Hybrid CNN Xception and Long Short-Term Memory Model for the Detection of Interpersonal Violence in Videos

David Calderon-Vilca
Department of Software Engineering
Universidad Nacional Mayor de San Marcos, Peru
hcalderonv@unmsm.edu.pe

Kent Cuadros-Ramos
Department of Software Engineering
Universidad Nacional Mayor de San Marcos, Peru
kent.cuadros@unmsm.edu.pe

Sergio Valcarcel-Ascencios
Department of Computer Science
Universidad Nacional Mayor de San Marcos, Peru
svalcarcela@unmsm.edu.pe

Igor Aguilar-Alonso
School of Systems Engineering
Universidad Nacional Tecnológica de Lima Sur, Peru
iaguilar@untels.edu.pe

Abstract: *It is common that interpersonal violence is recurrent in public spaces, these are manifested in different ways such as punching, slapping, kicking and pushing, being recorded by video surveillance cameras, these records of images are currently processed by algorithms that are able to detect interpersonal violence, but it is necessary to further improve performance. This paper proposes a hybrid model combining Xception Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) for the detection of violence in videos. We evaluate the effectiveness of our proposal using two datasets: the Hockey fight dataset and real life violence situations dataset. The results showed an accuracy of 93.90% and 98.50% respectively, highlighting that the best performance was achieved with the real life violence situations dataset, comparing the proposed hybrid model with other models proposed in related work, the one we propose shows better performance.*

Keywords: *Image processing, LSTM, patterns of violence, violence detection, xception.*

Received August 1, 2024; accepted June 16, 2025
https://doi.org/10.34028/iajit/22/5/14

1. Introduction

Interpersonal violence occurs very frequently and are direct attacks carried out individually or in groups, these acts take place in public and private spaces. It is estimated that approximately 29.2% of young people have experienced situations of violence, which often go unnoticed by the authorities. This nature of bullying makes it difficult to address, as it occurs in unexpected circumstances and regardless of place or time [36].

During the period from 2000 to 2019, interpersonal violence, in the form of homicide, has emerged as the leading cause of death in America and Latin America. An annual average of 54 515 deaths was recorded, with an age-adjusted rate of around 23,6 per 100000 inhabitants in the young population, where the Andean subregion had the highest rate. It is important to note that the risk of homicide among young men is notably higher, reaching a ratio of 8,1 times that of young women, while among young adults the risk is 2,5 times higher compared to adolescents. Venezuela and Colombia stand out as the two countries with the highest risk of homicide for the young population [25]. The Global Peace Index 2020, elaborated by the Institute for Economics and Peace (IEP), indicates, based on Eurostat data, the incidence of homicides per 100000 inhabitants in 31 European countries, occurred during the period of 2016-2018, highlighting the countries of Belgium, France and Spain with the highest homicide rates. According to the Global Peace Index of mid-2021, the Latin American countries with the highest levels of insecurity include Venezuela, Colombia, Brazil, Ecuador and Peru [13].

According to Instituto Nacional de Estadística e Informática (INEI) [15], in Peru there is a growing implementation of neighborhood-owned surveillance booths, in exclusive areas or popular neighborhoods. 80,6% have a perception of insecurity in the face of a criminal act in the next twelve months. Therefore, it is a demand of the population to install video surveillance systems in the face of insufficient police action against crime; however, a key element is the use of emerging technologies to increase the value of conventional video surveillance systems [14], the perception of citizen insecurity correlates with an increase in the victimization rate of 85,8% in people over 15 years of age. To mitigate interpersonal violence, video surveillance network systems have been implemented to observe behavioral movements in crime prevention or human traffic, recording the events as evidence in situations related to crime or disturbance of order, however, these video surveillance network systems require a lot of manpower and management costs

because the video recording must be continuously monitored in real time to activate it, its records do not identify people or characteristic details [30]. Efforts to maintain effective video surveillance, becomes noticeable when it is intended to cover all sensitive areas such as entrances, exits, offices, corridors, parking lots, border walls, sports fields and their other facilities, as proposed by Wasim *et al*. [33]. However, nowadays the requirements for surveillance have changed due to different factors, which require the use of artificial intelligence systems capable of facial recognition or detecting the aggressor's postural patterns.

With the accelerated advancement of artificial intelligence, technology in the detection of human behavior captured in videos has taken an important step forward, as Wu and Cheng [34] stated in this study, which proposed a framework for recognizing violent activities. In the computer vision subfield of artificial intelligence, data are processed as images in order to find violent traits [24]. There are several proposals that manage to detect violence with artificial neural network algorithms [7, 16, 20], but they focus on finding situations of violence limited to stored videos, when currently it is required to get recognition in videos or video surveillance cameras, thus creating a knowledge gap and the need to strengthen research to find better models and architectures for detecting interpersonal violence.

Vijayakumar *et al*. [31] proposed software elements aimed at automatic detection in contexts of citizen insecurity. These elements comprise voice recognition, postural patterns and natural language processing. This software provides an intelligent surveillance service capable of identifying various situations of insecurity, issuing security alerts and keeping a log of events. However, its effectiveness is limited in uncontrolled environments due to the low quality of features, such as visual noises present during recording. Jebur *et al*. [18] presented a system designed to detect anomalous events in surveillance videos. This system is based on a new multi-scenario violence detection framework that operates in two environments: Multi-venue fights and rugby stadiums, this framework allows the classification of multiple violence scenarios within a single classifier. Moreover, this framework is not limited to violence detection and can be adapted to different tasks. Akash *et al*. [1], Muiruri *et al*. [23], and Vijayakumar *et al*. [31] presented hybrid models that apply different detection techniques for image classification and the extraction of significant features from video frames that allow detecting violent acts in surveillance camera videos with favorable accuracy results, in a previous research by Calderon-Vilca *et al*. [4] the Xception, InceptionV3 and Visual Geometry Group (VGG-16) algorithms were explored together with the Long Short-Term Memory (LSTM) to classify interpersonal violence detection, being considered as a base work to detect violence.

In addition, a hybrid model in image processing combines different neural network architectures to exploit the strengths of each and improve performance on specific tasks, a recent example of the application of this approach is the study by Al-Dulaimi and Kurnaz [2], Convolutional Neural Networks (CNNs) are effective in extracting spatial features from images, while LSTMs capture temporal dependencies in data sequences.

The objective in this paper is to propose a hybrid model combining two algorithms, Xception and LSTM, for interpersonal violence detection in videos. We address this challenge using two datasets, the Hockey fight dataset and the real life violence situations dataset, results are shown and compared with the results of related work, it is evident that the proposed hybrid model offers better performance. Likewise, we consider that this advance is important to implement the analysis process in real time on videos from surveillance cameras, in addition the hybrid model is adaptable to monitoring and intelligent surveillance platforms, offering a more effective approach to detect interpersonal violence in the videos.

2. Related Work

In this section, a review of previous works that are properly related to the detection of human interpersonal violence is made, addressing key aspects such as the application of methods and descriptors, the use of datasets, methods and algorithms.

2.1. Methods and Descriptors for the Detection of Violence in Human Activity

In the research process, different methods and descriptors that aim to contribute to the identification of human violent behaviors were observed. One of the most popular methods is Human Activity Recognition (HAR), especially in extracting video features that capture movement patterns on physical activities of people, by classifying the actions performed by individuals supported by sensory data, which were successfully used according to Gruosso *et al*. [9], Huszar *et al*. [12], and Snoun *et al*. [27]. Multiple ways of achieving extraction were perceived, such as when performing work in color and depth combinatorics, where pattern capture happens by various ways such as background subtraction for human silhouette edges, images with violent movements from a reference background and any other in the contour paradigm, in as much as it uses local binary patterns in rotation and motion flow as a method.

Another way is representation of spatial-temporal actions in videos such as Local Histogram of Oriented Gradients (LHOG) and Local Histogram of Optical Flow (LHOF) are critical for further detection [31]. In that line, Histogram of Optical Flow (HOF) and HOF-Homogeneous (HOF-HOMO) were used to extract

motion features over dynamic RGB images in videos. We also used the Violent Flows (ViF) descriptor and then classified the output using a linear Support Vector Machine (SVM); the accuracy obtained was 81,3%. Using the same classification algorithm combined with HOF [28].

2.2. Datasets Applied towards the Detection of Human Violence

During the study, it was appreciated that the datasets took an important role in sustaining the experimentation phase and training the machine learning models. From the proposals, these used video logs and predefined datasets such as Real-World Web Video Face Recognition (RWF-200), which are recommended for the use of face recognition in web video logs, and which was experimented with models such as: Inception, Xception, VGG-16 and in 4D convolutional models in long-range space-time [19, 21]. Another dataset that was relevant in the study was the ViF through the crowd of violence, images are collected to analyze their context and incidences in crowds of people, to detect interpersonal violence. It was observed that it was integrated with the ViF descriptor by applying information from the optical flow, its datasets are created through the ingestion of data coming from various sources, such as, web services, video security cameras and recordings [3].

Other datasets that also proved interesting were the Kungliga Tekniska Hogskolan (KTH), whose stored images are video sequences of events in a controlled environment for motion recognition, and also the movies fight dataset that is related to human violence in movies or videos that was used by Huszar *et al*. [12] and Kang *et al*. [19]. Overall, the presented datasets not only contributed to the discovery of human violence patterns but were contributed to the effectiveness of the algorithms.

The Hockey dataset, distinguished by its ability to use scenarios, lighting conditions and camera angles, this dataset collects the fights of the national Hockey league games in a thousand video clips, they are balanced in 500 violent and 500 non-violent videos, at a rate of 360 by 288 frames fully balanced between videos with or without violence [37]. Another dataset used in this article, is the Hockey fight, which results from a collection of 1000 of fragmented videos or clips, at this point the Real-World Fighting dataset (RWF) is used, considered the most complete, however, they do not meet technical criteria to form by themselves a dataset for violence detection, that is why they are combined with other datasets such as Hockey fight that apply evaluation models such as Semi-Supervised Hard-Attention (SSHA), in order to achieve the expected results [22]

Throughout this study, it is observed that balanced data are used in the datasets. This means that classes such as “violent” and “non-violent” are represented in equal amounts. So, in practice, the number of samples (videos, images or sequences) for each class is approximately the same, which helps to avoid biases in the training of the machine learning models.

The Movies dataset is balanced and contains 200 videos conveying fight and non-fight scenes. They also introduce the Hockey fight dataset. Finally, the crowd dataset with 246 balanced videos is characterized by being used in real time. In common they used k-fold validation for the training and testing stage [28].

The Smart-City CCTV Violence Detection (SCVD) dataset is a database, a highlight feature of it is recognizing any offensive object that causes harm to people or property. In this context, the detection discriminates three possibilities, violent action, non-violent action and violence using weapons [32].

Another way was the construction of the surveillance camera fight dataset, which was formed from videos of fights taken from YouTube, the curious thing about this figure is that it mixes different forms of violence in light and shadow conditions, this dataset adds up to 300 videos and is balanced in the ratio 1:1, between videos with and without violence. For each group the authors trained with 80% of the videos and the difference for testing, continuing in this study, the Automatic Violence Detection (AVD) dataset contains 350 videos, of which 65.7% were characterized as violent and the difference as non-violent, it is prepared to face complex scenarios. Finally, the RWF-2000 dataset, also a collection of 2000 clips retrieved from video surveillance cameras and streamed from YouTube, is balanced into violent and non-violent actions [3].

Sun *et al*. [29] considered using public datasets. The first one, known as University of California, San Diego (UCSD) anomaly detection dataset, consists of 16 videos making a total of 6550 frames, 2550 frames were allocated for training and 2010 were used for testing, the second dataset called Chinese University of Hong Kong (CUHK) avenue dataset, was formed with 16 training videos and 21 videos for testing.

2.3. Hybrid Models and Algorithms Relevant for Interpersonal Violence Detection

The study reveals that CNN models have significant presence to obtain results detecting human violent actions, the most used CNN models are 2D, 3D and even Four-Dimensional Convolutional Neural Network (4DCNN), the 2DCNN, process two-dimensional data (length and width), such as images and videos, to extract relevant features from the images and then feed those features through successive connected layers and achieve the classification of objects in data frame along the video [21]. The Inception and Xception models, are clear examples of feature extraction with efficiency by design. Also, CNNs have been hybridized with LSTM to classify violence between people [17, 19]. On the

other hand, beyond violence detection, some proposals develop and implement violence detection in video surveillance [9, 11, 12, 21] that can be used in conglomerated public places.

CNN network variants, such as Alexnet, enable the extraction of datagram features in images as encoders, adding VGG-13 as a decoder [21]. They also experimented on dorsal CNN networks such as SqueezeNet, MobilNets and EfficientNet, hybridizing LSTM and Xception and Inception as frameworks for comparison of results, based on their own neural model [19]. Added to this is MobilNet, as a key element for feature extraction in CNN models [11].

It was also noticed that there is a proportionality between the temporal information and the growth of the sequence length, in acquiring an attractive accuracy rate on the 3DCNN model and its computational time, this due to the strategy of its own design [11], another aspect that is interesting is to modify a standard CNN network to insert a neural network model (MR CNN). Datasets with labels are adapted in the process of human violence detection used 3DCNN models [12]. In general, these models rely on learning algorithms such as SVM, Random Forest (RF), clustering and decision networks. The use of Rectified Linear Unit (ReLU) as an activation function was observed in the study.

At other times, images are taken and are diversified into training and test data, they used k-folds, these are forwarded to the ResNet50V2 model, compared to another technique Discrete Wavelet Transform (DWT), Principal Component Analysis (PCA), VGG-16, and VGG-19, the latter two are convolutional models that were used in relation to other architectures hybrids as LSTM and Inception and supported with algorithms such as Gated Recurrent Units (GRUs) [28].

The mix or hybrid models between Inception V3 and the typical CNN convolutional network architecture potentially strengthened the detection of fights through complex patterns and spatial dependencies in the image data, it can be seen in [31] the application of the 2DCNN algorithm, it is an alternative for the improvement in the pre-processing of the data, supported with the use of You Only Look Once (YOLO) for the extraction of skeletons from the images. YOLO-Pose was strategically used to recognize key frames in the recreation of the human skeleton by combining the 2DCNN network in real time by inserting weighting modules for each frame in relation to the RGB scale and the skeleton [37]. There is a symbiosis between deep learning architectures such as 3DCNN and pre-trained models are vital in the spatiotemporal relationship for pattern recognition in motion and temporal connections in discriminating violent and non-violent actions [8].

On the other hand, in the tests the training results are taken and classified by deep extraction. The application of GoogleNet, which represents a CNN based on Inception, allows to extract significant features from the image by various filters with max-pooling layers with stride of 2 to reduce the resolution of the network [3]

The proposed Appearance and Movement Features Cross-Fusion Block Memory-Network (AMFCFBMem-Net), is able to extract and recognize complex movements and decode images in RGB video, solving the over generalization in deep learning [29].

Xu *et al*. [35] proposed the conjunction of the Convolutional Block Attention Module (CBAM) enhanced by Gaussian Context Transformer (GCT) and the introduction of the ConverGed Attention Module (GSAM), which reduces dimensionality and improves image resolution, gaining higher person recognition and better management in occlusion processing. In another case, the use of the SSHA model, was visualized. It employs a strategy that optimizes vector spaces for a better result in the recognition of violent or non-violent actions. It works with two assumptions. First, the use of the dataset meets high data quality requirements. Second, it focuses on eliminating redundant elements in the neural network input [22].

3. Design of the Hybrid CNN Xception and Long Short-Term Memory Model for the Detection of Interpersonal Violence in Videos

The proposed hybrid model for the identification of interpersonal violence in videos is composed of three parts: data preprocessing, feature extraction and classification. In the first part, from the data preprocessing, image sequences are obtained from a video format file that is recorded in real time, simultaneously converting them to a format compatible with CNN Xception; in the second part, the features that discriminate violence or non-violence in the image sequences that pass through CNN Xception are extracted; in the third and last part, the important features obtained from the image sequences are treated as inputs for the LSTM algorithm by which spatial-temporal features are obtained. Next, the neural network architecture defines two hidden layers and an output layer that has 2 neurons with softmax activation function, managing to determine the acts of violence. Figure 1 details the components of the model.

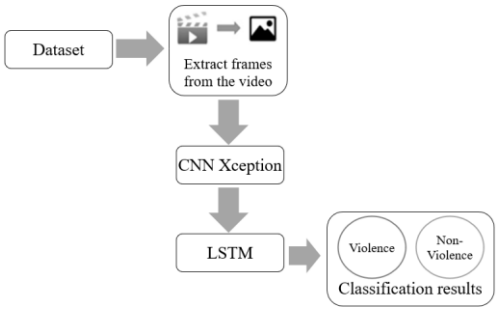


Figure 1. CNN Xception and LSTM hybrid model for detecting interpersonal violence in videos.

3.1. Dataset Collection

In the design of the model to analyze the image sequences, two video datasets were considered such as the Hockey fight dataset [26] and real life violence situations dataset taken from the repository [6]. These datasets were selected due to videos containing real-world characteristics, with a diversity of individuals in terms of race, gender, and age, as well as a variety of settings with attributes that make it possible to identify interpersonal violence.

1000 videos from the Hockey fight dataset with an approximate duration of 5 seconds were used, of which 500 were labeled as violence (marked with "V_") and 500 as non-violence (NV_) and were found to be balanced between the violence and non-violence classes. In the second case, 2000 videos from the real life violence situations dataset with a duration of approximately 5 seconds were used, of which 1000 were identified as violence (labeled with "V_") and 1000 as non-violence (NV_), also being considered balanced data.

The Hockey fight dataset contains violence situation activities such as fights and scenes showing physical confrontations between individuals in sports games; while the real life violence situations dataset contains physical violence such as physical altercations between people seen in the videos such as punching, pushing, physical aggression in the streets or public spaces, commercial establishments, including domestic fights. Both datasets contain activities of violence and non-violence situations, covering most scenarios of interpersonal physical violence.

Figure 2 show images sequences extracted from Hockey fight dataset, the same way in the Figure 3 show images sequences extracted from the real life violence dataset, containing violence and non-violence scenarios. In the case of scenes that present acts of violence, these are expressed in duration of few seconds in each video, on the other hand, in the non-violent videos, the content denotes non-threatening expressions such as the subject playing the violin. Image preprocessing techniques have allowed us to extract sequences, and to standardize the size of the images according to the algorithm used in the experiment.

a) Violence.

b) Non-violence.

Figure 2. Extracted sequences from the Hockey fight dataset [26].

a) Violence.

b) Non-violence.

Figure 3. Real life violence situations dataset [6].

3.2. Preprocessing

In the video preprocessing stage, the presence of illumination noise generated by abrupt changes in lighting conditions and motion noise caused by vibrations or rapid camera movements has been

identified. To attenuate illumination noise, histogram equalization has been applied to improve contrast in images with uneven illumination; to mitigate motion noise, temporal averaging of frames has been applied to smooth abrupt changes in the image.

On the other hand, spatial filtering has been applied with the median filter, Gaussian filter effective to remove Salt and Pepper noise and the one used to smooth intensity variations without compromising the edges of the objects. These techniques contributed to improve the clarity of the videos, facilitating their analysis and the subsequent detection of relevant events.

In particular, to mitigate Salt and Pepper, noise, characterized by the random appearance of scattered white and black pixels, the Adaptive Median Filter (AMF), which dynamically adjusts the size of the filtering window according to the noise density in the image, was applied.

3.3. Extract Frames from Video

Using both datasets, the extraction of the video data frames is performed by taking the first 20 image sequences of each video. Then, the size of the image sequences is resized to 299*299*3, an acceptable size for the CNN Xception model, then the "shuffle data" random reordering process is applied, which consists of shuffling the data to prevent the model from falling into overtraining. In Figure 4, an extraction of a video from the real life violence dataset [6] is shown, where it can be clearly observed that there is in each of the frames the state of violence, where one person is grabbing another by the neck.

Figure 4. Sequence of images extracted from the video.

3.4. Feature Extraction with CNN Xception

The CNN Xception architecture proposed by Chollet [5], implemented in three layers is based on depthwise separable convolution, which reduces computational processes compared to its predecessor Inception model and offers fast convergence in training. The first entry flow layer contains 2D convolutional layers that extract features such as edges, textures and colors, the internal separable convolutions use the ReLU activation function and apply max-pooling to represent the important features in less size. The second layer middle flow is mainly composed of residual blocks, each residual block includes depth separable convolutions and point-to-point convolutions, followed by batch normalization and ReLU activation function. The residual connections in these blocks allow information to be passed directly through the layers unchanged, which helps mitigate the gradient fading problem and

extracts deep features such as shapes, structures and objects. The third layer exit flow extracted is condensed and transformed into the classification output which includes separable convolutional layers for extraction of late features such as object variations in different frames, average global pooling is applied significantly reducing the amount of data to be processed avoiding overfitting. Then flattening is applied before the final output and reaching the two-dimensional feature maps produced by the convolutional layers and through global average pool is converted to a one-dimensional vector, which enables classification by the softmax activation function. From Figure 1, the CNN Xception is the main component, Figure 5 shows the application of the CNN Xception, which provides the inputs of preprocessed images with size of 299x299 in RGB color format, the introduction of 20 frames in each group of previously extracted images is performed, as a result we obtain violence features expressed in vector of 20x2048.

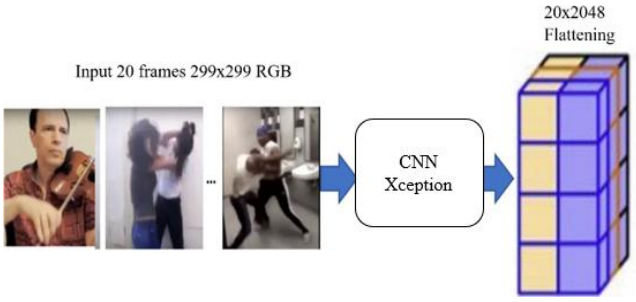


Figure 5. Application of CNN Xception in feature extraction.

Figure 6. Extraction of features as edges in the frames of a CNN Xception video.

In the entry flow layer, it performs convolutions that extract basic features of violence expressed in image edges of contact between people, textures and colors of the characters found in the sequence of images, as can be seen interpersonal violence in Figure 6, in this layer the ReLU activation function is implemented, which allows introducing the non-linearity that takes violence attributes and suppresses non-violence attributes, in addition, using the max-pooling function allows reducing the size of the grouped images, preserving more relevant aspects in a space of reduced dimensions

In the middle flow layer, shown in Figure 7, its input are the feature maps coming from the entry flow layer, residual blocks are controlled and managed, depth separable convolutions and point-to-point convolutions are applied to extract complex patterns based on filters.

batch normalization that stabilizes learning is also performed, the activation function used is ReLU that highlights the most important features of violence and discards the irrelevant ones. As output, it provides complex feature maps which are more detailed representations of shapes and structures associated with interpersonal violence.

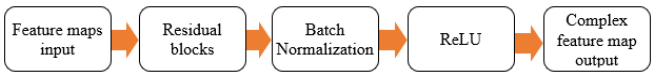


Figure 7. Complex features extraction in the middle flow layer of CNN Xception.

In the last exit flow layer finer features are extracted as variations of the same object or figure containing violence expressed in a different context, as shown in Figure 8-a) is a frame with violence captured in a video, then when another video is processed i.e. another file in this case sub Figure 8-b) a frame that also has violence, comparing both is the same scenario of violence between the two people taken with different approaches in the video shot. The work of CNN Xception is the application of average global pooling that generates as a result a vector of 2048 transfer values for each frame, being 20 sequences of images in each batch, expressed in 20x2048 transfer values. Finally, the flattening function converts the two-dimensional features to a one-dimensional vector, which is optimal for the input of the fully connected layer that comes configured with softmax. Considering that CNN Xception in its last fully connected layer allows switching with another algorithm, in this case we connect with the LSTM algorithm for its spatiotemporal feature extraction, finally we look for the classification of the presence and absence of violence, which is further substantiated.

a) Violence depicted in a scene from one video. b) Violence depicted in another scene from another video.

Figure 8. Finer feature extraction in the exit flow layer of CNN Xception.

3.5. Extraction of Spatiotemporal Features with Long Short-Term Memory

Considering that CNN Xception allows connecting with another algorithm on its last layer to achieve classification, in this case in the last layer of CNN Xception we hybridize connecting with the neural network algorithm LSTM, Figure 9 shows the hybrid model between CNN Xception and LSTM, for this we considered the vector of 20x2048 transfer values that generated the last fully connected layer of CNN Xception, in this case being the input for the LSTM the vector of 20x2048.

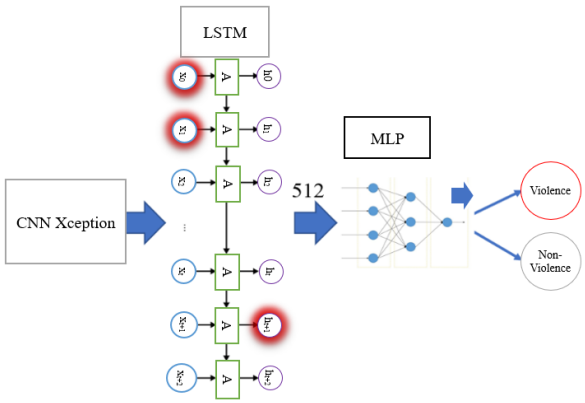


Figure 9. Hybridization between CNN Xception and LSTM.

In the implementation of the LSTM algorithm, in terms of hyperparameters, initially tests have been performed with 50, 100 and 150 epochs, observing that the model continued to improve its accuracy with no obvious signs of overfitting. Upon reaching 200 epochs the accuracy in the validation sets stabilized, the loss function showed very little change giving a constant convergence, further with 250 epochs and 300 epochs, no significant improvements in accuracy resulted, which indicated that 200 epochs was an optimal value to achieve a good balance for detecting violence in the videos. Regarding batch size, batch sizes of 128 and 256 were evaluated, resulting in larger fluctuations in the loss function, in addition to greater variability in accuracy during validation, increasing the batch size to 500, convergence and greater stability in the loss function was observed without compromising the model's ability to generalize. The Adaptive Moment Estimation (Adam) optimizer is widely used in LSTM neural network training due to its efficiency and effectiveness in convergence, it combines the advantages of the Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp) methods, allowing dynamic adjustments of the learning rates for each parameter. In a recent study by Gudla and Bhoi [10] investigated the impact of different learning rates when using the Adam optimizer in LSTM models, the results indicated that they allow to achieve better performance in terms of accuracy and stability in convergence.

As for the LSTM architecture, it was configured with 512 neurons for temporal feature extraction from image sequences derived from videos. The input 20x2048 was set, in this case 20 represents the number of image sequences taken from the videos and 2048 denotes the dimension of the vector containing the transfer values obtained in CNN Xception.

In LSTM each time step is a frame with violence content, capturing the temporal information and dependencies between frames achieves a dense 512-

dimensional layer of unique features. As a result, the LSTM algorithm generates a 512-dimensional vector with the unique spatiotemporal features that have violence and non-violence patterns in the image sequences.

3.6. Classification of Violence and Non-Violence

Considering Figure 10, to achieve the classification we have coupled the Multi-Layer Perceptron (MLP) neural network, it has as input 512 spatiotemporal features coming from the LSTM output, in the intermediate layers we consider 2 hidden layers each one with its own activation function, and finally an output layer.

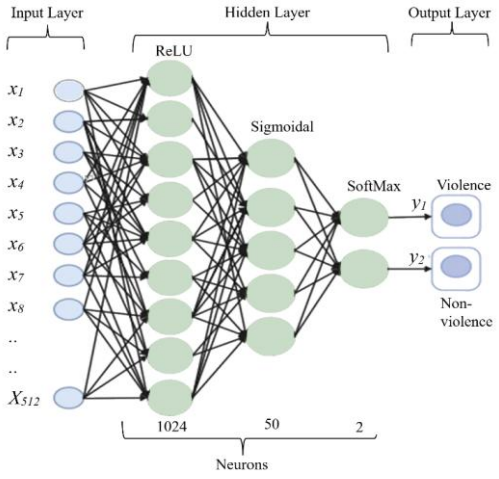


Figure 10. Fully connected MLP architecture.

Figure 10 shows the MLP architecture, the number of neurons in the first fully connected hidden layer was set to 1024 neurons each with the ReLU activation function preferred for the ability to maintain the gradient flow through its process. If the value is greater than zero, it returns the violence feature otherwise it sets it to zero directing that there is no violence. In the second hidden layer 50 neurons were designated, each of them with the sigmoidal activation function, which allows the model to interpret the activation of the neurons as the probability of the existence of violence in a specific region. The last layer is output with 2 neurons each with the softmax activation function, the first neuron expresses the probability of existence of violence and the second neuron expresses the probability of non-existence of violence, of both results the one with the higher probability is labeled as the final response of the model.

4. Results and Discussion

4.1. Dataset

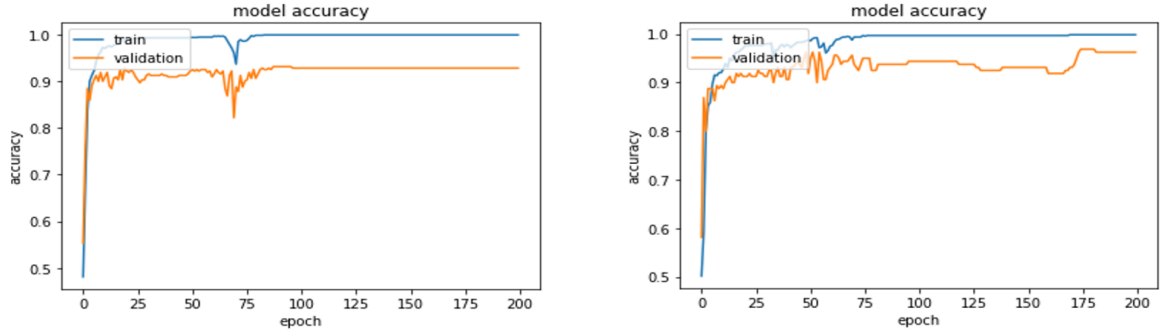
For testing and validation purposes, two different models were used. The first model was trained with Hockey fight dataset and the second model was trained with real life violence situations dataset. Table 1 shows the distribution of the two datasets, in order to evaluate their performance. Each dataset is divided into two parts, 80% for training and 20% for testing. In the first part, the training, the validation is included, that value was again distributed in two parts, 80% for training and 20% for validation. During training, the cross-validation technique was used to evaluate performance, breaking it down into subgroups of 10 folds to determine its validity. Finally, in the second part, the model is tested with the elements that did not participate in the training.

Table 1. Distribution of the dataset for training, validation and testing.

Data set	Training quantity (cross validation)				Training quantity (cross validation)		Total
	For training		For validation		Violence	Non-violence	
	Violence	Non-violence	Violence	Non-violence			
For first model with Hockey fight dataset	320	320	80	80	100	100	1000
For second model with real life violence situations dataset	640	640	160	160	200	200	2000

4.2. Training of the Hybrid Model

The training with the Xception model is shown in Figure 11, in general with the Hockey fight dataset, the training and validation values are shown in Figure 11-a). In its evolution it starts from its first epochs with a validation accuracy of 0.500, then when reaching epoch 40 it reveals an accuracy of 0.930, noting an accuracy stability of 0.935 when reaching epoch 200. Note that the training and validation curves are similar during the execution time.



a) Training of the first model with Hockey fight dataset.

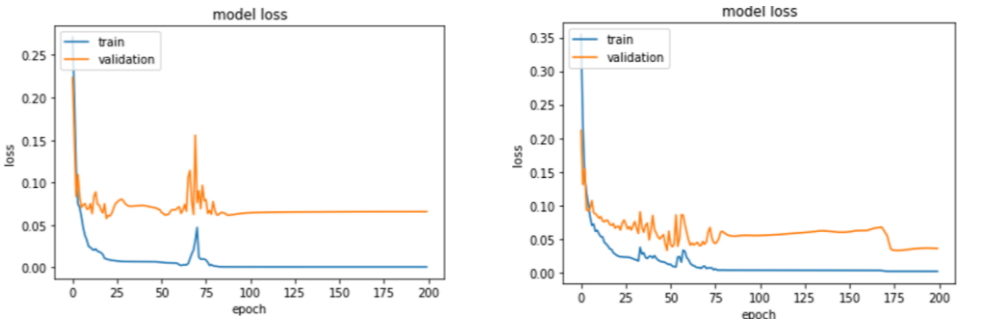
b) Training of the second model with real life violence situations dataset.

Figure 11. Results of training and accuracy in the detection of interpersonal violence.

Similarly, in Figure 11-b) with the real life violence situations dataset, the training and validation values are visualized.

In its evolution, it starts from its first epochs with a validation accuracy of 0.550, then when reaching epoch 40 it achieves an accuracy of 0.970 showing a significant change, then we can notice a stability of the accuracy of 0.9801 when reaching epoch 200. In both models it indicates that it is generalizing the patterns of violence and nonviolence correctly, in addition, in both models the linear trend is shown indicating that the more epochs of training would change very little, therefore, it is not necessary more epochs of training.

Regarding the loss function that is configured in the Xception model for training, it seeks to improve the learning of the model by reducing the error, it also reflects the distance between the classifications of the model with respect to the real labels of violence and non-violence, Figure 12-a) shows the loss function of the first training model with Hockey fight dataset and Figure 12-b) shows the loss function of the second training model real life violence situations dataset.



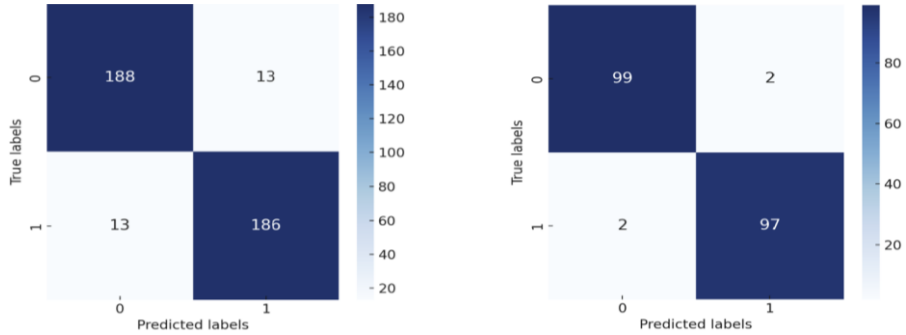
a) Loss function in training with Hockey fight dataset. b) Loss function in training with real life violence situations dataset.

Figure 12. Loss function in the validation of the training model.

In the training of both models, it can be appreciated that at the beginning the values of errors are higher, revealing that they do not manage to correctly classify the violence, then the errors are decreasing in each epoch, managing to stabilize in the last epochs. In the first model with the Hockey fight dataset note that in epoch 80 approximately the loss is less and less, in this case it shows us that their training is being stable until epoch 200, moreover, becoming more and more linear, suggesting that the model learned and generalized the characteristics of violence, finally its loss value was 5.88%. As for the second model with real life violence situations dataset up to epoch 75 shows that it manages to reduce much of the error, although during the following epochs it has ups and downs, from epoch 175 onwards it stabilizes with a linear trend, achieving its loss value of 2.00%. Finally, the behavior of the last epochs of both of them shows that it learns little or stops learning, suggesting that more training is not necessary.

4.3. Hybrid Model Testing and Evaluation Metrics

According to Table 1, the first model was tested with 20% of Hockey fight dataset through the confusion matrix, in Figure 13-a) we can appreciate the classification results of the first model, where the True Positives (TP) are 188 that correctly classified the videos containing violence, the False Positives (FP) 13 are those that incorrectly classified videos that actually do not contain violence, but classified as violence. True negatives (TN) 186 are those that were correctly classified videos that did not actually contain violence were classified as non-violence. False Negatives (FN) 13 are those that were incorrectly classified videos that actually contain violence, but the model classified as non-violence.



a) Result of the classification with Hocky fight dataset. b) Result of the classification with real life violence situations dataset.

Figure 13. Confusion matrix for the evaluation of the model with metrics.

As for the second model, it was tested with the 20% real life violence situations dataset, evaluating with the confusion matrix, Figure 13-b) shows the classification results, where the TP are 99 that were correctly

classified videos containing violence, the FP are 2 that were incorrectly classified videos that actually do not contain violence but classified as violence, TN are 97 that were correctly classified as videos that do not actually contain violence but were classified as non-violence, FN are 2 that were incorrectly classified as videos that actually contain violence but were classified as non-violence by the model.

Regarding the result of the confusion matrix with Hockey fight dataset, accuracy metric was used giving a result of 93.5%, in our context it tells us the proportion of correct classifications of violence and non-violence. Another metric is recall, providing a probability of 93.47%, where the model classifies showing evidence of violence in the videos. Next, the precision metric provides a probability of 94.16% that correctly predicts the existence of violence in a video when it is present. The model achieved an F1 score of 93.72%, indicating a solid balance between precision and recall when detecting violence. It also displayed an Area Under the Curve- Receiver Operating Characteristic (AUC-ROC) of 0.97, demonstrating an excellent ability to distinguish between videos with and without violence. Consequently, it can be seen that there is no overfitting, since the curves obtained in training and validation are similar.

Regarding the balanced life violence situations dataset, it has been measured with the accuracy metric, in this case the test gave a result of 98.50%, which informs us of a proportion of correct classification of violence and non-violence. Another metric, the recall gives us a 97.97% probability of correctly identifying violence in a video when it is present. Precision gives us a 98.11% probability that the model accurately predicts the existence of violence in a video when it actually occurs. The model achieved an F1 score of 98.01%, demonstrating a balance between precision and recall in violence detection. It also showed an AUC-ROC of 0.98, confirming is extremely high ability to distinguish between violent and non-violent videos.

5. Discussion

In this section, the findings are interpreted, contrasted with previous studies and their practical implications are discussed. The results obtained in this research demonstrate the effectiveness of violence detection models applied to videos with violent content.

Table 2 presents different research, with their respective models used, as well as the datasets and metrics used with their results.

Table 2. Comparison of the results of the models and datasets used versus other research.

Authors	Model or algorithm	Dataset	Metric	Result
Zhang *et al*. [37]	RTPNet (YOLO-Pose+2DCNN ACTION-Net)	RWF-2000	Accuracy	93.30%
Zhang *et al*. [37]	RTPNet (YOLO-Pose+2DCNN ACTION-Net)	Surveillance camera fight	Accuracy	93.40%
Mohammadi and Nazerfard [22]	Semi-supervised hard attention model (I3D backbone)	RWF	Accuracy	90.40%
Mohammadi and Nazerfard [22]	Semi-supervised hard attention model (I3D backbone)	Hockey	Accuracy	98.70%
Sumon *et al*. [28]	Pre-trained CNN+Transfer learning	YouTube dataset (Bangladesh-specific)	Accuracy	95.67%
Wankhade *et al*. [32]	Deep learning model (MATLAB-based)	SCVD	Accuracy	96.4%
Huszar *et al*. [12]	Smart networks with 3D convolutions	CV-1500	Accuracy	92.00%
Huszar *et al*. [12]	Smart networks with 3D convolutions	RWF-2K-1500	Accuracy	88.25%
Kang *et al*. [19]	MSM+Mobilnet V3	RWF-2000	Accuracy	90.00%
Kang *et al*. [19]	MSM+EfficientNet-Bo	RWF-2000	Accuracy	92.00%
Our research	**Xception+LSTM**	**Hockey fight dataset**	**Accuracy**	**93.90**
Our research	**Xception+LSTM**	**Real life violence situations dataset**	**Accuracy**	**98.50**

When analyzing the models presented in the state of the art, there is a strong presence of advanced architectures, such as Real-Time pose-based Network (RTpNet) (YOLO-Pose+2DCNN ACTION-Net) proposed by Zhang *et al*. [37], neural networks with 3D convolutions [12], and hybrid models such as Motion Saliency Map+Mobilnet V3 (MSM+Mobilnet V3) [19]. Likewise, the use of models with semi-supervised attention [22], evidences an interest in complex structures capable of capturing spatial and temporal relationships. Against these proposals, our study employs the combination of Xception with LSTM, integrating an efficient convolutional network with a recurrent network, allowing robust processing of both spatial and temporal features. This approach proves to be competitive with sophisticated models, standing out for its simplicity and effectiveness in image and video violence detection tasks.

The analyzed studies use several specific datasets, such as RWF-2000 and surveillance camera fight [37], RWF and Hockey [22], YouTube dataset [28], and the SCVD [32]. These datasets reflect violence scenarios in controlled or urban surveillance environments. In comparison, our research employs the Hockey fight dataset and the real life violence situations dataset, which present more diverse and realistic scenes, including variability in recording environments and conditions. This diversity presents a greater challenge for model generalization. Therefore, the use of these datasets in our study evidences an effort to validate model performance in less structured contexts closer to real-life situations.

The results reported in the state-of-the-art show high levels of accuracy, reaching values such as 98.70% [22] and 96.4% [32], with a general range between 88.25% and 98.70%. In this context, our model, based on Xception+LSTM, achieves 93.90% accuracy in the Hockey fight dataset and an outstanding 98.50% in the real life violence situations dataset.

These results demonstrate that our approach is not

only competitive, but also outperforms state-of-the-art work in some cases, especially when handling more realistic situations. The ability to maintain high performance on complex datasets validates the effectiveness and robustness of the proposed model against recognized architectures [12, 19, 37].

6. Conclusions and Future Work

For the detection of interpersonal violence, we propose the hybrid CNN Xception and LSTM model, in which, during the experiment, two datasets were used: Field Hockey fight dataset and real life violence situations dataset. The former achieved an accuracy of 93.90%, while the latter achieved an accuracy of 98.50%. After comparing both, it was determined that using the real life violence situations dataset showed a superior result, then when compared with other research, it again confirms better performance in the detection of interpersonal violence for real-time videos. Consequently, we recommend our hybrid model to be used as an artificial intelligence engine in video surveillance platforms for interpersonal violence detection.

For future work, it is suggested to investigate violence with serious injuries captured in real time through videos. It is also important to explore the classification of different types of violence present in the videos, as this area of research is growing in the scientific community.

Acknowledgment

This research was supported by Universidad Nacional Mayor de San Marcos-Resolucion Rectoral N. 005753-2021-R/UNMSM and Project number C21201361-PCONFIGI 2021.

References

[1] Akash S., Moorthy R., Esha K., and Nathiya N., "Human Violence Detection Using Deep Learning Techniques," *in Proceedings of the 8th International Virtual Conference on Biosignals, Images, and Instrumentation*, Online, pp. 1-12, 2022. DOI 10.1088/1742-6596/2318/1/012003

[2] Al-Dulaimi O. and Kurnaz S., "A Hybrid CNN-LSTM Approach for Precision Deepfake Image Detection Based on Transfer Learning," *Electronics*, vol. 13, no. 9, pp. 1-22, 2024. https://doi.org/10.3390/electronics13091662

[3] Basavaraj G. and Kodli Post S., "Violence Detection in Real Life Videos Using Pre-Trained Models," *International Journal of Creative Research Thoughts*, vol. 12, no. 10, pp. 670-670, 2024. https://www.ijcrt.org/papers/IJCRT2410662.pdf

[4] Calderon-Vilca H., Ramos K., Quiroz E., Rojas J., Vilca R., and Tarqui A., "The Best Model of Convolutional Neural Networks Combined with LSTM for the Detection of Interpersonal Physical Violence in Videos," *in Proceedings of the 29th Conference of Open Innovation Association*, Tampere, pp. 81-86, 2021. DOI: 10.23919/FRUCT52173.2021.9435563

[5] Chollet F., "Xception: Deep Learning with Depthwise Separable Convolutions," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, pp. 1800-1807, 2017. DOI: 10.1109/CVPR.2017.195

[6] Elesawy M., Hussein M., and Abd El Massih M., Real Life Violence Situations Dataset, 1000 Videos Containing Real Street Fight and 1000 Video from other Classes, https://www.kaggle.com/datasets/mohamedmustafa/real-life-violence-situations-dataset/data, Last Visited, 2024.

[7] Febin I., Jayasree K., and Joy P., "Violence Detection in Videos for an Intelligent Surveillance System Using MoBSIFT and Movement Filtering Algorithm," *Pattern Analysis and Applications*, vol. 23, no. 2, pp. 611-623, 2020. https://doi.org/10.1007/s10044-019-00821-3

[8] Gangu. and Bhadrashetty A., "Violence Detection in Real Life videos Using Pre-Trained Models," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 6, no. 6, pp. 1825-1830, 2024. https://www.doi.org/10.56726/IRJMETS59050

[9] Gruosso M., Capece N., and Erra U., "Human Segmentation in Surveillance Video with Deep Learning," *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 1175-1199, 2021. https://doi.org/10.1007/s11042-020-09425-

[10] Gudla S. and Bhoi S., "A Study on Effect of Learning Rates Using Adam Optimizer in LSTM Deep Intelligent Model for Detection of DDoS Attack to Support Fog Based IoT Systems," *in Proceedings of the 1st International Conference Computing, Communication and Learning*, Warangal, pp. 27-38, 2022. DOI: 10.1007/978-3-031-21750-0_3

[11] Hussain A., Muhammad K., Ullah Hayat., Amin Ullah., and et al., "Anomaly Based Camera Prioritization in Large Scale Surveillance Networks," *Computers, Materials and Continua*, vol. 70, no. 2, pp. 2171-2190, 2022. https://doi.org/10.32604/cmc.2022.018181

[12] Huszar V., Adhikarla V., Negyesi I., and Krasznay C., "Toward Fast and Accurate Violence Detection for Automated Video Surveillance Applications," *IEEE Access*, vol. 11, pp. 18772-18793, 2023. DOI: 10.1109/ACCESS.2023.3245521

[13] Institute for Economics and Peace, Global Peace Index: Measuring Peace in a Complex World, https://www.visionofhumanity.org/wp-content/uploads/2021/06/GPI-2021-web-1.pdf,

Last Visited, 2024.
[14] Instituto Nacional de Estadistica e Informatica (INEI), Estadísticas de Seguridad Ciudadana, https://www.gob.pe/institucion/inei/colecciones/6094-estadisticas-de-seguridad-ciudadana?sheet=2, Last Visited, 2024.
[15] Instituto Nacional de Estadistica e Informatica (INEI), Victimizacion en el Peru 2010-2019, Principales Resultados, https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1730/Libro.pdf, Last Visited, 2024.
[16] Janbi N., Ghaseb M., and Almazroi A., "ESTS-GCN : An Ensemble Spatial-Temporal Skeleton-based Graph Convolutional Networks for Violence Detection," *International Journal of Intelligent Systems*, vol. 2024, no. 1, pp. 1-19, 2024. https://doi.org/10.1155/2024/2323337
[17] Jaouedi N., Boujnah N., and Bouhlel M., "A Novel Recurrent Neural Networks Architecture for Behavior Analysis," *The International Arab Journal of Information Technology*, vol. 18, no. 2, pp. 133-139, 2021. https://doi.org/10.34028/iajit/18/2/1
[18] Jebur S., Hussein K., Hoomod H., and Alzubaidi L., "Novel Deep Feature Fusion Framework for Multi-Scenario Violence Detection," *Computers*, vol. 12, no. 9, pp. 14475-14482, 2023. https://doi.org/10.48084/etasr.7270
[19] Kang M., Park R., and Park H., "Efficient Spatio-Temporal Modeling Methods for Real-Time Violence Recognition," *IEEE Access*, vol. 9, pp. 76270-76285, 2021. DOI: 10.1109/ACCESS.2021.3083273
[20] Khan M., El Saddik A., Gueaieb W., De Masi G., and Karray F., "VD-Net: An Edge Vision-based Surveillance System for Violence Detection," *IEEE Access*, vol. 12, pp. 43796-43808, 2024. DOI: 10.1109/ACCESS.2024.3380192
[21] Magdy M., Fakhr M., and Maghraby F., "Violence 4D: Violence Detection in Surveillance Using 4D Convolutional Neural Networks," *IET Computer Vision*, vol. 17, no. 3, pp. 282-294, 2023. https://doi.org/10.1049/cvi2.12162
[22] Mohammadi H. and Nazerfard E., "Video Violence Recognition and Localization Using a Semi-Supervised Hard Attention Model," *Expert Systems with Applications*, vol. 212, pp. 118791, 2023. https://doi.org/10.1016/j.eswa.2022.118791
[23] Muiruri S., Okong'o M., and Mwathi D., "Enhancing Public Safety through Advanced Video Analysis: A Conv-LSTM-SVM Model for Violence Detection in Surveillance Footage," *East African Journal of Information Technology*, vol. 7, no. 1, pp. 202-214, 2024. https://doi.org/10.37284/eajit.7.1.2117
[24] Rendon-Segador F., Alvarez-Garcia J., Salazar-Gonzalez J., and Tommasi T., "CrimeNet: Neural Structured Learning using Vision Transformer for Violence Detection," *Neural Networks*, vol. 161, no. 1, pp. 318-329, 2023. https://doi.org/10.1016/j.neunet.2023.01.048
[25] Sanhueza A., Caffe S., Araneda N., Soliz P., San Roman-Orozco O., and Baer B., "Homicide among Young People in the Countries of the Americas," *Pan American Journal of Public Health*, vol. 47, pp. 1-11, 2023. https://doi.org/10.26633/RPSP.2023.108
[26] Shrief Y., Hockey Fight Dataset, Fight and Non-Fight Videos, https://www.kaggle.com/datasets/yassershrief/hockey-fight-vidoes/data, Last Visited, 2024.
[27] Snoun A., Jlidi N., Bouchrika T., Jemai O., and Zaied M., "Towards a Deep Human Activity Recognition Approach Based on Video to Image Transformation with Skeleton Data," *Multimedia Tools and Applications*, vol. 80, no. 19, pp. 29675-29698, 2021. https://link.springer.com/article/10.1007/s11042-021-11188-1
[28] Sumon S., Shahria M., Goni M., Hasan N., Almarufuzzaman A., and Rahman R., "Violent Crowd Flow Detection Using Deep Learning," *in Proceedings of the 11th Asian Conference, Intelligent Information and Database Systems*, Yogyakarta, pp. 613-625, 2021. https://doi.org/10.1007/978-3-030-14799-0_53
[29] Sun F., Zhang J., Wu X., Zheng Z., and Yang X., "Video Anomaly Detection Based on Global-Local Convolutional Autoencoder," *Electronics*, vol. 13, no. 22, pp. 1-18, 2024. https://doi.org/10.3390/electronics13224415
[30] Sung C. and Park J., "Design of an Intelligent Video Surveillance System for Crime Prevention: Applying Deep Learning Technology," *Periodicals Multimedia Tools and Applications*, vol. 80, no. 26-27, pp. 34297-34309, 2021. https://doi.org/10.1007/s11042-021-10809-z
[31] Vijayakumar E., Puviarasan A., Natarajan P., and Ganesan S., "Optical Flow-based Feature Selection with Mosaicking and FrIFrO Inception V3 Algorithm for Video Violence Detection," *Engineering, Technology and Applied Science Research*, vol. 14, no. 3, pp. 14475-14482, 2024. https://doi.org/10.48084/etasr.7270
[32] Wankhade A., Jaiswal S., and Tingane S., "Violence Detection in Surveillance Videos Using Artificial Intelligence," *International Journal of Engineering Research and Management*, vol. 11, no. 5, pp. 32-39, 2024. https://www.ijerm.com/download_data/IJERM1105009.pdf
[33] Wasim M., Ahmed I., Ahmad J., and Hassan M., "A Novel Deep Learning Based Automated Academic Activities Recognition in Cyber-Physical Systems," *IEEE Access*, vol. 9, pp.

63718-63728, 2021. DOI: 10.1109/ACCESS.2021.3073890

[34] Wu C. and Cheng Z., "A Novel Detection Framework for Detecting Abnormal Human Behavior," *Mathematical Problems in Engineering*, vol. 2020, no. 1, pp. 1-9, 2020. https://doi.org/10.1155/2020/6625695

[35] Xu F., Luo Y., Sun C., and Zhao H., "Improved Convolutional Neural Network for Traffic Scene Segmentation," *Computer Modeling in Engineering and Sciences*, vol. 138, no. 3, pp. 2691-2708, 2024. https://doi.org/10.32604/cmes.2023.030940

[36] Zhang L., Ruan X., and Wang J., "WiVi: A Ubiquitous Violence Detection System with Commercial WiFi Devices," *IEEE Access*, vol. 8, pp. 6662-6672, 2020. DOI: 10.1109/ACCESS.2019.2962813

[37] Zhang P., Zhao X., Dong L., Lei W., Zhang W., and Lin Z., "A Framework for Detecting Fighting Behavior Based on Key Points of Human Skeletal Posture," *Computer Vision and Image Understanding*, vol. 248, pp. 104123, 2024. https://doi.org/10.1016/j.cviu.2024.104123

David Calderon-Vilca abstained PhD in Computer Science, research Professor of the "Artificial Intelligence" Group of the Universidad Nacional Mayor de San-Peru, professor in the Universidad Nacional de Ingenieria. Professor of doctoral programs in other universities. Development projects related to Neural Networks, Machine Learning and Natural Language Processing, Advisor of undergraduate and graduate thesis.

Kent Cuadros-Ramos is a Software Engineer specializing in the Financial Industry. He is currently a DevOps Architect at Hapi, where he enhances Technological Infrastructure, Implements Managed Secrets, and Optimizes Performance. Previously, he was a Tech Leader at Interbank, Leading Application Management and Acting as a Liaison between Technology and Clients. He also served as a Software Architect at Global66, developing and refactoring APIs. He holds a degree in Software Engineering from the National University of San Marcos (UNMSM) and is pursuing a Diploma in Agile Methods and Innovation at the Pontifical Catholic University of Peru (PUCP). He is proficient in Technologies such as Java, Python, Spring Boot, Docker, Kubernetes, and AWS.

Sergio Valcarcel-Ascencios is currently a PhD candidate in Systems Engineering at the National University of San Marcos in 2023. He holds a Master's degree in Systems Engineering and Computer Science. Assistant Professor at the National University of San Marcos (2019), in undergraduate and graduate courses. He is a member of the AI research group, has published article and books in this line of research and is a reviewer of scientific articles in AI at the Technological Institute of Production. He is experienced in the implementation of information technology projects applying AI concepts at the National Service of Agricultural Health since 1995. He is an active member of the normative technical committee 086 of research, Technological Development and Innovation Management and Mirror Committee International ISO/TC 279. His research areas are Artificial Neural Network, Internet of Things and Knowledge Management.

Igor Aguilar-Alonso is a Senior Professor at the Professional School of Systems Engineering at the Universidad Nacional Tecnologica de Lima Sur and Professor in the Software Engineering department at Universidad Nacional Mayor de San Marcos, Peru. He received the M.Sc. degree in Industry 4.0 from Universidad Internacional de la Rioja (UNIR), Logroño, Spain, and the Ph.D. degree from Universidad Politécnica de Madrid (UPM), Spain, in 2013. He was an employee at IBM Global Service, Spain. His research papers have been published at conferences and international journals. He is currently serving as a Referee of research for national and international scientific journals, and conference proceedings, he is also a reviewing consultant of the National Council for Science and Technology, Peru. His research areas are Industry 4.0 Technologies, IT Governance, IT Services Management, and Business.