

Ginés Guerrero · Jaime San Martín ·
Esteban Meneses ·
Carlos Jaime Barrios Hernández ·
Carla Osthoff ·
Jose M. Monsalve Diaz (Eds.)

Communications in Computer and Information Science

2270

High Performance Computing

11th Latin American High Performance Computing Conference CARLA 2024
Santiago de Chile, Chile, September 30 – October 4, 2024
Revised Selected Papers



Springer

Communications in Computer and Information Science

2270

Series Editors

Gang Li , *School of Information Technology, Deakin University, Burwood, VIC, Australia*

Joaquim Filipe , *Polytechnic Institute of Setúbal, Setúbal, Portugal*

Zhiwei Xu, *Chinese Academy of Sciences, Beijing, China*



Web System for Recognizing Actions of Physical Violence in Urban Spaces Using CNN with Transfer Learning

José Edgar García Díaz¹ , Wilder Suárez Romero¹ , Ciro Rodríguez^{1,3},
Jorge Puga de la Cruz² , Victor Manuel Cabrejos Yalan¹ ,
Isis Ayme Moran Temoche¹ , and Jhon Charlie Martínez Carranza⁴ 

- ¹ Escuela Universitaria de Posgrado, Universidad Nacional Federico Villarreal (UNFV),
Lima 15001, Perú
- ² Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional de la Amazonia
Peruana (UNAP), Iquitos 16001, Perú
- ³ Department of Software Engineering, Universidad Nacional Mayor de San Marcos
(UNMSM), Lima 15081, Perú
- ⁴ Instituto de Investigaciones de la Amazonía Peruana (IIAP), Iquitos 16001, Perú

Abstract. Violence is a problem of great impact that affects the countries of the world, violence is a soft system that to date cannot be controlled, because it can manifest itself in different ways, represented by high levels of crime and delinquency. But now, there is Artificial Intelligence (AI), which uses innovative resources and will help close gaps in the Sustainable Development Goals (SDG), such as in health, traffic management, climate change and also violence. One way to apply AI is through image processing with Convolutional Neural Networks (CNN). This research evaluates the effectiveness of classifying actions of interpersonal violence such as: strangulation, grappling, kicking or punching, using CNN with Transfer Learning. First, a personalized dataset was developed using images with simulated violent actions, obtaining a total of 2000 images, which were distributed into 60% for training, 30% for validation and 10% for testing. Second, the pre-trained CNN models were trained: VGG16, MobileNetV2 and YOLOv8, subjected to 150 epochs, where the performance results were compared between them and it was determined that the best performance was for YOLOv8, obtaining an accuracy rate of 94.25% and an accuracy of 89%. In the end, a prototype web system was developed using Python with OpenCV and YOLOv8 to detect actions of interpersonal violence in real time and alert the authorities involved in a timely manner. This research will serve as a reference for future real-time video surveillance applications.

Keywords: Physical violence · Convolutional Neural Networks · Web System · Deep Learning

1 Introduction

On their website, the Pan American Health Organization (PAHO) together with the World Health Organization (WHO) announce that every year at least 470,000 people in the world die from homicide, a very alarming figure; even worse for the Americas which has one of the highest homicide rates, which is three times the global average, where almost 500 people have died per day as a result of interpersonal physical violence, which at least 1 in 3 women suffered physical or sexual violence by your partner [1]. Furthermore, the rates of violence in Latin America and the Caribbean generate a climate of uncertainty and insecurity in the population, since it is also a problem linked to the political instability of a government, the poor economy of a country makes it difficult to counteract it, as well as other factors. Associated [2–4].

In Peru, Law No. 30364 [5] establishes measures against violence against women and family members, which classifies violence into four types: physical, psychological, sexual and economic or patrimonial. The National Institute of Statistics and Informatics - INEI of Peru, in its technical report on crime and citizen security, announces that the most violent cases are: psychological and physical violence, and that in 2017, 65.4% of women between 15 and 49 years old, were raped at some point by their partner [6]. Furthermore, the INEI indicates that the majority of women were victims of psychological violence (61.5%), physical violence (30.6%) followed by sexual violence (6.5%).

Likewise, although governments have support and monitoring programs [7] together with the police to counteract this problem, through citizen control, security in strategic areas, integrated patrols as well as having video surveillance centers, these strategies are not enough due to the great need and/or dependence on the presence of the human factor to be able to identify and warn about violent actions, however, today AI makes it possible to streamline recognition processes of said physical violent actions [8–15] and automate your early warning by communicating in a timely manner to the authorities involved.

2 Related Works

The works related to this research were found in indexed databases such as: Scopus, IEEE, EBSCOhost, Wiley and ACM applying a Systematic Literature Review – RSL with key criteria of “deep learning” and “physical violence”, in addition, only works that used CNN with Transfer Learning were considered; finally, data collection was carried out taking into account the following fields: Authors, Pre-trained models used by the researchers, the pre-training Dataset, the proposed Hybrid Model and the result of the “Accuracy” metric in a way that allows carrying out a performance comparison between them, as detailed in Table 1.

In summary, Table 1 shows some CNN models with Transfer Learning used in similar research, where there are studies using only pre-trained CNN models [8–10, 13, 16] as well as hybrid models that can be observed, that is That is, they are investigations that study the combination of CNN models with Machine Learning algorithms [8, 11, 12, 14, 15] in order to find better performance for the proposal. In this research, tests were carried out using three (3) pre-trained CNN models, which are: VGG16, MobileNetV2

Table 1. References that use CNN models with Transfer Learning

Author	Models used	Dataset Pretraining	Proposed Model	Acc (%)
[8]	MobileNet	ImageNet	Conv LSTM	91.00
	YOLOv7	COCO	-	mAP@0.5 = 76%
[9]	YOLOv3	COCO	-	94.00
[10]	MobileNet	ImageNet	-	89.33
[11]	InceptionV1	ImageNet+Kinetics	2-Stream 3D-CNN	98.00
[12]	ResNet50	ImageNet	CNN+LSTM	93.63
[13]	MobileNet	ImageNet	-	95.41
[14]	VGG16	INRIA Person	2D BiGRU-CNN	94.58
[15]	VGG16	ImageNet	VGG16+LSTM	94.70
[16]	YOLOv2	COCO	-	88.00

and YOLOv 8, the latter needed to work with labeling techniques to detect objects in the image, but the first two work with image classification models. From its total dimension. Likewise, in this research a personalized dataset was used with images of actions of simulated interpersonal physical violence, and which were subjected to the three aforementioned models, in order to consider the best of them to be used within the development of the web system.

3 Proposed Design

3.1 Dataset Development

The first stage was the preparation of the dataset, it was decided to develop one based on simulated physical violence actions in scenarios in urban areas, for which various video recordings were made with body movements or behaviors of: strangulation, grappling, kicking and punching [9] which were developed with the participation of young university students, who provided support to this research voluntarily using their cell phones to make the various video recordings.

To make the recordings, high-end cell phones with twelve (12) megapixel resolution cameras were used, which were directed towards a single angle, because what is needed in this research is to try as much as possible to simulate a camera of video surveillance in an urban area, and normally these are always placed facing the same capture angle, for which the following was considered:

- The cell phone was placed on a stand or tripod with a downward inclination of 15°.
- The cell phone should have a vertical distance of 2 m high and a horizontal distance of 4 m maximum.
- The recordings were of simulated physical violent actions between a maximum of 2 participants (see Fig. 2) and without other people around.

- Each video made had an average duration of 10 min.
- The recordings were made in the morning with natural light and at the same time without much sunlight.
- The cell phone camera was not exposed to rainy conditions or at night since the images had to be as clear as possible and free of noise or distortions.

In the end, and with the help of the “VLC media player” software, a quantity of two thousand (2000) frames or images of size 1920×1088 pixels were selected, which represents a standard quantity of images with which it was possible to work without have overfitting problems, and that were distributed as detailed in Table 2 and placed in labeled folders, as shown in Fig. 1:

Table 2. Distribution of selected images

Classes	Training 60%	Validation 30%	Testing 10%	Total
Strangulation	300	150	50	500
Grappling	300	150	50	500
Kicking	300	150	50	500
Punching	300	150	50	500
TOTAL	1200	600	200	2000

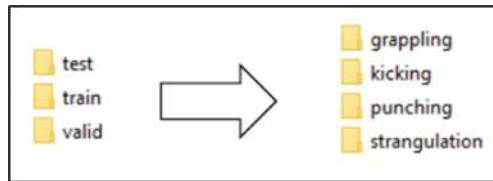


Fig. 1. Distribution of the dataset folders

3.2 Training of Selected Models

The selected models were trained with the dataset created using a Workstation with the following characteristics: Intel Core i9-9900K Processor, 128 GB RAM (4×32 GB) 3200 MHz, NVIDIA GPU model TITAN RTX 24 GB GDDR6 Professional, disk 1 TB M.2 Sata solid state hard drive (SSD) and one TB 7200 RPM hard drive. Which were executed using Anaconda and Jupyter Notebook software. In addition, before executing the training, Transfer Learning techniques were applied to take advantage of the pre-trained weights of the models.

Before executing the training for the VGG16 and MobileNetV2 models, Transfer Learning was first applied, adapting its last layers according to the parameters detailed in Table 3 so that they can only work with four (4) classes of physical violence, also,



Fig. 2. Example of images selected for the dataset: Strangulation (a), Grappling (b), Kicking (c) and Punching (d)

the hyperparameters of both models were configured as detailed in Table 4, however, it is necessary to indicate that the proposal of the values assigned to the hyperparameters result from the evaluation of the state of the art, but these values can be configured and adapted to other proposals or different combinations during the training tests in order to seek better performance of the results.

Table 3. Sequence of layers to apply Transfer Learning

Layer	Valor
Flatten	-
Dense	neurons: 128; activation: “relu”
Dense	neurons: 128; activation: “relu”
Dense	neurons: 4; activation: “softmax”

On the other hand, in the case of YOLOv8, the pre-trained weight of “yolov8x” available on the <https://github.com/ultralytics/ultralytics> page was used directly and the following training code was applied.

```
!yolo task = detect mode = train model = yolov8n.pt data  
= {dataset.location}/data.yaml epochs = 150 batch = 16 imgsz = 224.
```

Table 4. Hyperparameters of pre-trained CNN models

Hiperparámetro	Valor
Época	150
Bach size	16
Loss	“categorical_crossentropy”
Opmizer	“adadelata”
Metrics	“accuracy”

4 Results

4.1 Training Results of the Selected Models

In the second stage of the research, the training, validation and testing processes of the selected CNN models were carried out, for which Python code was applied in addition to using Keras libraries, to obtain the VGG16 and MobileNetV2 models, where their final layers were subjected to Transfer Learning technique and used the hyperparameters mentioned above. Likewise, at the end of each training, the views of the “Loss” and “Accuracy” graphs were developed and with the help of the Matplotlib and Sklearn libraries to display the Confusion Matrix. Below are the graphs obtained from the training and validation processes with respect to the loss and accuracy metrics, in addition to showing the Confusion Matrix of each selected model.

VGG16 was the first model to run, which lasted approximately 1 h and 30 min of training; Fig. 3a shows how the “Loss” level decreases as the epochs progress until reaching levels less than 2% and Fig. 3b shows how the “Accuracy” level increases, exceeding 60% at as the times advance.

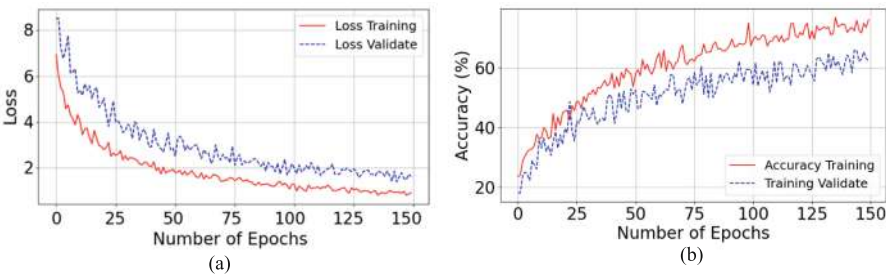


Fig. 3. Training and validation of loss (a) and accuracy (b) with VGG16

Figure 4 shows the Confusion Matrix of VGG16, as a result of the tests carried out with 50 violent images of each class, where it can be seen that 41 images were correctly classified for the “Punching” class, while the least accurate class was from “Strangulation” with only 6 images correctly classified. Likewise, it is observed in the matrix that the most common errors with this model fall on the actions of grappling and kick with

33 and 27 errors respectively, where their predictions were taken as strangulation and punching respectively, this may have happened because grappling and Strangulation are very similar body movements, the same thing happens between a kick and a punch when the algorithm confuses an arm with a leg or vice versa.

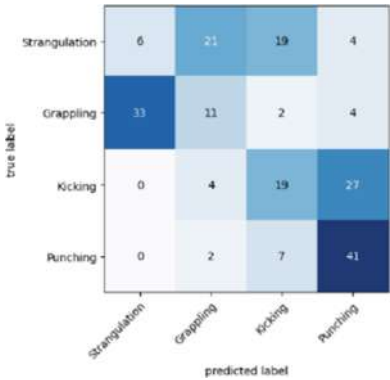


Fig. 4. VGG16 model confusion matrix

MobileNet was the second model to run, which lasted approximately 2 h of training; Likewise, Fig. 5a shows how the “Loss” level decreases to levels less than 1% and Fig. 5b shows how the “Accuracy” level increases and exceeds 80% as the epochs progress. of training.

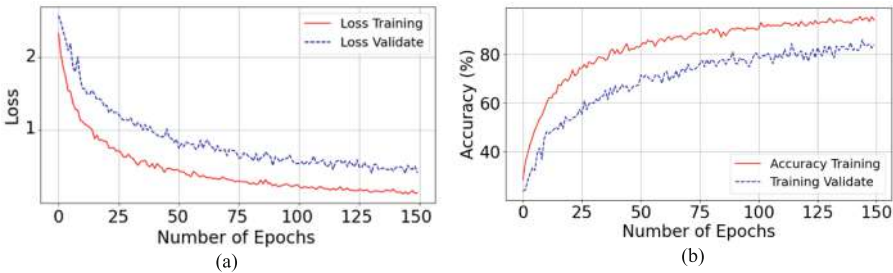


Fig. 5. Training and validation of loss (a) and accuracy (b) with MobileNet

Figure 6 shows the results of the MobileNet Confusion Matrix, where tests were also done with 50 new violent images from the study classes, but it is observed that the results are better than VGG16 since 49 correctly classified images were obtained from the “Punching” class and 48 correctly classified images of the “Grappling” classes, however, the “Strangulation” class is still the least accurate since it continues with only 6 correctly classified images. However, it is observed that there is a high level of error regarding the strangulation actions (41 errors), where their predictions were taken as a grappling, and this happens due to the great similarity between these two movements,

which is when it happens that the algorithm does not differentiates these actions very well.

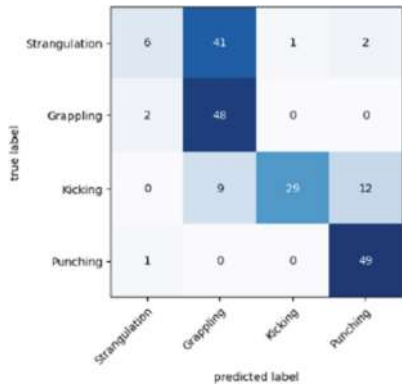


Fig. 6. MobileNet confusion matrix

YOLOv8 was the third model to run, which lasted approximately 2 h and 50 min of training; Fig. 7 shows how the “box_loss” level for training and validation decreases as the epochs progress until reaching levels less than 0.5% while the levels “Precision”, “mAP50” and “mAP50-95” They increase, exceeding 90% as the times progress.

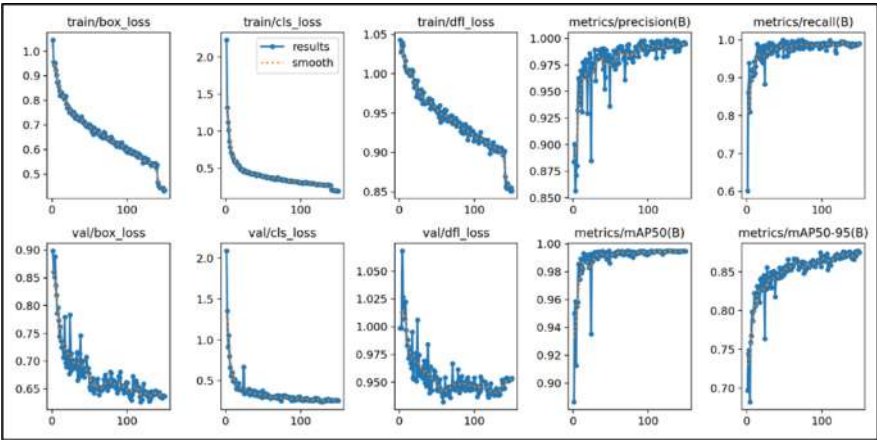


Fig. 7. Training results with YOLOv8

Figure 8 shows the results of the YOLOv8 Confusion Matrix where tests were also done with 50 new violent images of the study classes, but it is observed that the results are much better than VGG16 and MobileNetV2 since 49 classified images were obtained. Correctly of the “Punching” class and 50 correctly classified images of the “Grappling” class and 49 correctly classified images of the “Kicking” class. Likewise,

with YOLOv8 the classification of the “Strangulation” class improved with 41 correctly classified images.

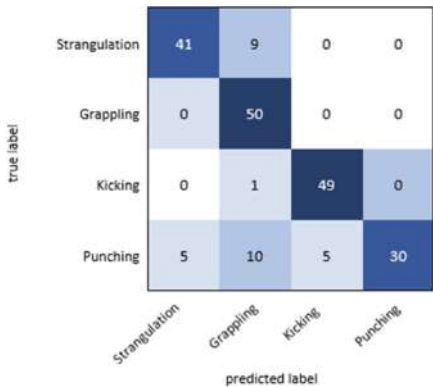


Fig. 8. YOLOv8 confusion matrix

In the end, the performance of each selected model was evaluated according to each class of study, taking into account the metrics of Precision, Sensitivity, Specificity and F1-score [17]. Likewise, the “Accuracy” metric was considered the highest. Importance, as detailed in Table 5.

After analyzing the results of Table 5, YOLOv8 was considered the ideal model to implement with the Web System since it achieved an average precision of 94.25% and an accuracy level of 89.00% and $mAP@0.5 = 91\%$, which indicates greater reliability to classify the different actions of physical violence, then, next, the system for real-time recognition was developed.

4.2 The Web System and YOLOv8

In the third stage of the research, the web system and its integration with the YOLOv8 model were developed, based on the support code from the GitHub section: <https://github.com/xavialex/Streamlit-TF-Real-Time-Object-Detection> which allowed creating an interactive website with Python language and Tensorflow, Keras and OpenCV libraries quickly and with a minimum of code, which also allowed integration with the selected model; Likewise, Fig. 9 shows a possible prototype or scheme to implement the web system as if it were a real-time video surveillance system.

Finally, it is considered that the operation of the web system is very simple, you simply have to place the web camera in the appropriate place at a maximum height of 2 m and having the capture area fixed at a distance of no more than 4 m, then in the main interface of the web system select the “Detection” option, then select the level of precision you want in the “Select Model Confidence” scroll bar, this level ranges from 25% to 100%, and then the “Video” option to analyze saved violent videos or the “Webcam” option to activate the web camera to capture violence in real time, at the end click on the “Detect Video Object” button and the system starts working, as shown in

Table 5. Summary of performance metrics for selected models

Models	Classes	Performance metrics in %				
		Prec.	Sens.	Spec.	F1	Acc.
VGG16	Strang.	15.38	12.00	68.26	13.48	38.5
	Grappling	28.95	22.00	70.97	25.00	
	Kicking	40.43	38.00	67.44	39.18	
	Punching	53.95	82.00	95.88	65.08	
	Avg	34.68	38.50	75.63	35.69	
MobileNetV2	Strang.	66.67	12.00	97.67	20.34	66.0
	Grappling	48.98	96.00	62.69	64.86	
	Kicking	96.67	58.00	99.04	72.50	
	Punching	77.78	98.00	85.57	86.73	
	Avg	72.53	66.00	86.31	61.11	
YOLOv8	Strang.	93.17	78.11	97.94	84.98	89.0
	Grappling	92.18	100.0	96.92	95.93	
	Kicking	95.38	99.35	97.98	97.32	
	Punching	96.28	76.38	99.06	85.18	
	Avg	94.25	88.46	94.75	90.85	

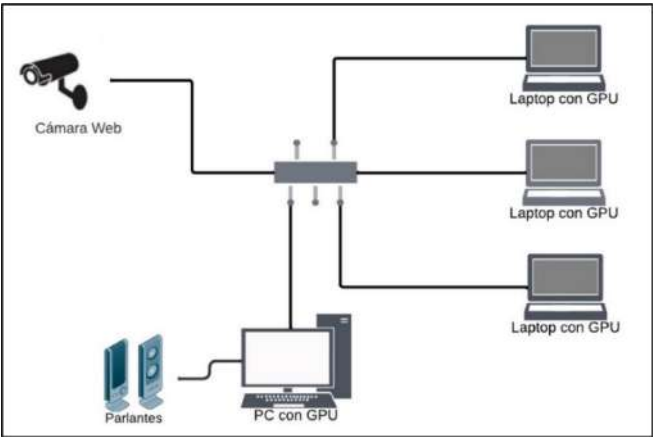


Fig. 9. Implementation scheme of the Web System based on YOLOv8

Fig. 10. The system will begin to recognize the different actions of interpersonal physical violence by showing labels around the detected images, as can be seen in Fig. 11 and when detecting at least four (4) violent actions uninterrupted within a period of 10 s, the

system will emit an audible alert through the speakers connected to the PC or Laptop to notify the people or authorities involved and act in accordance with the law.

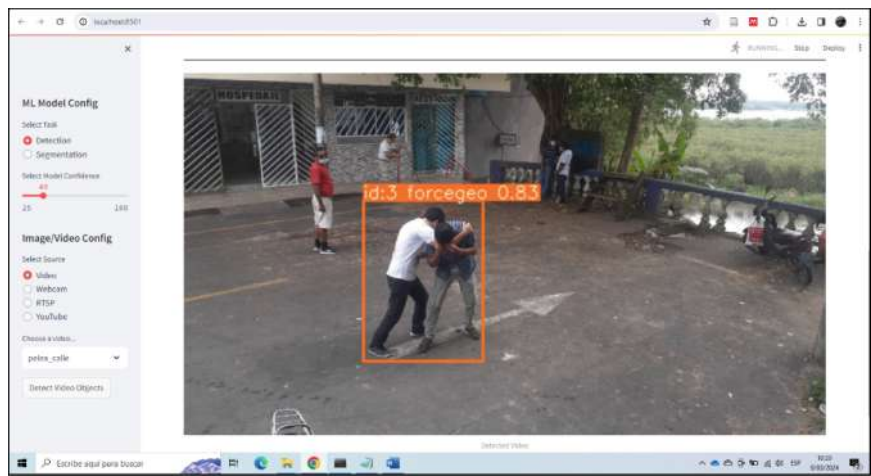


Fig. 10. Home page of the Web System based on YOLOv8

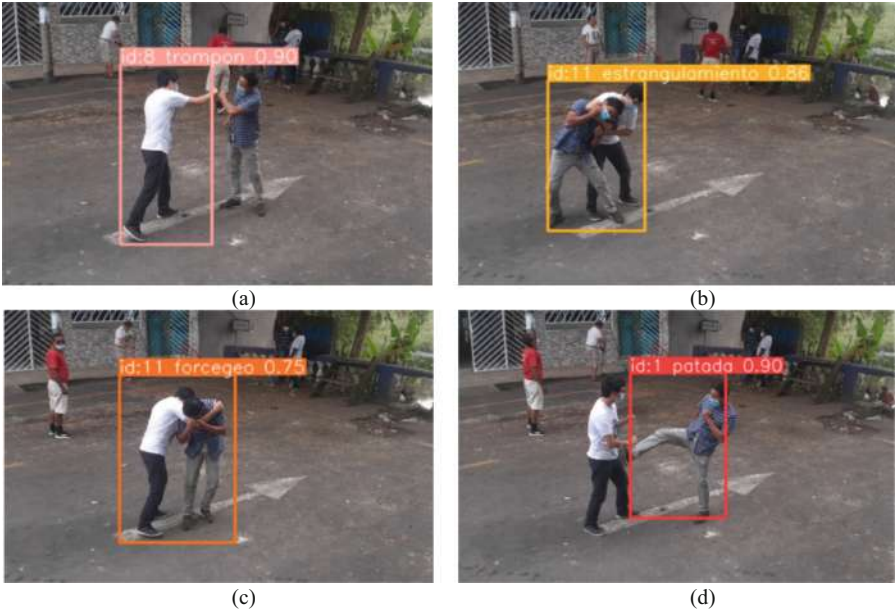


Fig. 11. Example of recognitions detected by the YOLOv8-based Web System: Punching 90% (a), Strangulation 86% (b), Grappling 75% (c) and Kicking 90% (d)

5 Conclusion

An investigation was carried out to compare three pre-trained CNN models based on Transfer Learning techniques to classify actions of physical violence in urban areas, the results show that the selected and tested models work adequately with a personalized dataset and provide significant advantages that helped deduce which of these models is the most appropriate for the proposed research.

Likewise, from the results of the performance metrics, it is concluded that the model that presents the best performance is YOLOv8 compared to its similar VGG16, MobileNetV2, however, we must take into account the handling that these models have to recognize and detect objects in the images since it is a determining factor and that is what we have verified in this investigation.

In addition, the Web System was implemented with YOLOv8, which is a technology that has been constantly updated and is considered one of the most optimal models to perform in scenarios to detect objects in real time, where its convolutional layers have better filters to extract features. Essentials of an image.

6 Recommendation

It is recommended that, for future work, a more robust personalized dataset should be generated with other actions of violence, and tests should also be carried out with other object detection models that use the bounding box labeling technique, such as Thunder, PicoDet, among others. Finally, it is recommended to implement an Intelligent Video Surveillance System in real time that issues other types of alarms automatically, such as sending alert signals or text messages to the cell phones of the competent authorities to ensure the tranquility of the areas. Urban.

References

1. Organización Panamericana de la Salud [OPS]. Prevención de la violencia - OPS/OMS (2021). <https://www.paho.org/es/temas/prevencion-violencia>
2. Latinoamérica 21. La violencia interminable en América Latina (2021). <https://latinoamerica21.com/es/la-violencia-interminable-en-america-latina/>
3. Deutsche Welle Homepage. <https://www.dw.com/es/crisis-pol%C3%ADtica-en-per%C3%BA-tensi%C3%B3n-e-incertidumbre/a-55627085>
4. Ediciones El País Homepage. <https://elpais.com/economia/2020-12-04/la-crisis-politica-en-peru-pone-a-prueba-la-inmunidad-de-la-economia.html>
5. El Peruano: Ley para prevenir, sancionar y erradicar la violencia contra las mujeres y los integrantes del grupo familiar - LEY - N° 30364. CONGRESO DE LA REPUBLICA (2015). <https://busquedas.elperuano.pe/dispositivo/NL/1314999-1>
6. Instituto Nacional de Estadística e Informática: Boletín de Seguridad Nacional (2020)
7. Gob.pe Homepage. <https://www.gob.pe/institucion/mimp/noticias/84829-loreto-respuesta-de-la-policia-ante-casos-de-violencia-hacia-las-mujeres-es-reforzada-con-aplicacion-de-nueva-ficha-de-valoracion-de-riesgo>
8. Sakiba, C., Tarannum, S.M., Nur, F., Arpan, F.F., Anzum, A.A.: Real-time crime detection using convolutional LSTM and YOLOv7. Doctoral dissertation, Brac University (2023)

9. Yahuarcani, I.O., et al.: Recognition of violent actions on streets in urban spaces using Machine Learning in the context of the Covid-19 pandemic. In: Proceedings of the 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), Cape Town, South Africa, pp. 1–4 (2021). <https://doi.org/10.1109/ICECET52533.2021.9698762>
10. Gaytán Aguilar, I., et al.: A comparative study of three pre-trained convolutional neural networks in the detection of violence against women. CIENCIA ergo-sum, [S.l.], v. 31, n. 2, abr. (2023). ISSN 2395-8782. <https://cienciaergosum.uaemex.mx/article/view/19352>
11. Mehmood, A.: Abnormal behavior detection in uncrowded videos with two-stream 3d convolutional neural networks. Appl. Sci. **11**(8), 3523 (2021)
12. Patel, M.: Real-time violence detection using CNN-LSTM. arXiv preprint [arXiv:2107.07578](https://arxiv.org/abs/2107.07578) (2021)
13. Haque, M.R., et al.: Crime detection and criminal recognition to intervene in interpersonal violence using deep convolutional neural network with transfer learning. Int. J. Ambient Comput. Intell. (IJACI) **12**(4), 154–167 (2021)
14. Traoré, A., Akhloufi, M.A.: 2D bidirectional gated recurrent unit convolutional neural networks for end-to-end violence detection in videos. In: Campilho, A., Karray, F., Wang, Z. (eds.) Image Analysis and Recognition. ICIAR 2020. Lecture Notes in Computer Science, vol. 12131, pp. 152–160. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50347-5_14
15. Soliman, M.M., Kamal, M.H., Nashed, M.A.E.M., Mostafa, Y.M., Chawky, B.S., Khattab, D.: Violence recognition from videos using deep learning techniques. In: Proceedings of the 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 80–85. IEEE, December 2019
16. Lumba, A.P.L., et al.: Computing solution for the recognition of basic actions of violence in real time, from the use of convolutional neural networks, video sequences and high performance computing. XLV Lat. Am. Comput. Conf. (CLEI) **2019**, 1–9 (2019). <https://doi.org/10.1109/CLEI47609.2019.235100>
17. Lopez-Betancur, D., Bosco Durán, R., Guerrero-Mendez, C., Zambrano Rodríguez, R., Saucedo Anaya, T.: Comparación de arquitecturas de redes neuronales convolucionales para el diagnóstico de COVID-19. Computación y Sistemas **25**(3), 601–615 (2021)