



# A Comparative Study of Three Pre-trained Convolutional Neural Networks in the Detection of Violence Against Women

Gaytán Aguilar, Ivan; Aguilar Contreras, Alejandro; Alejo Eleuterio, Roberto; Rendón Lara, Eréndira; Miranda Piña, Grisel; Granda Gutiérrez, Everardo E.

A Comparative Study of Three Pre-trained Convolutional Neural Networks in the Detection of Violence Against Women

CIENCIA *ergo-sum*, vol. 31, 2024 | e232

Ciencias Exactas y Aplicadas

Universidad Autónoma del Estado de México, México

Esta obra está bajo una Licencia Creative Commons Atribución-NoComercial-SinDerivar 4.0 Internacional.



Gaytán Aguilar, I., Aguilar Contreras, A., Alejo Eleuterio, R., Rendón Lara, E., Miranda Piña, G. y 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. <http://doi.org/10.30878/ces.v31n0a17>

# A Comparative Study of Three Pre-trained Convolutional Neural Networks in the Detection of Violence Against Women

## Estudio comparativo de tres redes neuronales convolucionales preentrenadas en la detección de violencia contra las mujeres

*Ivan Gaytán Aguilar*

*Tecnológico Nacional de México, Campus Toluca, México*

mm20280290@toluca.tecnm.mx

 <https://orcid.org/0000-0002-4101-5351>


Recepción: 29 de agosto de 2022

Aprobación: 2 de diciembre de 2022

*Alejandro Aguilar Contreras*

*Tecnológico Nacional de México, Campus Toluca, México*

acontrerasa@toluca.tecnm.mx

 <https://orcid.org/0000-0003-1493-2987>

*Roberto Alejo Eleuterio*

*Tecnológico Nacional de México, Campus Toluca, México*

ralejoe@toluca.tecnm.mx

 <https://orcid.org/0000-0002-7580-3305>

*Eréndira Rendón Lara*

*Tecnológico Nacional de México, Campus Toluca, México*

erendonl@toluca.tecnm.mx

 <https://orcid.org/0000-0003-4581-6022>

*Grisel Miranda Piña\**

*Tecnológico Nacional de México, Campus Toluca, México*

mm22280266@toluca.tecnm.mx

 <https://orcid.org/0000-0001-7122-0658>

*Everardo E. Granda Gutiérrez*

*Universidad Autónoma del Estado de México, México*

eegrandag@uaemex.mx

 <https://orcid.org/0000-0002-9316-9627>

### RESUMEN

Se presenta una comparación de rendimiento entre tres modelos de redes CNN preentrenadas (VGG16, ResNet50 y MobileNet) en la detección en video de violencia física contra la mujer. Para llevar a cabo la clasificación de imágenes que incluyan violencia física contra la mujer y aquellas que no, se recolectaron 2 800 imágenes (1 400 violentas y 1 400 no violentas) de un Dataset público y posteriormente fueron divididas en entrenamiento (1 200 imágenes), validación (1 000 imágenes) y prueba (600 imágenes). Para evaluar su rendimiento, se tomaron en cuenta los valores de exactitud para cada modelo; al respecto, la red MobileNet se posiciona como el clasificador con mejor rendimiento para esta tarea de clasificación con 89% de exactitud.

**PALABRAS CLAVE:** inteligencia artificial, aprendizaje profundo, violencia contra la mujer, transferencia de aprendizaje, CNN, VGG16, ResNet50, MobileNet.

---

\*AUTORA PARA CORRESPONDENCIA

mm22280266@toluca.tecnm.mx

## ABSTRACT

This paper presents a performance comparison between three models of pre-trained CNN networks (VGG16, ResNet50, and MobileNet) in detecting physical violence against women in video. To carry out the classification of images that include physical violence against women and those that do not, 2 800 images (1 400 violent and 1 400 non-violent) were collected from a public dataset and subsequently divided into training (1 200 images), validation (1 000 images) and test (600 images). To evaluate their performance, accuracy values for each model were considered, positioning the MobileNet network as the best-performing classifier for this classification task with 89% accuracy.

**KEYWORDS:** artificial intelligence, deep learning, violence against women, transfer learning, CNN, VGG16, ResNet50, Mobile Net.

## INTRODUCTION

Globally, women are subjected to physical, sexual, and psychological abuses that transcend classes and cultures. These kinds of violence are recognized as human rights violations and discrimination against women (McQuigg, 2018). The 1993 Declaration on the Elimination of Violence against Women called on States to promote research, collect data and develop statistics about different forms of violence against women, especially domestic violence. It also encouraged research on the causes, nature, and consequences of violence against women and the effectiveness of measures to prevent and redress it. Likewise, this declaration called to end all forms of violence against women (Fried, 2003; Schwartz, 2000).

Available statistics issued by police, women's centers, and other formal institutions often underestimate levels of violence due to underreporting; the percentage of women seeking help is less than 10% in almost all countries (WHO, 1997). For this reason, there is a need to obtain accurate and comparable data on violence against women at the community, national and international levels to strengthen promotion efforts that can measure the true prevalence of violence. However, it is a complex task; first, the lack of consistent methods and subjective definitions of violence makes statistical analysis difficult (Castorena *et al.*, 2021); second, the severity of recorded violence may vary between victims (Qureshi, 2020).

Physical violence is any act that injures the victim with actions such as pushing, grabbing, twisting the arm, pulling hair, slapping, kicking, biting, or hitting with a fist or any object, attempting to strangle, suffocate, burning, scalding on purpose or attack with some weapon (Lea, 1993).

This problem is an everyday situation because no monitoring systems alert about the event and provide help to the victim, especially in the street or public spaces. Although surveillance cameras have been placed to increase the security for citizens and many people are employed by multiple institutions to monitor all events that are recording the cameras, there is still a probability that the person in charge of a specific area will miss some events that include abnormal things due to human error. That supposes a loss of time and energy. Moreover, monitoring all of them is complex due to the number of people required for real-time monitoring (Peixoto *et al.*, 2018; Vosta & Yow, 2022).

To deal with the security problem in surveillance cameras, many researchers have worked on multiple proposals to detect abnormal events reducing incidents such as kidnappings, robberies, murders, rapes, domestic violence, and gender violence, among others, whose recommendations include the implementation of machine and deep learning-based modern techniques, mainly: speech recognition, image processing, and pattern recognition (Alexandrie, 2017; Roa *et al.*, 2018; Ye *et al.*, 2022). Then, a surveillance camera system may be able to detect if something abnormal is happening and notify to get help immediately. Based on this, we present an implementation and comparison of three well-known pre-trained Convolutional Neural Network (CNN) models for detecting physical violence against women in video scenes.

## 1. RELATED WORKS

Physical violence against women has emerged as a critical problem that violates the rights of millions of women worldwide and has economic, health, and social repercussions for victims, perpetrators, and society (Ferrari *et al.*, 2022).

Multiple investigations to address this issue suggest using new and more advanced techniques capable of immediately detecting scenes of physical violence without omitting data. Some proposals to address this problem are based on implementing algorithms based on machine learning. For example, Hanson *et al.* (2019) use a novel method for detecting violence and fights in detention centers and psychiatric hospitals. The novelty of the approach consists of applying a transformation to a sequence of consecutive video frames, a technique known as extreme pattern acceleration. This work showed that it is possible to increase violence identification by up to 12% concerning the state-of-the-art results.

Patel (2021) shows a model that uses deep learning to recognize if a violent movement is taking place in real-time and, if so, alert on a said event to act. His work showed accuracy results of 92.3 % for the Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) hybrid model with the pre-trained model ResNET50.

A method based on convolutional networks to detect abuse against women in a surveillance system is proposed by Sandhiya *et al.* (2020). This method first identifies that the man and the woman are in the place and then detects the movements of abuse. Their article concludes that their method correctly detects the event using the MCF (Motion Co-occurrence Feature) technique.

Likewise, Dandamudi *et al.* (2020) present an application using a drone with a CNN based on an image processing model to detect multiple issues in society, such as robberies, weapons usage, fire accidents, violence against women (VAW), and sexual gender-based violence (SGBV). They implemented the application with a Raspberry Pi, a GPS Neo 6m, an HC-SR04 ultrasonic sensor, and an Arduino connected with a PI camera.

Another work related to the detection of physical violence is presented by Vijeikis *et al.* (2022); the authors introduce a novel architecture to detect violence in surveillance cameras using MobileNet as the pre-trained model, showing results of 82% accuracy and 81% precision.

Based on the state-of-art papers addressing the detection of violence in video scenes, this work compares three pre-trained models to obtain a better method to detect violence against women in video using CNN as the base architecture to carry out the experiments.

## 2. THEORETICAL FOUNDATION

Since the second half of the twentieth century, machine learning (ML) has revolutionized scientific fields, technology, and our day-to-day lives; it has evolved as a subfield of artificial intelligence thanks to the increase in quantity and diversity of data measurements (Deng & You, 2014). Today, one of the most used ML-based methods is deep learning (DL), characterized by models like neural networks, as it can learn patterns in images automatically and have high levels of abstraction (Arel *et al.*, 2010).

One of the dominant architectures in image analysis and processing is the convolutional network (CNN), explicitly designed for images. It had its most significant impact in 2012 with the ImageNet challenge (Soffer *et al.*, 2021). The CNN architecture (see Figure 1) is comprised of at least one layer with convolutional operations and a pooling layer to reduce the dimensionality of the features by saving the most critical features, whose output will be passed to a fully connected neural network (multilayer perceptron) to obtain the final and desired output (Bilbro *et al.*, 2019; Naranjo-Torres *et al.*, 2020). The main advantage of this model that makes it a suitable feature extractor for sequential and image data sets is the extraction of complex hidden features from high dimensional data with a complex structure (Vosta & Yow, 2022). Extracted deep features have been used in lesion classification, person identification, and image quality assessment tasks. Although they have been implemented for various deep-learning tasks, they are primarily employed in computer vision to solve image classification and anomaly detection in images and video (Alzubaidi *et al.*, 2021).

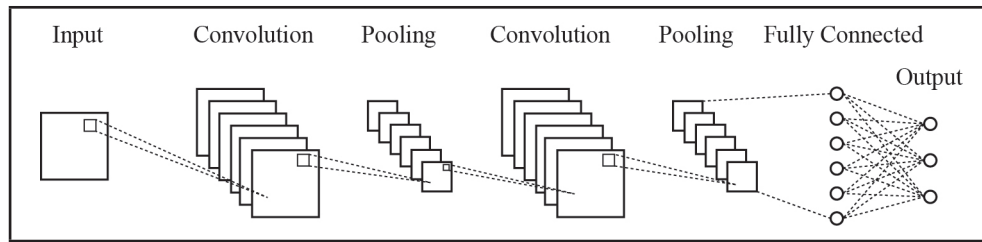


FIGURE 1

A general CNN architecture with two convolutional layers and two pooling layers.

Source: self made.

CNN networks have been generating more efficient alternatives that optimize the computational cost of training a network from scratch over the years. These alternatives are based on transfer learning, a technique used to learn new patterns from new data when there is not enough data to train (Abu *et al.*, 2021). Some of the best-known CNN-based methods in transfer learning are:

- a) VGG16 Model. It is the first deep neural architecture predecessor of the AlexNet model, which is comprised of multiple convolutions and fully connected layers. The VGG16 (see Figure 2) CNN model counts with many image recognition parameters. In 2014, the VGG16 model with 16 layers achieved a test accuracy of 92.7 % in the top 5 on ImageNet classification with an error of 7.3 % (Gujjar *et al.*, 2021).
- b) ResNet50 model. The ResNet50 model (see Figure 2) is part of the ResNet family, whose architecture was introduced to avoid dead connections and reduce the effect of layers on performance and the number of deep layers depends on its version, e.g., in this case, ResNet50 has 50 deep layers (Abu *et al.*, 2021). This deep residual CNN model has produced ground-breaking performances in image recognition, object detection, face recognition, and image classification. Its architecture has four stages that fix the degradation problem in the deeper CNN (Gujjar *et al.*, 2021; Theckedath & Sedamkar, 2020).
- c) MobileNet model. Models that require fewer parameters and reduce the computational cost are gaining relevance, one of them is MobileNet (see Figure 2), a model designed to use a depth-separable convolution in addition to using only two hyperparameters, a length multiplier, and a resolution multiplier to obtain a balance between accuracy and latency (Chen & Su, 2018). MobileNet is a class of CNN that was open source by Google, giving an excellent starting point for training the classification, detection, embeddings, and segmentation (Gujjar *et al.*, 2021).

One way to visualize the performance of a binary classifier is the binary confusion matrix (Luque *et al.*, 2019). This tool provides a detailed view of performance with correct and incorrect predictions for each of the classes involved; the classifier performance is obtained by comparing actual and predicted values (Tangirala, 2020).

Table 1 graphically describes the binary confusion matrix, where *tp* corresponds to correctly predicted positive samples, *fn* is the number of positive samples indicated as negative, *fp* is the number of the negative samples predicted as positive, and *tn* is the number of the correctly predicted negative samples.

TABLE 1  
Confusion matrix for binary classification

		Predicted Class	
		Positive	Negative
True Class	Positive	True Positive ( <i>tp</i> )	False Negative ( <i>fn</i> )
	Negative	False Positive ( <i>fp</i> )	True Negative ( <i>tn</i> )

Source: self made.

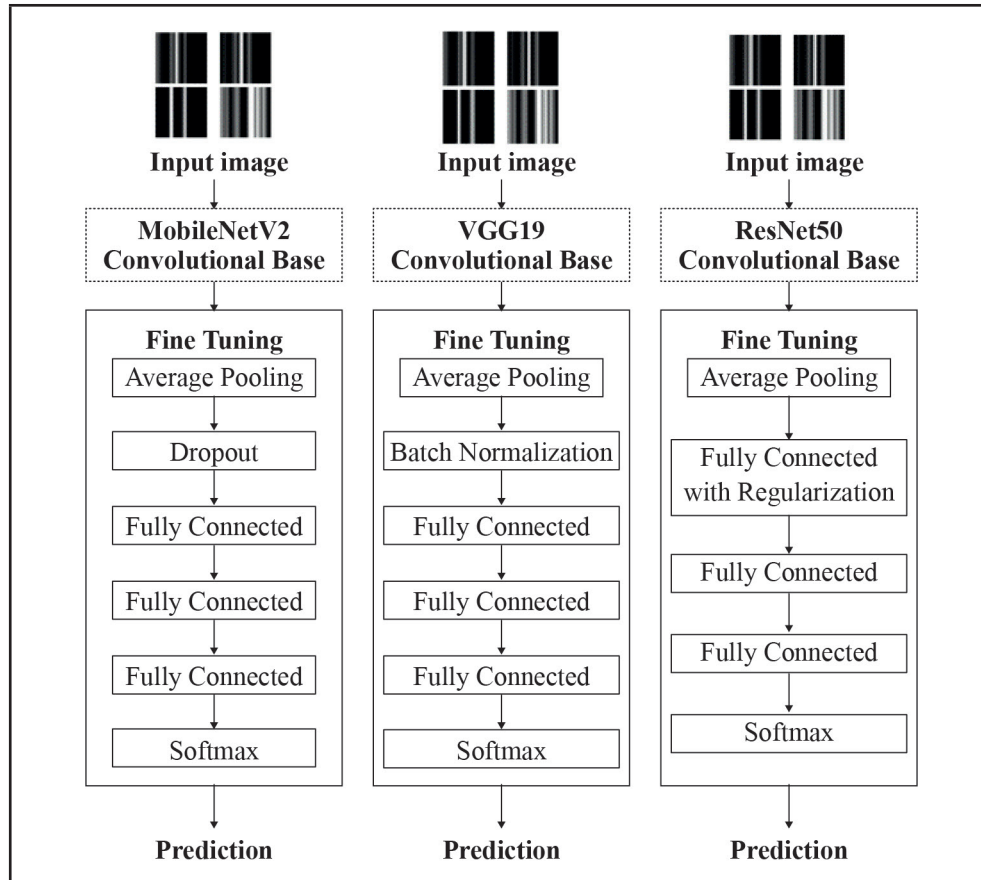


FIGURE 2  
 General architecture of models: VGG16, ResNet50, MobileNet  
 Source: Gayathri *et al.*, 2020.

An advantage of confusion matrices is that they can be used to measure the main parameters to evaluate the classification performance: recall or sensitivity (eq. 1), specificity (eq. 2), precision (eq. 3), accuracy (eq. 4), F1-score (eq. 5), and AUC-ROC curve, which represents the Area Under de Curve (AUC) of the Receiver Characteristic Operator (ROC) and means the evaluation metric of the probabilistic plot of the true positive rate (sensitivity) against the false positive rate (1-sensitivity), being helpful to identify the performance of the model to distinguish between positive and negative (Beauxis-Aussalet & Hardman, 2014).

$$recall = \frac{tp}{tp + fn} \tag{1}$$

$$specificity = \frac{tn}{tn + fp} \tag{2}$$

$$precision = \frac{tp}{tp + fp} \tag{3}$$

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{4}$$



$$F1 = \frac{2(\text{recall})(\text{precision})}{\text{recall} + \text{precision}} \quad (5)$$

### 3. METHODOLOGY

This section presents and describes the methodological aspects, including the hardware and software used, the image processing method, and the training and validation stages for the models proposed in this study.

#### 3. 1. Software and Hardware

The operating system used to perform the necessary experiments was Linux (Ubuntu distribution), where Anaconda software was installed to create a development environment for the Python programming language. All tests were carried out on a computer that has a 4 GB NVIDIA GeForce GTX 1650TI graphics card with CUDA (Unified Computing Device Architecture) support. TensorFlow (an open source machine learning library that can run on multiple CPUs and GPUs, with CUDA support) (Stančin & Jović, 2019), with the programming interface Keras to the design, configuration, and training of artificial neural networks (traditional or deep). Likewise, JupyterLab was chosen as the web-based interactive development environment since it covers the flexibility needed to configure and organize the user interface to support the wide range of data workflows thanks to its plugins that add new components and integrate with existing ones (Bisong, 2019).

#### 3. 2. Dataset and image processing

Dataset was built from different videos; they were obtained by filtering keywords related to physical violence against women. Each frame of the videos was extracted using the Python programming language and the OpenCV open source library (Bradski, 2000). Next, each video frame was manually classified according to where violence was evident and where it was not, yielding a dataset of 2 800 samples consisting of 1 400 violent and 1 400 non-violent images. Then, images were normalized in size and grayscale. Afterward, they were divided into three sets: the training set (1 200 images), the validation set (1 000 images), and the test set (600 images); it is essential to mention that all subsets are composed of 50% of each class, i.e., the training set has 600 violent and 600 non-violent samples. Figure 3, shows the example for both sample cases, either scene with violence against the woman (left) or a non-violent scene (right). It is worth mentioning that the dataset cannot be provided to protect personal data and the right to privacy and to obey national and international data privacy regulations that could apply. However, images or screenshots from typical image or video repositories could be used to evaluate this approach.

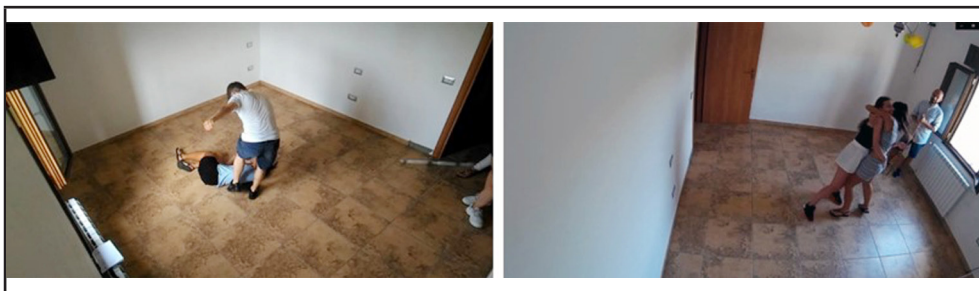


FIGURE 3

Examples of images for the two classes involved in the classification in scenes with violence against the woman (left) or a non-violent scene (right).

Source: Bianculli *et al.*, 2020.

### 3.3. Deep learning models

Deep learning models on image classification, especially CNN, have been successfully applied in many computer vision applications to classify object recognition, image analysis, image colorization, scene labeling, etc. Once the data had been previously divided and classified manually, the next step was to make a workflow diagram (see Figure 4), where the activities that the program encoded with the algorithm must carry out until reaching its validation were established; in this way, define the instructions to be programmed in each model are clear and precise.

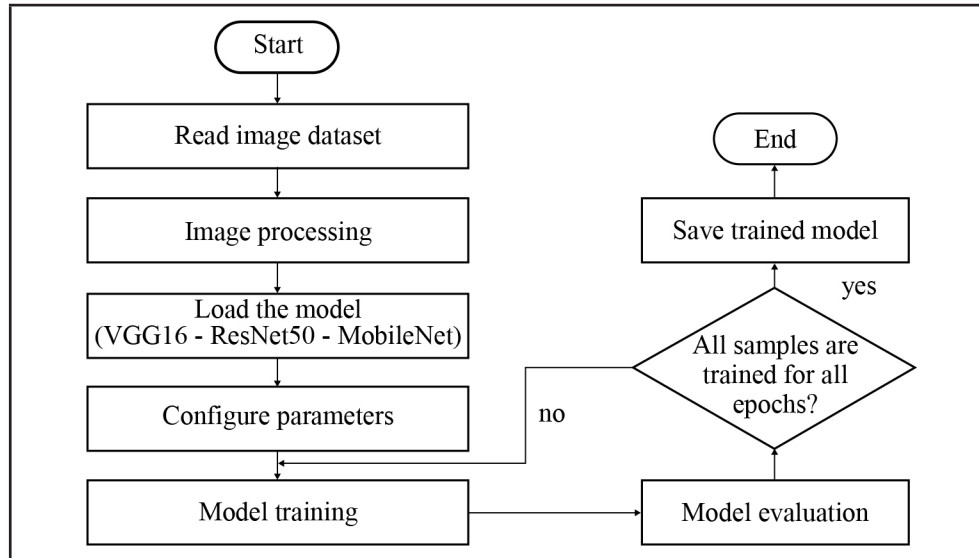


FIGURE 4  
Workflow diagram for each proposed classifier model.  
Source: self made.

The components of each algorithm share the number of input and output layers; two convolutional layers were placed for the input and two dense layers for the output. In this case, the difference lies in the hidden layers of each model; for the VGG16 model, there were 16 layers (13 convolutional layers and 3 dense layers), ResNet with 152 convolutional layers, and MobileNet with 19 convolutional layers. The models mentioned above are available as part of the OpenCV libraries, but the tuning and configuration must be performed as depicted here to test the results.

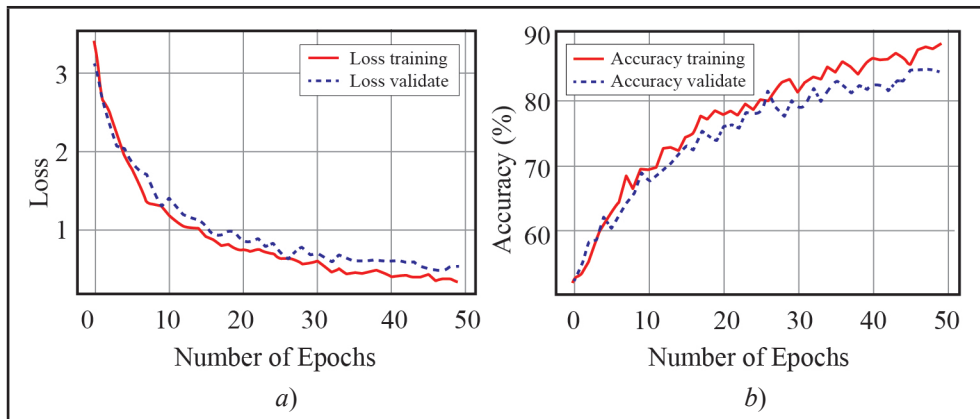
Finally, the performance of the deep neural network was evaluated using two metrics commonly used in deep learning: the confusion matrix and the accuracy (a number between 0 and 1 indicating how successful the evaluation was).

## 4. RESULTS AND DISCUSSION

This section shows the results obtained from the carried-out experimentation for each model used.

The first neural network model verified was VGG16, whose training was carried out at an execution time of 38 min 49 sec. The corresponding tests were carried out using the test images, i.e., images the model has never seen. The neural network performs using the pre-trained VGG16 algorithm, as depicted in Figure 5a, which shows how the loss level decreases. Figure 5b shows how the precision percentage increases when the training epochs evolve.





**FIGURE 5**  
 Graphs of loss and accuracy for the VGG16 model  
 Source: self made.  
 Note: a) loss, b) accuracy.

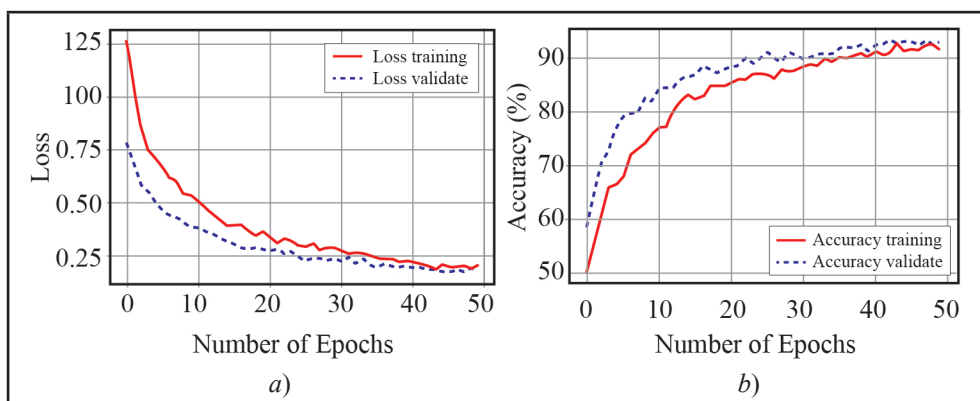
Table 2 shows the confusion matrix results, where 300 violent images were tested. Still, only 259 pictures were correctly classified, and 41 were not, while with the 300 non-violent photos, 217 were correctly classified, and 83 images did not.

**TABLE 2**  
 Confusion matrix for VGG16 model

		Predicted Class	
		Nonviolence	Violence
True Class	Nonviolence	217	83
	Violence	41	259

Source: self made.

The second neural network evaluated was the ResNet50 model, reporting a training time of 50 min 20 sec. The performance evaluation was performed similarly to the first model, using the same images. In more detail, Figures 6a and 6b show how the accuracy percentage goes up and the loss goes down as the epochs increase; this model had better values than the first one.



**FIGURE 6**  
 Graphs of loss and accuracy for the ResNet50 model  
 Source: self made.  
 Note: a) loss, b) accuracy.

For violent image samples, the confusion matrix shows that 287 were correctly classified while 13 were not. In contrast, for non-violent images, 245 were correctly classified, and 55 were not (see Table 3).

TABLE 3  
Confusion matrix for ResNet50 model

		Predicted Class	
		Nonviolence	Violence
True Class	Nonviolence	245	55
	Violence	13	287

Source: self made.

Finally, the last model used was MobileNet, which lasted 1 h, 20 min, and 55 sec in its training stage. Following the evaluation process carried out in the previous models, the loss (Figure 7a) and precision (see Figure 7b) values were plotted. Like the previous examples, the precision percentage increases, and the loss falls as the epoch number increases.

The confusion matrix for the third experiment shows that of the 300 samples with violence, 294 were correctly classified, while of the 300 scenes with non-violence, 58 were incorrectly classified, offering better performance for the kind of violence compared to previous models (see Table 4)

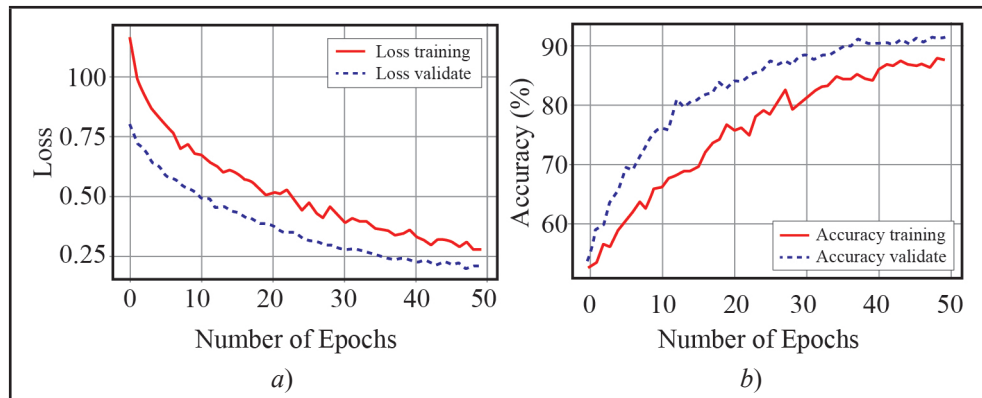


FIGURE 7

Graphs of loss and accuracy for the MobileNet model

Source: self made.

Note: a) loss, b) accuracy.

TABLE 4  
Confusion matrix for MobileNet model

		Predicted Class	
		Nonviolence	Violence
True Class	Nonviolence	242	58
	Violence	6	294

Source: self made.

Other metrics such as precision, recall, F1-score, and accuracy were calculated as supplemental performance measures to obtain more details about the performance of each model and to choose the best pre-trained model concerning the best performance for both classes. Table 5 shows that in the accuracy values, the violent class obtains the highest values with MobileNet in the lead; however, for the recall and the F1-score, the non-violent

class also gives the highest values, with MobileNet at the helm. Likewise, analyzing the scores of the accuracy metric, the MobileNet model obtains the highest value, again showing that this model performs better in classifying images with violence and non-violence.

The metrics proposed in this work are known as the most common indicators of the performance of classifiers because they complement the visual information provided by the confusion matrix (Canbek, 2021). Mainly F1-score and accuracy have been used in evaluating balanced and imbalanced datasets and are considered general quantitative indicators for the performance of models (Chicco, 2020; Brown, 2018). These metrics represent the overall performance of the tested models; consequently, the results obtained in this work identify the MobileNet as the best-performing model of the three evaluated.

TABLE 5  
Results of 5 measures obtained from the 3 pre-trained model

Model	class	precision	recall	F1-score	accuracy
VGG16	nonviolence	0.7573	0.8633	0.8069	0.7933
	violence	0.8411	0.7223	0.7778	
ResNet50	nonviolence	0.8392	0.9567	0.8941	0.8867
	violence	0.9496	0.8167	0.8781	
Mobile Net	nonviolence	0.8352	0.9800	0.9018	0.8933
	violence	0.9758	0.8067	0.8832	

Source: self made.

It is worth noticing that the class “violence” is the most representative class to be identified by the models because it is directly related to the phenomenon to be addressed (the presence of violence against women), as it is confirmed by the F1-score for the class mentioned earlier and by the overall value of the accuracy. However, the inaccurate identification of the “nonviolence” class could be related to a false alarm, i.e., an excess in the false positive cases. Thus, evaluating the recall, precision, and F1-score for the individual classes is relevant to evaluate the model that accurately identifies not only the true existence of violence but also avoids the emission of false alarms. In this sense, from Table 5, it can be observed that, again, the MobileNet model obtains the best results in the identification of the “nonviolence” class because it obtains the best values in all the evaluated parameters, in comparison to the other two models.

## PROSPECTIVE

Artificial Neural Networks (ANN) have been used in various commercial applications such as natural language manipulation through electronic devices or facial recognition, advanced web searches, and so on. Likewise, the latest generation ANN networks, such as convolutional or transformer networks, are currently the main architectures of technological development, whose models are showing a greater capacity to imitate human intelligence and carry out activities that were previously seen exclusively by humans, such as creativity (see GAN models). All the above highlights how powerful ANNs are in novel solutions to everyday problems.

Physical violence against women is a daily problem in all social and economic contexts. This paper tries to demonstrate how the ANNs can be used to treat physical violence against women. The main idea is that images obtained from videos can identify these types of violence and thus be the prelude to a more sophisticated application capable of sending real-time alerts to the corresponding authority when this type of activity is detected. Likewise, the results of this work can be extended to the detection of violence against groups of vulnerable people, such as immigrants or people with disabilities. However, this work is an initial stage of a broader and more far-reaching project in which technology is an ally in the fight for a society free of violence and fairer.

## CONCLUSION AND FUTURE WORKS

This article compares three pre-trained models based on the CNN architecture to detect physical violence against women in video images. The experimental results show that using pre-trained models to solve tasks such as violence detection in images with a limited data set provides a significant advantage by reducing the computational cost and the number of parameters to configure. Furthermore, based on the values generated from the metrics used, the MobileNet model had the best performance compared to VGG16 and ResNet50, but had the most expensive runtime. Its performance can be attributed to the fact that it is one of the later models than the first ones. Therefore, the convolutions have better filters that extract the essential characteristics of each image.

For future work, we consider using a dataset with more samples and other pre-trained CNN-based models to increase the accuracy value and, if possible, mount this project on a real-time video surveillance system to identify scenes where physical violence against women occurs and create an automatic alert.

## GRATITUDE

We want to thank the reviewers for their comments, which have served to improve the work.

## REFERENCES

- Abu, M., Amir, A., Lean, Y. H., Zahri, N. A. H., & Azemi, S. A. (2021). The performance analysis of transfer learning for steel defect detection by using deep learning. *Journal of Physics: Conference Series*, 1755(1), 12041.
- Alexandrie, G. (2017). Surveillance cameras and crime: A review of randomized and natural experiments. *Journal of Scandinavian Studies in Criminology and Crime Prevention*, 18(2), 210-222.
- Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1), 1-74.
- Arel, I., Rose, D. C., & Karnowski, T. P. (2010). Deep machine learning-a new frontier in artificial intelligence research [research frontier]. *IEEE Computational Intelligence Magazine*, 5(4), 13-18.
- Beauxis-Aussalet, E., & Hardman, L. (2014). *Visualization of Confusion Matrix for Non-Expert Users (Poster)*.
- Bianculli, M., Falcionelli, N., Sernani, P., Tomassini, S., Contardo, P., Lombardi, M., & Dragoni, A. F. (2020). A dataset for automatic violence detection in videos. *Data in Brief*, 33, 106587. <https://doi.org/10.1016/j.dib.2020.106587>
- Bilbro, R., Ojeda, T., & Bengfort, B. (2019). *Applied text analysis with Python*. O'Reilly Media Inc.
- Bisong, E. (2019). JupyterLab Notebooks. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform* (pp. 49-57). Springer.
- Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 25(11), 120-123.
- Brown, J. B. (2018). Classifiers and their Metrics Quantified. *Molecular informatics*, 37(1-2). 1700127. <https://doi.org/10.1002/minf.201700127>
- Canbek, G., Taskaya Temizel, T., & Sagioglu, S. Bench (2021). BenchMetrics: a systematic benchmarking method for binary classification performance metrics. *Neural Computing & Applications*, 33, 14623-14650. <https://doi.org/10.1007/s00521-021-06103-6>

- Castorena, C. M., Abundez, I. M., Alejo, R., Granda-Gutiérrez, E. E., Rendón, E., & Villegas, O. (2021). Deep neural network for gender-based violence detection on Twitter messages. *Mathematics*, 9(8), 807.
- Chen, H.-Y., & Su, C.-Y. (2018). An enhanced hybrid MobileNet. *2018 9th International Conference on Awareness Science and Technology (ICAST)*, 308-312.
- Chicco, D. & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(6), 13. <https://doi.org/10.1186/s12864-019-6413-7>.
- Dandamudi, A. G. B., Vasumithra, G., Praveen, G., & Giriraja, C. V. (2020). CNN Based Aerial Image processing model for Women Security and Smart Surveillance. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 1009-1017.
- Deng, L., & Yu, D. (2014). *Deep learning: Methods and applications*. Foundations and Trends.
- Ferrari, G., Torres-Rueda, S., Chirwa, E., Gibbs, A., Orangi, S., Barasa, E., Tawiah, T., Dwommoh Prah, R. K., Hitimana, R., Daviaud, E., &... Vasal, A. (2022). Prevention of violence against women and girls: A cost-effectiveness study across 6 low-and middle-income countries. *PLOS Medicine*, 19(3), e1003827.
- Fried, T. S. (2003). *Violence against Women, Health and Human Rights Violence, Health, and Human Rights*.
- Gayathri, R. G., Sajjanhar, A., & Xiang, Y. (2020). Image-based feature representation for insider threat classification. *Applied Sciences*, 10(14), 4945.
- Gujjar, J. P., Kumar, H. R. P., & Chiplunkar, N. N. (2021). Image classification and prediction using transfer learning in Colab notebook. *Global Transitions Proceedings*, 2(2), pp. 382-385.
- Hanson, A., PNVR, K., Krishnagopal, S., & Davis, L. (2019). Bidirectional Convolutional LSTM for the Detection of Violence in Videos. L. Leal-Taixé & S. Roth, (eds.), Computer Vision–ECCV 2018 Workshops. ECCV 2018. *Lecture Notes in Computer Science*, 11130, 280-295. [https://doi.org/doi.org/10.1007/978-3-030-11012-3\\_24](https://doi.org/doi.org/10.1007/978-3-030-11012-3_24)
- Lea, R. A. (1993). *World development report 1993. Investing in Health. World Development Indicators*. New York: Oxford University Press.
- Luque, A., Carrasco, A., Martín, A., & De Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91, 216-231.
- McQuigg, R. J. A. (2018). Is it time for a UN treaty on violence against women? *The International Journal of Human Rights*, 22(3), 305-324.
- Naranjo-Torres, J., Mora, M., Hernández-García, R., Barrientos, R. J., Fredes, C., & Valenzuela, A. (2020). A review of convolutional neural network applied to fruit image processing. *Applied Sciences*, 10(10), 3443.
- Patel, M. (2021). *Real-Time Violence Detection Using CNN-LSTM*. ArXiv Preprint ArXiv:2107.07578.
- Peixoto, B. M., Avila, S., Dias, Z., & Rocha, A. (2018). Breaking down violence: A deep-learning strategy to model and classify violence in videos. *Proceedings of the 13th International Conference on Availability, Reliability and Security*, 1-7.
- Qureshi, S. (2020). The recognition of violence against women as a violation of human rights in the United Nations system. *South Asian Studies*, 28(1).
- Roa, J., Jacob, G., Gallino, L., & Hung, P. C. K. (2018). Towards smart citizen security based on speech recognition. *2018 Congreso Argentino de Ciencias de La Informática y Desarrollos de Investigación (CACIDI)*, 1-6.
- Sandhiya, R., Gokul Prasad, A. R., Gokul Krishnan, D., & PrajethBalan, S. (2020). Women Abuse Detection in Video Surveillance using Deep Learning. *GRD Journals-Global Research and Development Journal for Engineering*, 5(4), 5.

- Schwartz, M. D. (2000). Methodological issues in the use of survey data for measuring and characterizing violence against women. *Violence Against Women*, 6(8), 815-838.
- Soffer, S., Klang, E., Shimon, O., Barash, Y., Cahan, N., Greenspan, H., & Konen, E. (2021). Deep learning for pulmonary embolism detection on computed tomography pulmonary angiogram: a systematic review and meta-analysis. *Scientific Reports*, 11(1), 1-8.
- Stančin, I., & Jović, A. (2019). An overview and comparison of free Python libraries for data mining and big data analysis. *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 977-982.
- Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, 11(2), 612-619.
- Theckedath, D., & Sedamkar, R. R. (2020). Detecting affect states using VGG16, ResNet50, and SE-ResNet50 networks. *SN Computer Science*, 1(2), 1-7.
- Vijeikis, R., Raudonis, V., & Dervinis, G. (2022). Efficient violence detection in surveillance. *Sensors*, 22 (6), 2216.
- Vosta, S., & Yow, K.-C. (2022). A CNN-RNN combined structure for real-world violence detection in surveillance cameras. *Applied Sciences*, 12(3), 1021.
- WHO (World Health Organization). (1997). *Violence against women: Definition and scope of the problem*. Geneva: The World Health Organization.
- Ye, L., Yan, S., Zhen, J., Han, T., Ferdinando, H., Seppänen, T., & Alasaarela, E. (2022). Physical violence detection based on distributed surveillance cameras. *Mobile Networks and Applications*, 1-12.

**CC BY-NC-ND**