





## Classification of Physical Violence Actions Using Convolutional Neural Networks with Transfer Learning

José Edgar García Díaz<sup>1,2\*</sup>, Ciro Rodríguez<sup>1,3</sup>

<sup>1</sup> Escuela Universitaria de Posgrado, Universidad Nacional Federico Villarreal (UNFV), Lima 15001, Peru

<sup>2</sup> Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional de la Amazonia Peruana (UNAP), Iquitos 16001, Peru

<sup>3</sup> Department of Software Engineering, Universidad Nacional Mayor de San Marcos (UNMSM), Lima 15081, Peru

Corresponding Author Email: [2021007985@unfv.edu.pe](mailto:2021007985@unfv.edu.pe)

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ijse.140501>

### ABSTRACT

**Received:** 29 June 2024

**Revised:** 20 August 2024

**Accepted:** 4 September 2024

**Available online:** 31 October 2024

#### Keywords:

*classification, physical violence, deep learning, transfer learning*

Violence in the world is a problem of great impact that affects any developing country; it represents a soft system that to date cannot be controlled due to its different manifestations with high rates of crime and delinquency. Artificial Intelligence (AI) uses different innovative resources that help close gaps related to the Sustainable Development Goals (SDGs) such as health, traffic management, climate change, among others, being a way of applying AI to through image processing with Convolutional Neural Networks (CNN). This research evaluates the effectiveness of classifying violent actions such as: strangulation, grappling, kicking or punching, using CNN with Transfer Learning. First, a personalized dataset was created with simulated images of violent actions, made up of 2000 images distributed in 60% for training, 30% for validation and 10% for testing. Second, the pre-trained CNN models were trained: VGG16, MobileNetV2, ResNet50 and InceptionV3 applying Transfer Learning, subjected to 150 epochs and using the same hyperparameters. In the end, the performance results were compared between them, where it was determined that the best performance is from MobileNet, which obtained a precision rate of 72.53%, and an accuracy level of 66%. This research will serve as a reference for future applications.

## 1. INTRODUCTION

The large population growth concentrated in the cities of the world is related to urban social disorder, according to data from Urban Social Disorder v.3 that studies 186 national capitals and urban centers from 1960 to 2014, demonstrating a relationship between population size and frequency of events related to violence [1] it can be said that these events are the product of lack of employment, income inequality or the difficult economic situation that are factors that contribute significantly to violence, such as in the regions of Latin America where the highest rates of violence are experienced [2].

Likewise, the Pan American Health Organization (PAHO) and the World Health Organization (WHO), in their article "Prevention of Violence," define violence as the "intentional use of actual or threatened physical force or power against self, a person, group or community that results in the likelihood of psychological harm, injury, death, deprivation or maldevelopment" [3]. PAHO/WHO announced that every year at least 470,000 people in the world die from homicide, a very alarming number; 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 violence; It also indicates that at least 1 in 3 women suffered physical or sexual violence from their

partner [3].

Furthermore, the rates of violence in Latin America and the Caribbean generate a climate of uncertainty and insecurity in the population, which is a problem related to the instability of the government and the poor economy that makes a good analysis to counteract it difficult, among other associated factors [2-4]. In Peru, Law No. 30364 [5] is about taking measures against violence against women and family members, where violence is classified into four types: physical, psychological, sexual, and economic or patrimonial. Meanwhile, 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. Likewise, the INEI indicates that the majority of women were victims of psychological violence (61.5%), as well as physical violence (30.6%), followed by sexual violence (6.5%) [6].

Although governments always produce support programs for the population through the police, citizen control centers, security of strategic urban areas, integrated patrols as well as municipal video surveillance centers throughout the day, these strategies are limited due to the need and/or dependence on the human factor to analyze, identify and automatically alert violent actions [7], making it necessary to apply AI

technological strategies that allow streamlining the processes of recognizing violence without dependence on the human factor [8-10].

It is important to consider the challenges and transparency needs of the research considering the current limitations, such as the capacity of the proposed system that adopt the CNN model, transfer learning, and data visualization that are easier to interpret for image processing, making its results understandable that facilitate the identification of acts of violence and the protection of people from these acts, allowing users to understand and trust the process [11-13]. Likewise, the system must adapt to respond resiliently to changes in its environment due to unforeseen challenges that require accommodating new circumstances of violence with new technical skills in response to environmental changes [14, 15].

Therefore, the technologies of video surveillance systems must be combined with AI to take advantage of the advantages they provide [16] and establish a mechanism that guarantees better citizen security through automated monitoring [17]. Today, AI through deep learning based on computational models such as Convolutional Neural Networks (CNN) can perform image processing and computer vision using different architectures such as Visual Geometry Group 16 (VGG16), Mobile Network (MobileNet), Residual Network 50 (ResNet50) [18] and Google's InceptionV3.

This research aims to solve the problem of identifying a CNN architecture capable of detecting violent actions quickly and efficiently, for which performance tests were performed on four models to evaluate their ability to classify violent actions such as strangulation, grappling, kicking or punching, combined with the application of the Transfer Learning technique with each of them.

One of the gaps identified in the literature reviewed is that research is more oriented to health-related effects [18-20]; however, prevention is very important in anticipating and warning of an act of violence, which is what is sought with this research.

After the evaluation, the model with the highest accuracy and best average of its performance metrics was selected [19]. This information obtained is intended to establish a contrast

between the different CNN models so that they can be considered in future research related to detecting interpersonal physical violence.

## 2. CNN ARCHITECTURES

### 2.1 VGG16

The Visual Geometry Group (VGG) proposed by Simonyan and Zisserman [21] from the University of Oxford, reduced the margin of error by 7.3%. The contribution of the work demonstrated that the depth of a network is a critical component for improving recognition or classification accuracy in CNN. The VGG architecture consists of two convolutional layers, which use the ReLU activation function; after activation, there is a single max pooling layer and several fully connected layers also using ReLU, as shown in Figure 1. The last layer of VGG is a Softmax layer for classification, the convolution filter size is changed to a  $3 \times 3$  filter with a step of 2. The VGG models VGG11, VGG16, and VGG19 have 11, 16, and 19 layers, respectively, hence their names [22].

### 2.2 MobileNetV2

MobileNetV2 was proposed in 2017, with a lightweight architecture designed for mobile and end-to-end vision applications. Google researchers developed it as an enhancement to the original MobileNet model; a notable aspect of this model is its ability to strike a balance between size and accuracy, making it ideal for resource-constrained devices. The MobileNetV2 architecture (see Figure 2) incorporates several key features that contribute to its efficiency and effectiveness in image classification tasks, these features include depth-separable convolution, inverted residuals, bottleneck design, linear bottlenecks, and compression and excitation blocks; each feature plays a crucial role in reducing the computational complexity of the model while maintaining high accuracy [23].

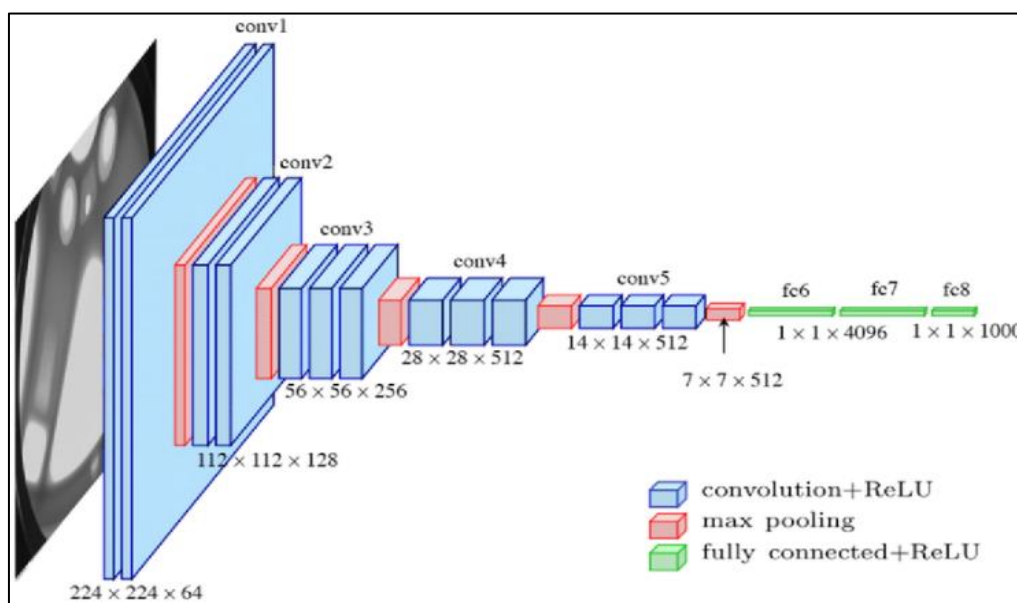


Figure 1. VGG16 model architecture

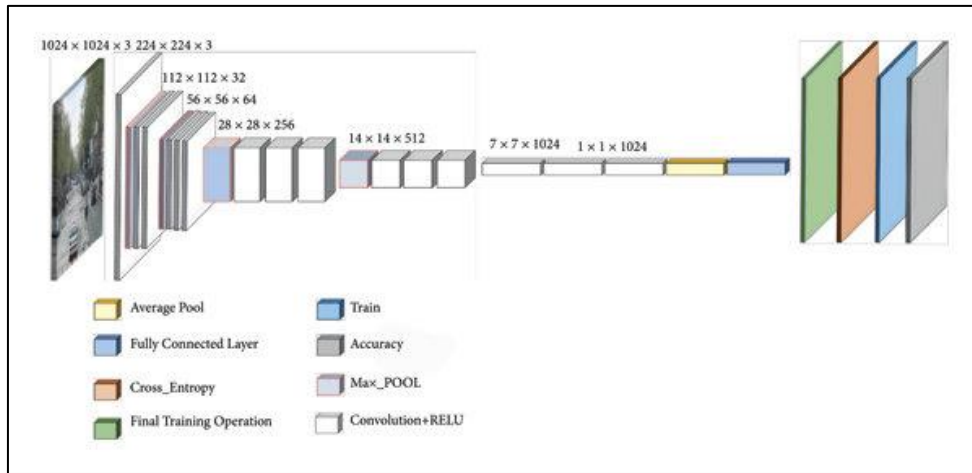


Figure 2. MobileNetV2 model architecture

### 2.3 ResNet50

ResNet50 is a Residual Neural Network of 50 layer with a simple, effective and beneficial structure of wide application containing a residual structure block, within the identity mapping it is necessary to match boundaries and use a  $1 \times 1$  convolutional layer to add dimensions. ResNet50 [24], solves the problem of gradient fading, being too deep, gradients can pass directly through the return hopping connections of the layers downstream of the initial filters, demonstrating great performance in image classification and object detection tasks. ResNet50 has influenced the design of other architectures and consists of over 23 million parameters, being much smaller than VGG16 and consists of a series of residual blocks that learn features at different levels of abstraction. In the end, it uses clustering layers, fully connected layers and an output layer to perform specific image classification tasks (see Figure 3).

achieved by freezing and unfreezing its layers, converting them into a retrainable network that can be used in any specific problem and adapted to our needs (see Figure 4). This model was pre-trained with the ImageNet dataset and is adjusted with a Batch size of 8 and a Learning rate of 0.0001 [25].

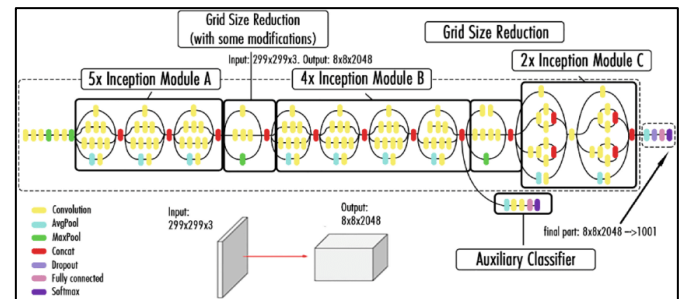


Figure 4. InceptionV3 model architecture

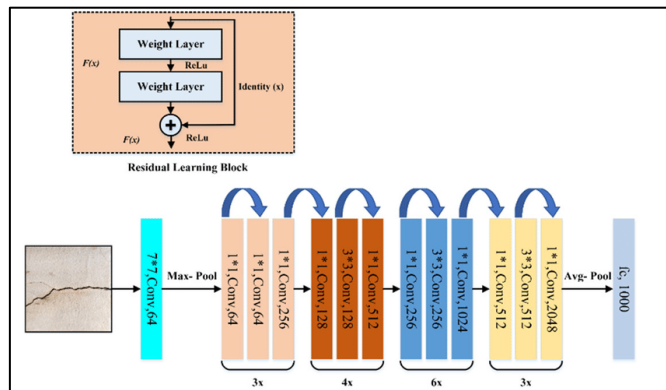


Figure 3. ResNet50 model architecture

### 2.4 InceptionV3

InceptionV3 uses modules called Inception, which act as multiple filters applied to the same input value through convolutional and pooling layers, which allows taking advantage of the extraction of patterns that provide different sizes in the filters; the result of these filters is concatenated and used as the output value of the module. InceptionV3 increases the trainable parameters and the computation required; however, it greatly improves accuracy. It applies three convolution architectures such as  $1 \times 1$ ,  $3 \times 3$ , and  $5 \times 5$ ; this is

## 3. RELATED WORK

To obtain works related to the research, a search was carried out in the indexed databases of Scopus, IEEE, EBSCOhost, Wiley, and ACM using the Systematic Review of Literature – RSL methodology, applying keywords of “deep learning” and “violence” in addition to considering only the works that used CNN models with Transfer Learning techniques dedicated to classifying violent actions, at the end, a summary was made with the most important data, such as the pre-trained models used by the authors, the original pre-training dataset, the model proposed by the researcher and the result obtained with the “accuracy” metric that allowed perform a comparative analysis on this metric, as detailed in Table 1.

From the analysis of the data in Table 1, it is estimated that the most used CNN models with Transfer Learning within similar research are VGG16, MobileNet, ResNet50, and InceptionV3 in addition to YOLO; however, the latter needs to work with labeling techniques to detect objects in the image, but this research is limited to working with image classification models based on their total dimension. For this purpose, a personalized dataset is used with images of simulated violent actions, which were trained with the four models, to analyze and consider the performance results obtained.

**Table 1.** References that use CNN models with transfer learning

References	Models	Dataset Pretraining	Proposed Model	Acc (%)
[11]	MobileNet	ImageNet	Conv LSTM	91.00
[12]	MobileNet	ImageNet	-	89.33
[13]	InceptionV1	ImageNet + Kinetics	2-Stream 3D-CNN	98.00
[14]	ResNet50	ImageNet	CNN+ LSTM	93.63
[15]	MobileNet	ImageNet	-	95.41
[16]	VGG16	INRIA Person	2D BiGRU-CNN	94.58
[17]	VGG16	ImageNet	VGG16+ LSTM	94.70
[18]	YOLOv2	COCO	-	88.00

## 4. METHODS AND PROPOSED DESIGN

### 4.1 Dataset development

In this stage of preparation and collection of personalized data, it was decided to develop a dataset with simulated violent actions in urban areas, taking into account the parameters of behaviors or body movements classified as strangulation, grappling, kicking, and punching [26], which were developed with the participation of young university students, who provided support for this research voluntarily using their cell phones to make the various video recordings, which were then processed to obtain the most appropriate images for each classification and which finally became part of the training dataset.

**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
<b>TOTAL</b>	<b>1200</b>	<b>600</b>	<b>200</b>	<b>2000</b>

During this process, the challenge of choosing the most appropriate recording setting was presented: finding a free and spacious urban area, finding a location and support for the cell phone, and trying to find an adequate height. and safe; since different models of high-end cell phones with twelve (12) megapixel resolution cameras were used, the idea was to simulate capturing the images as if a camera did it from a municipal urban video surveillance system. In addition, the following measures and actions were considered to guarantee diversity within the collection of data:

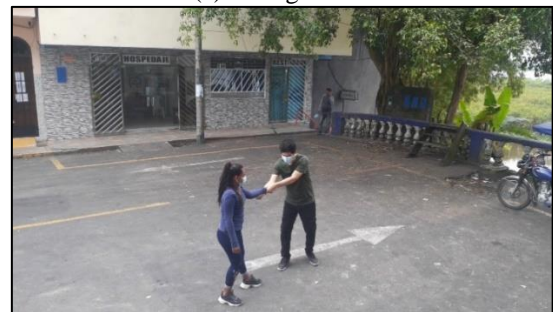
- The cell phone was placed on a stand or tripod at a downward inclination of 15°.
- The cell phone was placed on a support vertically at a height of 2 meters and the reception area horizontally at a maximum of 4 meters.
- Video recordings were made with simulated violent actions for an average of 10 minutes and with a maximum of 2 participants.
- The maximum distance between the camera and the violence zone for detection is 4 meters.
- The cell phone camera was not exposed to rainy conditions, making recognition difficult.
- No captures were made in areas with crowds of people that obstruct the view of the cell phone camera.

In total, 100 short videos were selected (25 videos for each type of action) of approximately three (3) minutes each, from

which two thousand (2000) images with a size of 1920×1088 pixels were selected with the help of the “VLC Media Player” software, as you can see an image of each type of violent action in Figure 5, finally, the 2000 images were distributed as seen in Table 2 and placed in folders as seen in Figure 6.



(a) Strangulation



(b) Grappling



(c) Kicking



(d) Punching

**Figure 5.** Example of images selected for the dataset

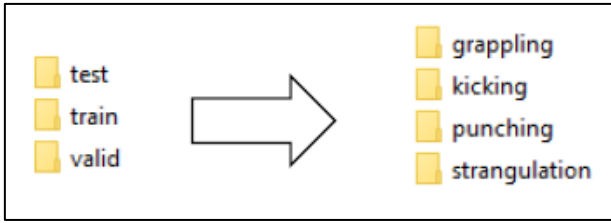


Figure 6. Distribution of the dataset folders

#### 4.2 Training CNN models with transfer learning

The four selected CNN models were trained with the customized dataset. At first, training began through Python, but when there were disconnection problems during the process, it was decided to do it locally using a 16 GB laptop. of RAM and GPU model GeForce GTX 1660 Ti with 8 GB of memory, for which the necessary Anaconda and Jupyter software libraries were installed. Before executing the training, the Data Augmentation technique was also applied to obtain more variants on the dataset's images, thus avoiding overfitting problems. Likewise, the Transfer Learning technique was then applied to each CNN model, taking advantage of its original weights acquired in its pre-training with ImageNet, where only its last network layers were adapted so that these models show one of the four as a response predefined type of physical violence. The sequence to be considered for the last additional layers of each model is detailed in Table 3.

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"

For training, the image inputs were resized to 224x224 pixels and 3 color channels (RGB) where each model was configured with the following characteristics or hyperparameters detailed in Table 4.

Table 4. Hyperparameters of pre-trained CNN models

Hyperparameter	Valor
Época	150
Bach size	16
Loss	"categorical_crossentropy"
Opmizer	"adadelata"
Metrics	"accuracy"

It is necessary to indicate that the hyperparameters considered result from evaluating the best state of the art to standardize each model. However, it can also be said that they can be different, and there is the possibility of creating a combination of values. of evidence between them, which would be a reason to continue conducting another research.

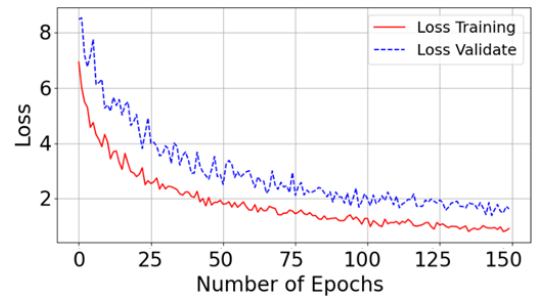
In the next research stage, the results obtained from the training, validation and testing of the pre-trained CNNs are disclosed. At the end, an analysis of the most important metrics or performance indicators is carried out through a comparative table.

## 5. RESULTS AND DISCUSSION

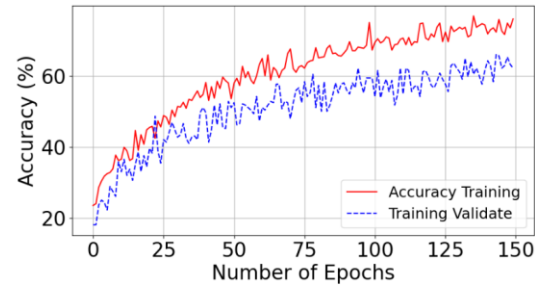
To execute the training, Python code was applied using Keras libraries, which allowed us to obtain each pre-trained model and where its final layers were subjected to the Transfer Learning technique using the same hyperparameters for each of them as mentioned above. At the end of each training, views of the "loss" and "accuracy" graphs were developed with the help of the Matplotlib and Sklearn libraries to display the confusion matrix.

Below, the graphs obtained from the training and validation processes are shown with respect to the loss and accuracy metrics, in addition, the confusion matrix of each pre-trained model is shown.

The VGG16 model was the first model executed, which lasted approximately 4 hours and 30 minutes of training; Figure 7(a) shows how the loss level decreases until it reaches a level less than 2 and Figure 7(b) shows how the accuracy level increases to more than 70% as the epochs progress.



(a) VGG16 loss chart



(b) VGG16 accuracy graph

Figure 7. Results of model VGG16

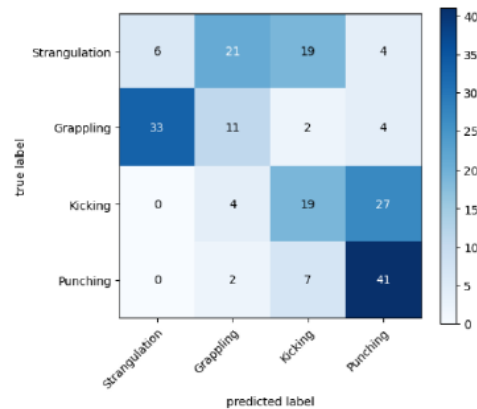


Figure 8. VGG16 confusion matrix

Figure 8 shows the results of the VGG16 confusion matrix, where 50 new violent images were tested for each class, of which, 41 images were correctly classified for the blown class, while the least accurate class was strangulation with only 6

images classified correctly. 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.

The MobileNet model was the second model executed, which lasted approximately 6 hours of training. Figure 9(a) shows how the loss decreases to levels less than 1 and Figure 9(b) shows how the accuracy level increases exceeding 80% as the training epochs progress.

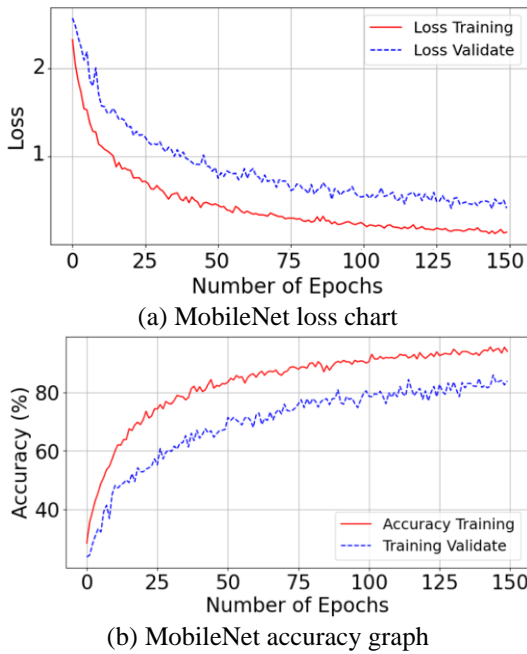


Figure 9. Results of model MobileNet

Figure 10 shows the results of the MobileNet confusion matrix, which was also tested with 50 new violent images of each class, but the results are better than the previous one since 49 and 48 images were correctly classified for the punching and wrestling classes respectively, however, the strangulation class remains the least accurate and continuous with only six images classified correctly. 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 differentiate these actions very well.

The third model to examine was ResNet50, which lasted approximately 6 hours and 18 minutes of training. Figure 11(a) shows how the training loss decreases to levels less than 0.5 and Figure 11(b) shows how the training accuracy level increases exceeding 90% as the training epochs progress, apparently it is one of the best performances.

Figure 12 shows the confusion matrix of the ResNet50 model and, as in all cases, the same 50 test images were also used, but it can be seen that the recognition results for the kick action are much better than the other models, since a total of 50 correctly classified images were obtained, however, there are several false positives related to the trumpet class, ResNet50 detected 26 and 33 classes of strangulation and

grappling respectively as if they were trumpet. Although, in this case false positives could help with the activation of an early warning, instead of having false negatives that would not help activate any type of alarm, however, it does not guarantee efficient work.

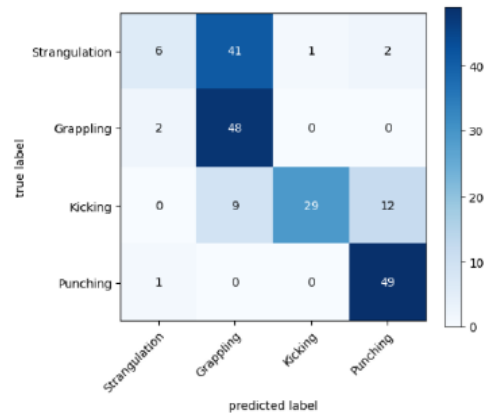


Figure 10. MobileNet confusion matrix

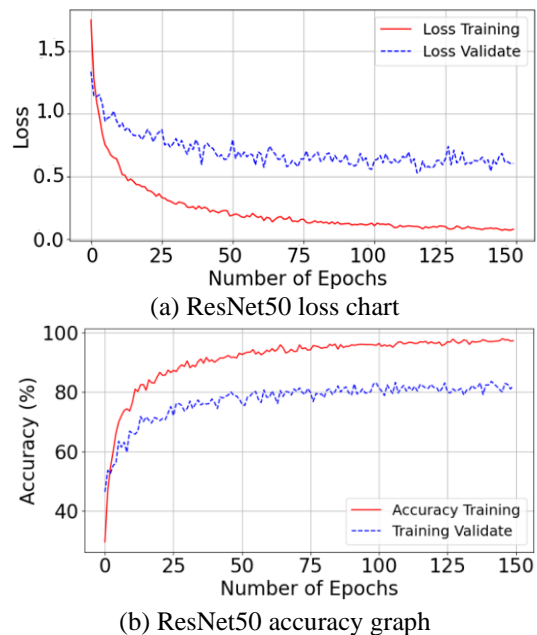


Figure 11. Results of model ResNet50

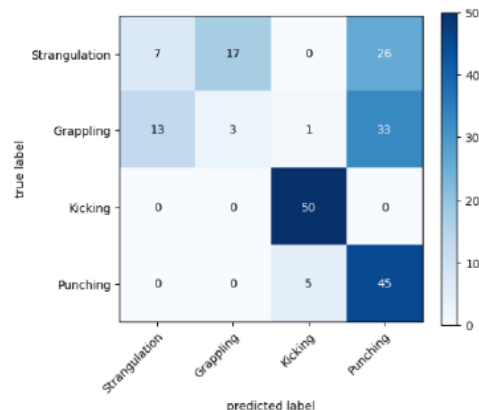


Figure 12. ResNet50 confusion matrix

The last model to examine was InceptionV3, which lasted approximately 5 hours and 24 minutes of training. Figure 13(a)

shows how the level of loss in both training and validation go closely together, decreasing to scales less than 1, while Figure 13(b) also shows a very favorable performance as accuracy levels increase and exceed 90%. As training iterations progress, it is also apparently one of the best performers.

Likewise, Figure 14 shows the data from the confusion matrix of InceptionV3, with the same 50 tests carried out, where it can be seen that the results are not very encouraging, although it is observed that the wrestling action correctly classified 50 out of 50, but there is a total confusion of false positives with the strangulation class, that is, for InceptionV3 the strangulation class is always taken as if it were a grappling, which can occur due to the similarity between both movements, likewise, the kick class is not recognized by the model since it does not have any correct classification.

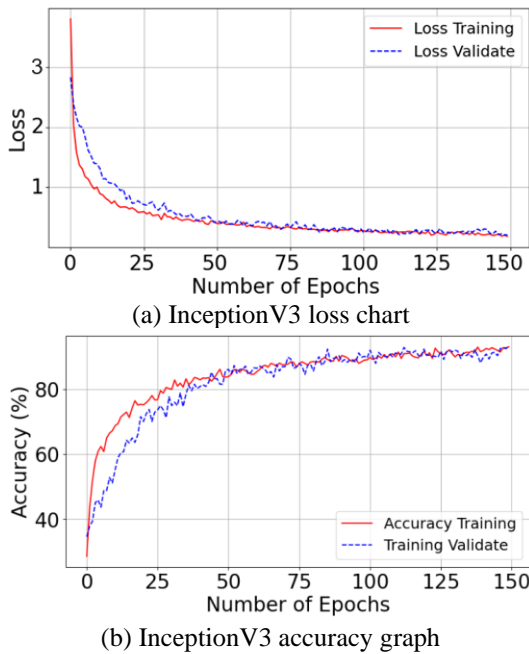


Figure 13. Results of model InceptionV3

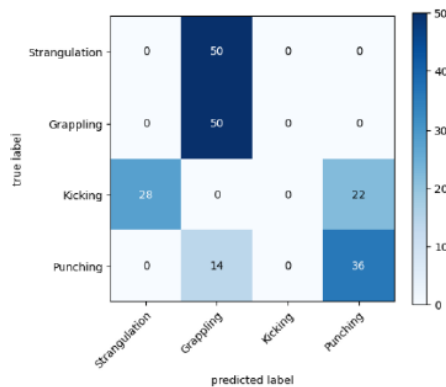


Figure 14. InceptionV3 Confusion matrix

The performance of each pre-trained CNN model was evaluated according to each class, taking into account the metrics as indicated by the state of the art [18, 19]. The accuracy metric was considered the most important metric; however, the other metrics must also be considered since they may be important for other types of studies. Table 5 below shows the summary of the results.

Finally, after evaluating and analyzing each of the performance results of the CNN models (see Table 5), we have

the following points of comparison: with an average level of precision of 72.53% and an accuracy level of 66.00%, the most effective model for classifying violent actions is the MobileNet model, with a training time of six hours and although the VGG16 model was trained with an hour and a half less, it was not considered very effective in providing a better solution, since, It only obtained 38.5% accuracy, placing it last in the ranking among the four. Likewise, it is necessary to mention the ResNet50 model, which placed itself in second place in the ranking for achieving an accuracy rate of 52.5%.

Then, with the results of the performance metrics averages, the research objective is demonstrated, which was to evaluate the speed and effectiveness of classifying violent actions using CNNs pre-trained using Transfer Learning techniques, being largely addressed by having considered the most important evaluation metrics.

Table 5. Results of the CNN models' performance metrics

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	
Mobile NetV2	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	
Res Net50	Strang.	35.00	14.00	88.28	20.00	52.5
	Grappling	15.00	6.00	85.71	8.57	
	Kicking	89.29	100	90.16	94.34	
	Punching	43.27	90.00	50.42	58.44	
	Avg	45.64	52.50	78.64	45.34	
Inception V3	Strang.	0.00	0.00	75.43	0.00	43.0
	Grappling	43.89	100	36.00	60.98	
	Kicking	0.00	0.00	100	0.00	
	Punching	62.07	72.00	69.44	66.67	
	Avg	26.48	43.00	70.22	31.91	

This research included a personalized dataset of two thousand (2000) images of interpersonal violence, which were developed under the same scenario. This can be a limitation when testing images from other urban areas. However, there are few investigations with a personalized dataset with more images; almost all the research reviewed works with videos of violence taken from Kaggle or another public repository.

In this context, it can be said that the use of personalized datasets is convenient as long as there are images within a variety of scenarios, simulating as much as possible to be realistic; that is, the movements of the simulated violent actions must be as spontaneous as possible, images should also be obtained from other urban areas or other countries, given that each area has its characteristics, all of this will help to generalize the algorithms, obtain better performance results and improve recognition of the different types of tests.

## 6. CONCLUSIONS

The development and use of a customized dataset provide significant advantages to reducing computational costs and better managing the hyperparameters of the algorithm configuration, as well as to identifying local violence actions

that are specific to the study area.

The performance metrics results estimate that the most effective pre-trained CNN model is MobileNet because it performs better than its peers VGG16, ResNet50, and InceptionV3, which helps to identify movements more accurately in violence recognition. However, its training requires more time, which is a limitation that requires future studies. Therefore, this research considers MobileNet to be the most suitable model to act as a violent action classifier, even for use on mobile devices, because its convolutional layers have better filters to extract essential features from an image such as those evaluated.

## 7. FUTURE SCOPE

For the next stage of this research, it is recommended to develop an intelligent video surveillance Web System by integrating it with the winning model, for which Python and AI libraries such as TensorFlow, Keras and OpenCV can be used that allow creation of applications quickly and with a code minimum. Its operation would be simple, the web camera would simply be located in a suitable place at a maximum height of 2 meters, calibrating the capture area at a distance of no more than 4 meters, in the end with just one click the system would begin to recognize the actions of physical violence showing labels around the detected images and when detecting actions of uninterrupted physical violence for 10 seconds, the system should emit an audible alarm through the PC or Laptop speakers.

It is also recommended to examine other CNN models pre-trained with Transfer Learning for object detection, using labeling or bounding box techniques, such as, for example, can evaluate the YOLOv8, Thunder, and PicoDet models, among others, which also provide their pre-trained weights in order to increase the values of the performance metrics, in addition, after demonstrating these tests, a prototype of an intelligent and real-time video surveillance system could begin to be implemented, to detect scenes of physical violence and automatically issue an alarm via text messages or emails.

## REFERENCES

- [1] Thomson, H., Bahgat, K., Urdal, H., Buhaug, H. (2023). Urban social disorder 3.0: A global, city-level event dataset of political mobilization and disorder. *Journal of Peace Research*, 60(3): 521-531. <https://doi.org/10.1177/00223433221082991>
- [2] Tavassoli, A., Soltani, S., Jamali, S.M., Ale Ebrahim, N. (2022). A research on violence against women: Are the trends growing? *Iranian Rehabilitation Journal*, 20(3): 425-440. <https://doi.org/10.32598/irj.20.3.1664.1>
- [3] Vidal, J., Mejía, L., Curiel, R. (2021). La violencia como fenómeno social: Dimensiones filosóficas para su evaluación. *Revista de Filosofía*, 38(99): 179-189. <https://doi.org/10.5281/zenodo.5644261>
- [4] Rettberg, A. (2020). Violence in Latin America today: Manifestations and impacts. *Social Studies Magazine*, 1(73): 2-17. <https://doi.org/10.7440/res73.2020.01>
- [5] Defense of the People. (2018). VIOLENCE AGAINST WOMEN: Victims' perspectives, obstacles and quantitative indices. <https://www.defensoria.gov.pe/wp-content/uploads/2018/09/Reporte-de-Adjunt%C3%ADa-2-2018-Violencia-contra-las-mujeres-Perspectivas-de-las-v%C3%ADctimas-obst%C3%A1culos-e-%C3%ADndices-cuantitativos.pdf>
- [6] Morales, S.K., Morales, M.K. (2024). Cultura de paz en medio de una creciente Violencia Social Peruana. *Revista Venezolana de Gerencia: RVG*, 29(105): 36-48. <https://doi.org/10.52080/rvgluz.29.105.3>
- [7] Jagtap, S., Chopade, N.B. (2024). Object-based image retrieval and detection for surveillance video. *International Journal of Electrical and Computer Engineering*, 14(4): 4343-4351. <https://doi.org/10.11591/IJECE.V14I4.PP4343-4351>
- [8] Sakiba, C., Tarannum, S.M., Nur, F., Arpan, F.F., Anzum, A.A. (2023). Real-time crime detection using convolutional LSTM and YOLOv7. Doctoral dissertation, Brac University.
- [9] Gaytán Aguilar, I., Aguilar Contreras, A., Alejo Eleuterio, R., Rendón Lara, E., Miranda Piña, G., Granda Gutiérrez, E.E. (2024). A comparative study of three pre-trained convolutional neural networks in the detection of violence against women. *Ciencia Ergo Sum*, 31. <https://doi.org/10.30878/ces.v31n0a17>
- [10] Mehmood, A. (2021). Abnormal behavior detection in uncrowded videos with two-stream 3d convolutional neural networks. *Applied Sciences*, 11(8): 3523. <https://doi.org/10.3390/app11083523>
- [11] Patel, M. (2021). Real-time violence detection using CNN-LSTM. *arXiv preprint arXiv:2107.07578*. <https://doi.org/10.48550/arXiv.2107.07578>
- [12] Haque, M.R., Hafiz, R., Al Azad, A., Adnan, Y., Mishu, S.A., Khatun, A., Uddin, M.S. (2021). Crime detection and criminal recognition to intervene in interpersonal violence using deep convolutional neural network with transfer learning. *International Journal of Ambient Computing and Intelligence (IJACI)*, 12(4): 154-167. <https://doi.org/10.4018/IJACI.20211001.oa1>
- [13] Traoré, A., Akhloufi, M.A. (2020). 2D bidirectional gated recurrent unit convolutional neural networks for end-to-end violence detection in videos. In *International Conference on Image Analysis and Recognition*, pp. 152-160. [https://doi.org/10.1007/978-3-030-50347-5\\_14](https://doi.org/10.1007/978-3-030-50347-5_14)
- [14] Soliman, M.M., Kamal, M.H., Nashed, M.A.E.M., Mostafa, Y.M., Chawky, B.S., Khattab, D. (2019). Violence recognition from videos using deep learning techniques. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pp. 80-85. <https://doi.org/10.1109/ICICIS46948.2019.9014714>
- [15] Lumba, A.P.L., Rios Nunez, R.R., Yahuarcani, I.O., Vigo, R.C., Cortegano, C.A.G., Pezo, A.R., Satalaya, A.M.N., Gomez, E.G., Llaja, L.A.S. (2019). 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. In *2019 XLV Latin American Computing Conference (CLEI)*, pp. 1-9. <https://doi.org/10.1109/CLEI47609.2019.235100>
- [16] Omarov, B., Narynov, S., Zhumanov, Z., Gumar, A., Khassanova, M. (2022). State-of-the-art violence detection techniques in video surveillance security systems: A systematic review. *PeerJ Computer Science*, 8: e920. <https://doi.org/10.7717/peerj-cs.920>
- [17] Jadhav, C., Ramteke, R., Somkunwar, R.K. (2023).



- Smart crowd monitoring and suspicious behavior detection using deep learning. *Revue d'Intelligence Artificielle*, 37(4): 955-962. <https://doi.org/10.18280/ria.370416>
- [18] Enciso Ortiz, S.E. (2024). Determination of the best Convolutional Neural Network Architecture: VGG16, ResNet50 or MobileNet for Pneumonia detection 2023.
- [19] Lopez-Betancur, D., Bosco Durán, R., Guerrero-Mendez, C., Zambrano Rodríguez, R., Saucedo Anaya, T. (2021). Comparison of convolutional neural network architectures for COVID-19 diagnosis. *Computing and Systems*, 25(3): 601-615. <https://doi.org/10.13053/cys-25-3-3453>
- [20] Alrubaie, H.D., Aljobouri, H.K., Aljobawi, Z.J. (2023). Efficient feature selection using CNN, VGG16 and PCA for breast cancer ultrasound detection. *Revue d'Intelligence Artificielle*, 37(5): 1255-1261. <https://doi.org/10.18280/ria.370518>
- [21] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. In 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. <https://arxiv.org/abs/1409.1556v6>
- [22] Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Van Esesn, B.C., Awwal, A.A.S., Asari, V.K. (2018). The history began from AlexNet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*. <https://arxiv.org/abs/1803.01164v2>
- [23] Dong, K., Zhou, C., Ruan, Y., Li, Y. (2020). MobileNetV2 model for image classification. In 2020 2nd International Conference on Information Technology and Computer Application (ITCA), Guangzhou, China, pp. 476-480. <https://doi.org/10.1109/ITCA52113.2020.00106>
- [24] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778. <https://doi.org/10.1109/CVPR2016.90>
- [25] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 2818-2826. <https://doi.org/10.1109/CVPR.2016.308>
- [26] Yahuarcani, I.O., Diaz, J.E.G., Satalaya, A.M.N., Noriega, A.A.D., Cachique, F.X.L., Llaja, L.A.S., Pezo, A.R., Rojas, A.E.L. (2021). Recognition of violent actions on streets in urban spaces using Machine Learning in the context of the COVID-19 pandemic. In 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), Cape Town, South Africa, pp. 1-4. <https://doi.org/10.1109/ICECET52533.2021.9698762>