 2.14.0 of Keras and TensorFlow has been used, which is the latest version available at the time of development. It was not possible to generate the submodel with the two outputs mentioned above, given the persistent error: *ValueError: Graph disconnected: cannot obtain value for tensor*.

Therefore, an architecture that divides the process into two distinct parts, as discussed in Section 2, based on spatial feature concatenation between VGG19 and the Bi-LSTM layers is used. The training and testing in multiple datasets of this violence detection architecture is published in [58]. It's arquitecture is exposed in Figure 3.

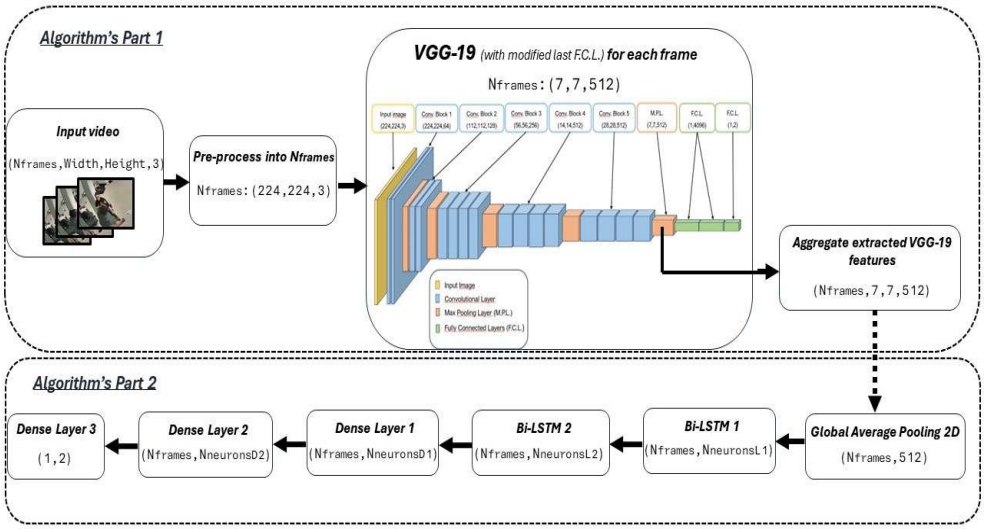


Figure 3: Violence detection arquitecture based on Spatial Features Concatenation [58]

The proposed architecture for this work, is exposed in Figure . Initially, the video undergoes division into N_{frames} and resizing to the dimensions (224, 224), compatible with VGG-19. Subsequently, the frames are fed individually into VGG-19, unlike the architecture illustrated in Figure 2, where the Time Distributed Layer incorporating the VGG-19 model received the frames aggregated within a single unit. Following the calculation of spatial features for each frame by VGG-19, these features are jointed into a singular entity upon exiting the final convolutional layer, taking the shape (7,7,512), thereby forming a structure of the type: (N_{frames},7,7,512). Upon aggregation of the pooled features, a *Global Average Pooling 2D* layer is applied, reducing the dimensionality to the format (N_{frames}, 512), and subsequently passed through two Bi-LSTM layers and three densely connected layers with Sigmoid activation, akin to the architecture depicted in Figure 2. However, the primary distinction in this architecture lies in the division of the process into two discrete segments, as opposed to a singular layered structure of successive neural networks. This implies that for training certain layers of the pre-trained CNN to adapt it for violence detection, as undertaken in this study, VGG-19 must be trained separately from the Bi-LSTM layers and dense layers. There is no indication in the analyzed state of the art whether this combination decreases accuracy results due to the CNN and Bi-LSTM being trained independently rather than as a unified architecture.

3.2. Explainable violence detection in video algorithm arquitecture

This section presents the architecture of the video violence detection model shown in Figure 3.1, together with the application of explainable artificial intelligence techniques. This architecture is shown in Figure 4, where it is implemented once the violence detection model is trained.

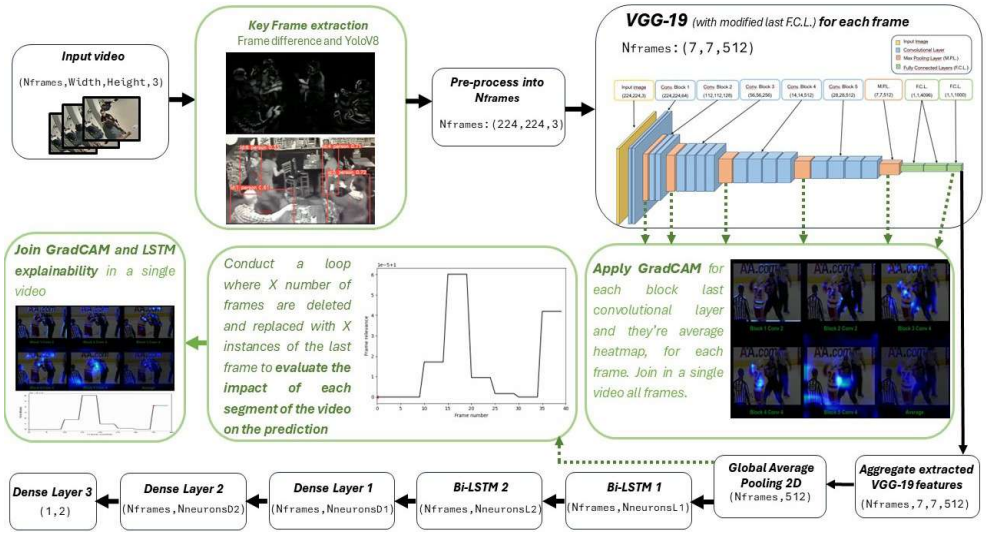


Figure 4: Explainable violence detection arquitecture

As a first element that provides explainability, a Key Frame extraction process based on the use of Frame Difference and YoloV8 is applied. It has been exposed in the state of the art as real-time video analysis is very computationally expensive, so the application of lightweight models that select frames that are likely to contain violence is relevant. Yolo verison 8 is used as a fast and optimised people detection algorithm that keeps a count of the people appearing in each frame, as well as being able to track and count them. The use of the frame difference consists of highlighting the parts of the video where there is more movement. Both algorithms have been used separately in the state of the art and the combination of both is proposed to minimise the number of key frames analysed by the violence detection algorithm. Depending on the type of scene that the video camera is analysing, a minimum number of people to be detected by YoloV8 and a minimum frame difference value should be established in order to pass these frames to the

violence detection algorithm when both conditions are met. Este primer elemento supone comprender porque la arquitectura supone relevante ciertos frames como candidatos para la detección de violencia.

The second element of explicability provided by the proposed structure is GradCAM used on VGG19. Gradients are obtained from each last convolutional layer of each block, allowing us to understand which parts of the frames are more relevant as VGG19 computes spatial features. Consequently, a total of 5 heatmaps are obtained, which are overlaid onto the original frames. Additionally, the average of the 5 heatmaps is calculated, although initially, the last convolutional layer is presumed to contain the most knowledge, thereby maximizing the most prominent areas of each block. Ultimately, a video with 6 sub-videos divided into grids for each heatmap overlaid with the original video is obtained. Applying GradCAM in this manner signifies a detailed understanding of how VGG19 processes video frames and where violence is spatially located.

The third element of explicability is providing information about which of the frames of the video is more relevant for violence detection. It has been noted in the state of the art that there are no algorithms that provide explainability about the relationship between events, in this case between frames, except for TimeSHAP. In this case, a different solution is intended to be provided to understand which frames are more relevant for violence detection. For this purpose, a loop is executed that removes 5 frames from the video in each iteration, starting from the beginning and ending at the end of the video, and adds the last frame of the video 5 times. This operation involves making predictions without those frames for each combination of removed frames, so if the accuracy percentage of that combination is lower than the original, those frames are relevant for the prediction; moreover, the lower the accuracy of the prediction, the more relevant they will be.

Five frames are chosen because it was found that a single frame did not yield very clear results, and since violence is a quick action, five frames constitute around one-sixth of a second, making it possible to differentiate key parts of the video. A time series is represented where the X-axis is the number of frames of the video and the Y-axis is the accuracy of the original video (in the numerator) divided by the accuracy of that combination without the 5 frames, thus maintaining the proportion that a higher value of the division implies greater importance of those 5 frames, and vice versa. This iterative application of the model is not computationally costly since it is only applied to the Bi-LSTM layers and dense layers from the spatial

features extracted by VGG19, making its calculation very fast. This method is relevant as it allows understanding which parts of the video are temporally relevant for violence detection.

Finally, the results of the six GradCAM heatmaps are unified into a single video along with the time series, which is animated with a red circle moving along the time series, making it easy to see where in the image and which moments of the video are most relevant for violence detection. Automatically, this video is stored along with the results of YoloV8 and Frame difference in a folder for further analysis. This explainability process provides valuable information to the user from the beginning to the end of the process, especially considering that CNNs and LSTMs are completely opaque algorithms.

4. Results

This section is going to present the results obtained during the training and testing phase of the architecture presented in Section .

4.1. Datasets

This section presents the datasets used for this work. Those datasets are: "Hockey Fights dataset", "Violent Flow dataset" and "Real World Fight 2000 dataset (RWF-2000)"; which are three of the four most used datasets in the recent state of the art [33]. "Hockey Fights dataset" [59] is chosen for being the most frequently utilized dataset and consists of action videos from field hockey games of the National Hockey League (NHL). While the altercations are genuine, the videos often feature close-ups, are well-illuminated, and do not portray a real-life scenario where the victim flees for assistance. Nonetheless, the ability to compare findings with a vast number of articles is a compelling rationale for its selection. Secondly, "Aggressive Flow dataset" [60] comprises YouTube real-world crowded videos with accompanying audio. Lastly, the "Real World Fight 2000 dataset (RWF-2000)" contains footage of violence captured by security cameras. It is arguably the most extensive dataset owing to its content and the quantity of 2000 videos.

It has been decided to exclude the "Action movies dataset", despite its status as the second most frequently employed dataset. This dataset includes scenes from action films characterized by close-ups and optimal lighting, ultimately depicting unreal scenarios. With the selected datasets, the objective is to juxtapose the findings derived from numerous state-of-the-art articles

Table 2: Test VGG19 and Bi-LSTM results [58]

Dataset	Test Accuracy	Learning Rate	Bi-LSTM 1	Bi-LSTM 2	F.C.L 1	F.C.L 2
Hockey Fights	0.97	0.01	64	256	64	256
RWF-2000	0.73	0.001	1024	1024	256	512
Violent Flow	0.90	0.01	64	512	64	128

and address three distinct forms of violence: a well-illuminated and documented setting like a hockey game, congested environments, and genuine occurrences captured by security cameras.

The datasets were divided into 70% for the training phase, 10% for validation and 20% for testing.

4.2. Training and testing pre-Trained VGG-19 + Bi-LSTM results

The training and testing of this architecture together with other proposals is developed more extensively in another paper, so the results obtained are summarised here. All calculations will be 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.

First, VGG19 is trained since the last dense layer has been replaced by a two-neuron output. Testing accuracys of 94%, 92% and 62% are obtained for the Hockey Fights, Violent Flow and RWF-2000 dartaset frames, respectively, for each frame of the test videos.

Once VGG19 is trained, the second part of the algorithm, consisting of two Bi-LSTM layers and three densely connected layers, is trained. The training is performed for the three datasets with 100 epochs. The values of the learning rate hyperparameters (0.1, 0.01 and 0.001) and the number of neurons in the Bi-LSTM and densely connected layers (64, 128, 256, 512 and 1024) are varied. All in all, the results of the best models obtained in the article where the model it's trained and tested for each dataset are shown in Table 2 [58].

4.3. XAI algorithm results

In this Section the results obtained during the application of the Key Frame extraction, GradCAM algorithms and the importance of the video frames will be presented. Although state-of-the-art research has been done on ways to quantify the results of the application of explainable artificial

intelligence algorithms, it is not a widely developed sector. Furthermore, quantifying how important an area or frame is for the detection of violence can be subjective to a certain extent. Therefore, observations made on the prediction of violence on certain videos based on the proposed architecture will be discussed.

The application of Key Frame extraction using the combination of YoloV8 and Frame difference has yielded positive results in several aspects. On one hand, both algorithms have very fast computation, which meets their purpose of real-time application. The computation of frame difference is on the order of 10^{-5} seconds, while YoloV8 calculates person detection in an average of 45 milliseconds. The condition that has been imposed is that YoloV8 must recognize a minimum number of people on the scene, and Frame Difference must obtain a minimum value of frame difference to pass the video fragment to the violence detection algorithm from that moment.

It has been observed how these two conditions can be useful and complement each other for different types of scenes. For example, scenes where the camera records highly crowded scenes do not make sense to impose a minimum number of people detected by YoloV8, but rather a higher than average frame difference movement. On the other hand, scenes recording a sparsely populated street at the pedestrian level, but with considerable traffic of cars or other vehicles, do not require a minimum number of frames since the vehicles are moving at high speed.

It has also been observed that depending on the position of the camera, the value of the frame difference varies significantly, with two main cases. The first case is if the camera is located far from the action, where it appears smaller and therefore the value of the frame difference will be smaller. The second case is that if the hits (punches, kicks, etc.) are perpendicular to the axis that joins the camera lens with the people, their amplitude is detected with their maximum value, but these decrease with the sine of the angle traced between the camera elements, the center of gravity, and the striking limb. However, while the position of the camera relative to that of the aggressors is not modifiable, the minimum value of frame difference can be modified by knowing the size of the people relative to the camera position.

Regarding the detection of a minimum number of people in the scene by YoloV8, it has been observed how occlusion between people and with other objects prevents correct counting in frames where this occurs. However, since YoloV8 has been implemented with the option of object tracking, when YoloV8 detects another person, it takes it into account in the scene, even if it

is hidden by another object, until the person completely exits it. An example is shown in Figure 5, where although the second Hockey player is occluded by the player with the number 24 jersey, YoloV8 detects the second hockey player, including him in the count of people inside the scene, and identifying the second referee as the fourth person within the action.

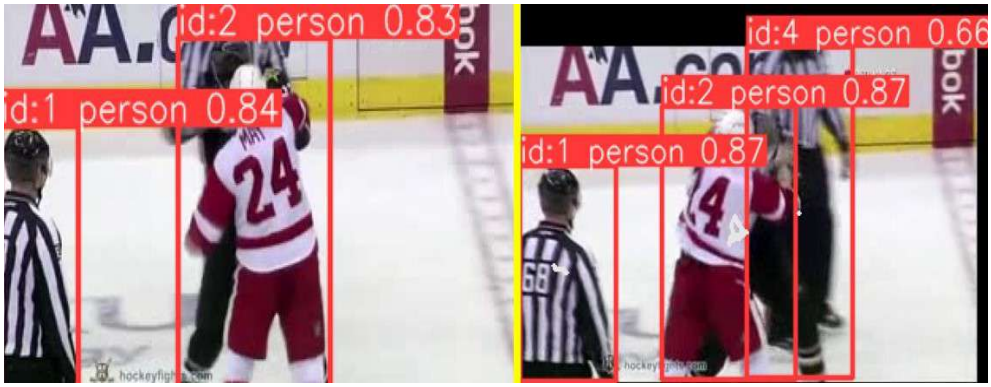


Figure 5: Hockey Fights dataset detection with YoloV8 on stage with occlusion between subjects

Moving on to the application of GradCAM on the prediction of VGG19, it proves to be very useful in understanding how VGG19's focus of attention in the last layer of each convolutional block. Although GradCAM is applied frame by frame, producing a video of heatmaps, a very representative example of what was observed with its results is attached in Figure 6.

It can be observed how the first and second convolutional blocks mainly focus on the contours and patterns of the bodies, clothes and signs that appear. The third and fourth convolutional blocks highlight the arm and body position of the players, also erroneously focusing on the signs and the referee to the left of the scene. The fifth (and last) convolutional block highlights the area of connection between the two hockey players, which is right at the center of the violence, although it also partially and erroneously highlights the hockey referee on the left side of the frame. The last frame, which is the average, helps to highlight not only the intersection area that highlights block 5 but also the bodies of the hockey players, although highlighted areas containing no violence are also included. One of the interesting utilities considered for the GradCAM application is to be able to verify that the al-

Figure 6: GradCAM result for the last convolutional layer of each block of VGG19 and the average of heatmaps for one frame of a video from the Hockey Fights dataset.

gorithm correctly focuses on the areas where violence occurs when predicting a violent act, and to observe in which areas it does not focus when making incorrect predictions. This can facilitate understanding the weaknesses of the model and balancing the training dataset accordingly.

Regarding the explainability method implemented to quantify the relevance of frames in violence detection, the speed of calculation stands out. Once the spatial features are extracted by VGG19, the manipulation of arrays and repeated execution of predictions by the Bi-LSTM layers is very fast. It has been observed that the variation in accuracy when varying groups of 5 frames iteratively is within the range of $10^{-}3$ to $10^{-}6$. This is very interesting as it demonstrates how robust the model's prediction is to variations, although it may also mean that the explainability values obtained are not highly reliable. There seems to be a trend of more significant frames when the camera axis with respect to the fight is perpendicular, that is, when

the aggressors are observed without overlapping each other in the plane, as depicted in the temporal series of Figure 7

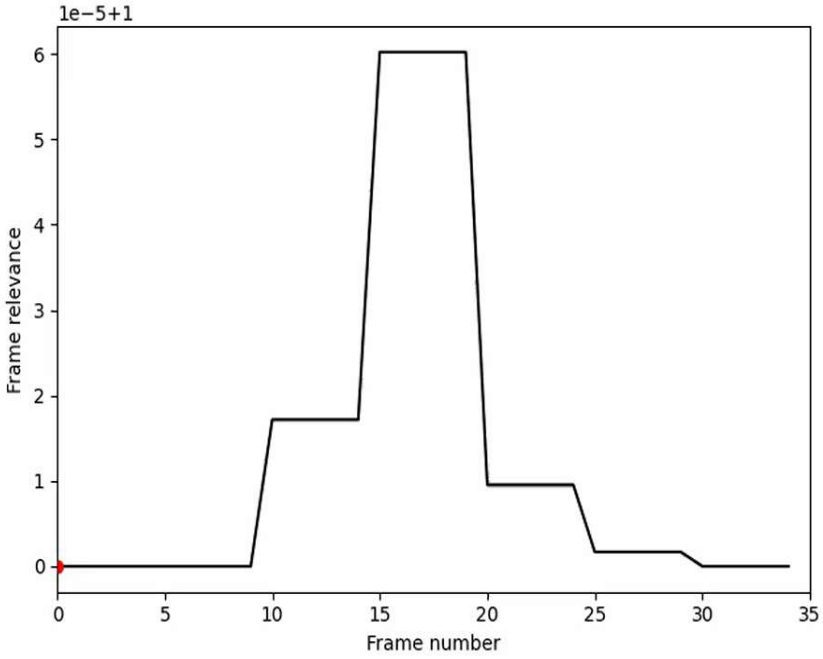


Figure 7: Frame relevance of "fi5_xvvid" Hockey Fights video dataset

All in all, the model's robustness against frame variations could imply the unnecessary recording of the entirety of frames, resulting in a significant reduction in computational cost. As a future study, it is proposed to analyze the quantification of the temporal relationship across a variety of scenes to extract patterns regarding frame relevance.

5. Conclusions and Future Work

Physical assaults are a serious issue occurring in all societies worldwide. Apart from causing physical and mental harm to the victims, they also disrupt the functioning of society itself. While multiple solutions have been employed to address this problem, the ultimate barrier is real-time violence detection, where artificial intelligence has proven to be an excellent tool for this task. On the other hand, major organizations such as the European

Union have begun to define and create legislation surrounding trustworthy artificial intelligence, which is based on its legitimacy, ethics, and robustness. Within ethics, one of its components is explainability, which has received significant attention in recent years.

First, this work provides a brief review of the state-of-the-art in video violence detection using artificial intelligence, focusing on the combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM). Also an explanation of what trustworthy artificial intelligence entails, the categorization of explainable artificial intelligence algorithms, and a brief description of the most commonly used algorithms in the literature. Finally, the explanation of how explainability techniques have not been used in video violence detection, although there are algorithms that can provide some intrinsic explainability.

The main objective of this work has been the development of an algorithm for detecting violence in video using artificial intelligence by applying explainable artificial intelligence techniques. For this, the combination of pretrained VGG19 with Bi-LSTM layers has been used, which has been trained on the Hockey Fights, Violent Flow and RWF-2000 datasets in article [58]. The architecture using explainability techniques includes the use of YoloV8 and Frame difference as Key Frame extraction method. On the other hand, GradCAM is applied to each frame of the video, creating a video with 6 different heatmaps superposed with the original video, which correspond to the heatmaps of the last convolutional layer of the five convolutional blocks of VGG19, the sixth subvideo being the average of the five heatmaps.

Finally, a method has been developed to quantify the relevance of a frame within a video for the detection of violence, given the lack of existing methods with this objective. The method consists of the iterative elimination of groups of frames, replaced by a repetition of the last frame. A decrease in accuracy in the prediction with a combination of frames eliminated at a key moment in the video with respect to the original prediction means that those frames are relevant in making the decision.

All in all, the development of an architecture for detecting violence in video with explainable artificial intelligence has been achieved that uses intrinsically explainable methods for Key Frame extraction and that indicates which area and which moment of the video are key for the detection of violence. As future work, an in-depth study is proposed based on the proposed model, classifying the results obtained according to the type of scenes and their characteristics, so that quantifiable conclusions can be drawn from the

explanatory algorithms applied. On the other hand, the use of Skeleton-based algorithms is proposed, since they are based on the detection of violence based on the posture of the people involved, which provides intrinsic explainability to the process.

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] A. Muarifah, R. Mashar, I. H. M. Hashim, N. H. Rofiah, F. Oktaviani, Aggression in adolescents: The role of mother-child attachment and self-esteem, Behavioral Sciences 12 (5) (2022).

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

[3] S. Z. Nurisma, B. Astuti, Peace sociodrama: A strategy to reduce junior high school aggression, International Journal of Social Service and Research 3 (5) (2023) 1319–1324.

[4] A. Enaifoghe, M. Dlelana, A. A. Durokifa, N. P. Dlamini, The prevalence of gender-based violence against women in south africa: A call for action, African Journal of Gender, Society & Development 10 (1) (2021) 117.

[5] O. M. de la Salud, Violence against women (2021).

[6] S. Hillis, J. Mercy, A. Amobi, H. Kress, Global prevalence of past-year violence against children: a systematic review and minimum estimates, Pediatrics 137 (3) (2016).

[7] Crime, safety and victims' rights. fundamental rights survey 2023 (Feb 2023).

[8] M. B. Martínez-González, Y. Turizo-Palencia, C. Arenas-Rivera, M. Acuña-Rodríguez, Y. Gómez-López, V. J. Clemente-Suárez, Gender, anxiety, and legitimation of violence in adolescents facing simulated physical aggression at school, Brain Sciences 11 (4) (2021).

[9] F. Fekih-Romdhane, D. Malaeb, A. Sarray El Dine, S. Obeid, S. Hallit, The relationship between smartphone addiction and aggression among lebanese adolescents: the indirect effect of cognitive function, BMC pediatrics 22 (1) (2022) 735.

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

[11] F. Jing, L. Liu, S. Zhou, J. Song, L. Wang, H. Zhou, Y. Wang, R. Ma, Assessing the impact of street-view greenery on fear of neighborhood crime in guangzhou, china, International journal of environmental research and public health 18 (1) (2021) 311.

[12] H. Yue, H. Xie, L. Liu, J. Chen, 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 (3) (2022) 151.

[13] D. Ding, Z. Ma, D. Chen, Q. Chen, Z. Liu, F. Zhu, Advances in video compression system using deep neural network: A review and case studies, Proceedings of the IEEE 109 (9) (2021) 1494–1520.

[14] J. y. R. E. y. G. S. y. M. A. H. y. N. H. S. Mugunga, Israel y Dong, Un modelo de funciones basado en fotogramas para la detección de la violencia desde cámaras de vigilancia que utilizan la red convlstm, in: 2021 Sexta Conferencia Internacional sobre Imagen, Visión y Computación (ICIVC), 2021, pp. 55–60.

[15] ISO/IEC, Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence, Technical Report ISO/IEC TR 24028:2020 (2020).

[16] K. Abhishek, D. Kamath, Attribution-based xai methods in computer vision: A review, arXiv preprint arXiv:2211.14736 (2022).

[17] Ethics guidelines for trustworthy ai, European Commission (2019).

[18] P. Negre, R. S. Alonso, J. Prieto, A. G. Arrieta, J. M. Corchado, Review of physical aggression detection techniques in video using explainable artificial intelligence, in: International Symposium on Ambient Intelligence, Springer, 2023, pp. 53–62.

[19] F. U. M. Ullah, M. S. Obaidat, K. Muhammad, A. Ullah, S. W. Baik, F. Cuzzolin, J. J. Rodrigues, V. H. C. de Albuquerque, An intelligent system for complex violence pattern analysis and detection, International Journal of Intelligent Systems 37 (12) (2022) 10400–10422.

[20] F. U. M. Ullah, K. Muhammad, I. U. Haq, N. Khan, A. A. Heidari, S. W. Baik, V. H. C. de Albuquerque, Ai-assisted edge vision for violence detection in iot-based industrial surveillance networks, IEEE Transactions on Industrial Informatics 18 (8) (2021) 5359–5370.

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

[22] I. Mugunga, J. Dong, E. Rigall, S. Guo, A. H. Madessa, H. S. Nawaz, 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, 2021, pp. 55–60.

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

[24] A. Jayasimhan, P. Pabitha, 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, 2022, pp. 1–5.

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

[26] B. Lohithashva, V. M. Aradhya, Violent video event detection: a local optimal oriented pattern based approach, in: Applied Intelligence and Informatics: First International Conference, AII 2021, Nottingham, UK, July 30–31, 2021, Proceedings 1, Springer, 2021, pp. 268–280.

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

[28] Y. Su, G. Lin, J. Zhu, Q. Wu, 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, 2020, pp. 74–90.

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

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

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

[32] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, Z. Yang, 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, 2020, pp. 322–339.

[33] P. Negre, R. S. Alonso, J. Prieto, C. N. Dang, J. M. Corchado, Systematic mapping study on violence detection in video by means of trustworthy artificial intelligence, SSRN (March 13 2024). doi:10.2139/ssrn.4757631.
URL `https://ssrn.com/abstract=4757631`

[34] R. Vijeikis, V. Raudonis, G. Dervinis, Efficient violence detection in surveillance, Sensors 22 (6) (2022) 2216.

[35] R. Halder, R. Chatterjee, Cnn-bilstm model for violence detection in smart surveillance, SN Computer science 1 (4) (2020) 201.

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

[37] H. M. B. Jahlan, L. A. Elrefaei, 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 (9) (2021) 8549–8563.

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

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

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

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

[42] S. Sharma, B. Sudharsan, S. Naraharisetti, V. Trehan, K. Jayavel, A fully integrated violence detection system using cnn and lstm., International Journal of Electrical & Computer Engineering (2088-8708) 11 (4) (2021).

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

[44] N. Mumtaz, N. Ejaz, S. Aladhadh, S. Habib, M. Y. Lee, Deep multi-scale features fusion for effective violence detection and control charts visualization, Sensors 22 (23) (2022) 9383.

[45] A. Srivastava, T. Badal, P. Saxena, A. Vidyarthi, R. Singh, Uav surveillance for violence detection and individual identification, Automated Software Engineering 29 (1) (2022) 28.

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

[47] C. Burger, L. Chen, T. Le, "are your explanations reliable?" investigating the stability of lime in explaining text classifiers by marrying xai and adversarial attack, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 12831–12844.

[48] S. Sahay, N. Omare, K. Shukla, An approach to identify captioning keywords in an image using lime, in: 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), IEEE, 2021, pp. 648–651.

[49] A. Bennetot, G. Franchi, J. Del Ser, R. Chatila, N. Diaz-Rodriguez, Greybox xai: A neural-symbolic learning framework to produce interpretable predictions for image classification, Knowledge-Based Systems 258 (2022) 109947.

[50] K. Sahatova, K. Balabaeva, An overview and comparison of xai methods for object detection in computer tomography, Procedia Computer Science 212 (2022) 209–219.

[51] J. Bento, P. Saleiro, A. F. Cruz, M. A. Figueiredo, P. Bizarro, Timeshap: Explaining recurrent models through sequence perturbations, in: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, 2021, pp. 2565–2573.

[52] Q. Liang, C. Cheng, Y. Li, K. Yang, B. Chen, Fusion and visualization design of violence detection and geographic video, in: Theoretical Computer Science: 39th National Conference of Theoretical Computer Science, NCTCS 2021, Yinchuan, China, July 23–25, 2021, Revised Selected Papers 39, Springer, 2021, pp. 33–46.

[53] U. Rachna, V. Guruprasad, S. D. Shindhe, S. Omkar, Real-time violence detection using deep neural networks and dtw, in: International

Conference on Computer Vision and Image Processing, Springer, 2022, pp. 316–327.

[54] Z. Zhang, D. Yuan, X. Li, S. Su, Violent target detection based on improved yolo network, in: International Conference on Artificial Intelligence and Security, Springer, 2022, pp. 480–492.

[55] Z. Islam, M. Rukonuzzaman, R. Ahmed, M. H. Kabir, M. Farazi, Efficient two-stream network for violence detection using separable convolutional lstm, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–8.

[56] S. M. Mohtavipour, M. Saeidi, A. Arabsorkhi, A multi-stream cnn for deep violence detection in video sequences using handcrafted features, The Visual Computer (2022) 1–16.

[57] N. Appavu, et al., Violence detection based on multisource deep cnn with handcraft features, in: 2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET), IEEE, 2023, pp. 1–6.

[58] P. Negre, R. S. Alonso, J. Prieto, P. Novais, J. M. Corchado, Violence detection in video models implementation using pre-trained vgg19 combined with manual logic, lstm layers and bi-lstm layers, papers.ssrn.com (2024). doi:10.2139/ssrn.4832475.
URL `https://doi.org/10.2139/ssrn.4832475`

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

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