



**GOBIERNO DE
MÉXICO**

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA



DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

MODALIDAD A.1, A.2 Y A.3 PROYECTO DE INVESTIGACIÓN CIENTÍFICA, TECNOLÓGICA O EDUCATIVA.

Título:

**Detección de Violencia Física en Imágenes Provenientes de Video, empleando
Redes Neuronales Artificiales**

Nombre del Docente:

ITZEL MARÍA ABUNDEZ BARRERA

Duración: Doce Meses

Fecha de Inicio: 1 de febrero 2023

Fecha de Terminación: 31 de enero 2024

1. Resumen del proyecto

Introducción: En este proyecto se abordó la problemática de identificación de violencia física en videos utilizando redes neuronales convolucionales.

Objetivos: los objetivos del proyecto fueron cumplidos ya que se logró publicar un artículo en una revista arbitrada e integrar a dos estudiantes uno con residencia profesional y otro con servicio social.

Metas:

Las metas establecidas fueron:

- Publicación de un artículo en revista arbitrada
- Incorporación de un residente
- Incorporación de un servicio social

Desarrollo y resultados del proyecto: En el proyecto se inició descargando el dataset AirtLab, repositorio público con 350 imágenes violentas y 350 no violentas, desarrollando en lenguaje Python utilizando tres redes preentrenadas VGG16, ResNet50, e Inception V3, una vez entrenadas la red se procedió a la clasificación y validación.

Conclusiones/Observaciones: Los resultados obtenidos en la investigación resultaron favorables, por lo que es la base para futuros trabajos utilizando otros modelos de redes y dataset.

2. Objetivo del proyecto:

El objetivo del proyecto investigación fue incrementar el rendimiento en la tarea de clasificación de imágenes provenientes de vídeo con dos clases violencia y no violencia, obtener la eficiencia de cada red y así establecer cual es la red que mejor clasifica este tipo de imágenes.

3. Metas: cumplimiento de Metas

PRODUCTOS ENTREGABLES			
Metas Cuantificables	Cantidad programada	Cantidad lograda	Observaciones.
Incorporación de estudiantes de licenciatura al proyecto, créditos complementarios:	1	1	
Incorporación de estudiantes de licenciatura al proyecto, servicio social:	1	1	
Artículos científicos en revistas arbitrada publicados:	1	1	

Incorporación de estudiantes de licenciatura al proyecto, servicio social:	
01-marzo-2023	Axel Rodrigo Lara Cruz (Ingeniería en Sistemas Computacionales)
Estudiantes residentes participantes en el proyecto	
15-enero-2023	Elizabeth Mejía Ibarra (Ingeniería en Sistemas Computacionales)

Metas cumplidas

Artículos científicos publicados o enviados en revistas arbitradas	
8-diciembre 2023	Empirical analysis on the effectiveness of pre-trained models in the identification of physical violence against women in videos for a multi-class approach

4. Metodología:

La metodología propuesta inicia con una entrada de un conjunto de imágenes por vídeo del dataset utilizar tres redes neuronales para extraer las características de las imágenes y clasificar cada imagen en violenta o no violenta. En la figura 1 se muestra la metodología propuesta.

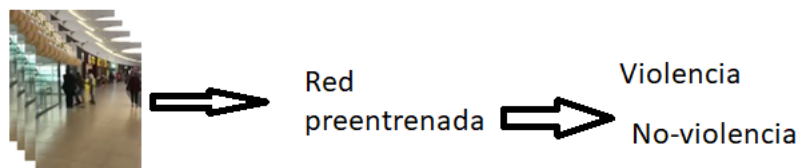


Figura 1 Metodología propuesta

Dataset utilizado

Para probar la eficiencia de los modelos se utilizó la base de datos publica, las cuales ya se en cuentan etiquetadas. AirtLab [1] es una colección de 350 vídeos etiquetados como violentos y 350 no violentos. En la Figura 2 muestra un ejemplo de imagen con violencia y no violencia de la base de datos.



Figura 2. Imágenes del dataset AIRTLab

Aprendizaje por transferencia

El entrenar una red neuronal convolucional desde cero tiene un alto costo computacional, por lo que se optó utilizar transferencia de aprendizaje, esta técnica se ajusta a como los seres humanos utilizamos el conocimiento en diferentes tareas. Utilizando el conocimiento previo en una nueva tarea. Existen diferentes modelos de redes preentrenadas como las utilizadas en esta investigación: VGG16, Resnet50, Inception. Cada modelo cuenta con diferente estructura lo que les permite utilizarlas en tareas con diferentes problemáticas. A continuación, se describe brevemente los modelos de utilizados.

Red Vgg16

Es una red preentrenada debe su nombre por estar compuesta por 16 capas de profundidad teniendo como tamaño de entrada 224 por 224, pudiendo clasificar 1000 categorías de objetos. Ver figura 3.

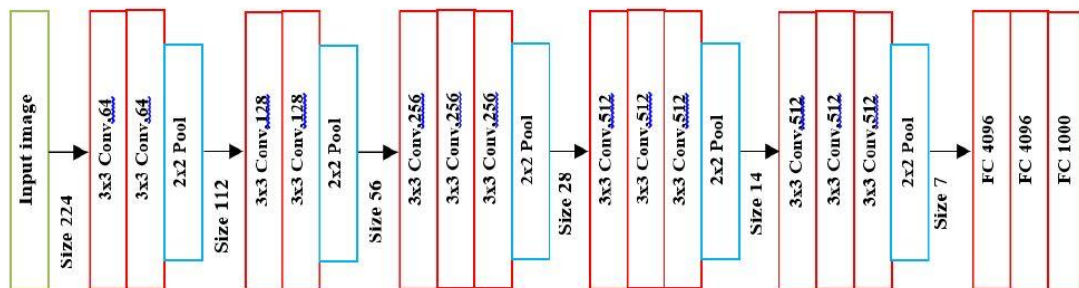


Figura 3. Arquitectura Vgg16 [2]

Red Inception -V3

Es una red preentrenada con 48 capas de profundidad, modelo compuesto por bloques de compilación simétricos y asimétricos los cuales incluyen convoluciones, reducción promedio, reducción máxima, concatenaciones y capas conectadas.

Red ResNet50

Como su nombre lo indica es una red convolucional de 50 capas. ResNet50 es un potente modelo de clasificación de imágenes que puede entrenarse con grandes conjuntos de datos y obtener los mejores resultados. Una de sus principales innovaciones es el uso de conexiones residuales, que permiten a la red aprender un conjunto de funciones residuales que asignan la entrada a la salida deseada. Estas conexiones residuales permiten a la red aprender arquitecturas profundas de lo que era posible hasta ahora, sin sufrir el problema de la desaparición de gradientes. En la figura 4 se muestra la arquitectura del modelo.

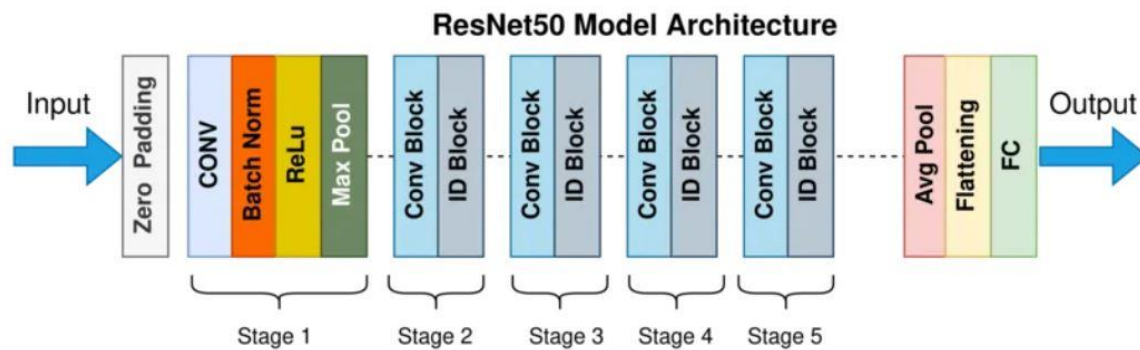


Figura 4. Arquitectura de Resnet50 [3]

Métrica de evaluación Accuracy

La métrica utilizada para medir la eficiencia de los modelos fue accuracy. La cual calcula el porcentaje de predicciones correctas realizadas por el modelo, ver ecuación 1.

$$accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

7. Información adicional

Referencias

[1] Bianculli, M.; Falcionelli, N.; Sernani, P.; Tomassini, S.; Contardo, P.; Lombardi, M.; Dragoni, A.F. A dataset for automatic violence detection in videos. Data in brief 2020, 33, 106587

[2] Ilyas, B. R., Beladgham, M., Merit, K., & taleb Ahmed, A. (2019). Illumination-robust face recognition based on deep convolutional neural networks architectures [J]. Indonesian Journal of Electrical Engineering and Computerence, 18(2), 1015-1027.

[3] 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).

A continuación, se anexa la publicación aceptada.

Subdirector(a) Académico(a)	Docente
América Gómez Sandoval	Itzel María Abundez Barrera
Nombre y Firma	Nombre y Firma

DMBD' 2023 <dmbd2023@easychair.org>

Sáb 11/11/2023 5:35

Para:Grisel Miranda Piña <MM22280266@toluca.tecnm.mx>

ADVERTENCIA, ESTE REMITENTE ES EXTERNO

ESTE MENSAJE DE CORREO ELECTRÓNICO SE ORIGINÓ FUERA DE LOS SERVICIOS INSTITUCIONALES DEL INSTITUTO TECNOLÓGICO DE TOLUCA. POR SU SEGURIDAD NO ABRA ARCHIVOS ADJUNTOS, NI HAGA CLIC EN ENLACES. NUNCA PROPORCIONE INFORMACION NI DATOS SENSIBLES, A MENOS QUE RECONOZCA AL REMITENTE Y TENGA LA CERTEZA DE QUE EL CONTENIDO ES SEGURO.

Dear Authors: Itzel Abundez, Grisel Miranda-Piña, Roberto Alejo, E. Everardo Granda-Gutiérrez, Alejandro Cisniega, Otniel Portillo-Rodríguez,

Congratulations!

On behalf of the DMBD'2023 program committee and technical committee, we are pleased to inform you that your paper

Paper ID: 6614:

Author(s): Itzel Abundez, Grisel Miranda-Piña, Roberto Alejo, E. Everardo Granda-Gutiérrez, Alejandro Cisniega, Otniel Portillo-Rodríguez

Title: Empirical analysis on the effectiveness of pre-trained models in the identification of physical violence against women in videos for a multi-class approach:

has been accepted for presentation at the Eighth International Conference on Data Mining and Big Data (DMBD'2023) which will be held on December 09-11, 2023, at Shanya, China, and for publication in the conference proceedings published by Springer's Book Series in Communications in Computer and Information Science (CCIS) (Electronic ISSN:1865-0937, Print ISSN:1865-0929), at <https://www.springer.com/series/7899>.

Empirical analysis on the effectiveness of pre-trained models in the identification of physical violence against women in videos for a multi-class approach ^{*}

I. Abundez¹[0000–0002–5095–342X], G. Miranda-Piña^{1**}[0000–0001–7122–0658], R. Alejo¹[0000–0002–7580–3305], E. E. Granda-Gutiérrez²[0000–0002–9316–9627], A. Cisniega¹[0009–0004–5055–2169], O. Portillo-Rodríguez³[0000–0002–2198–477X]

¹ Division of Postgraduate Studies and Research, National Institute of Technology of Mexico (TecNM) Campus Toluca, Av. Tecnológico s/n, Agrícola Bellavista, 52149, Metepec, México
{iabundezb, mm22280266, ralejoe, mm22281243}@toluca.tecnm.mx

² UAEM University Center at Atlacomulco, Universidad Autónoma del Estado de México, Carretera Toluca-Atlacomulco Km. 60, 50450, Atlacomulco, México
eegrandag@uaemex.mx

³ Faculty of Engineering, Universidad Autónoma del Estado de México, Cerro de Coatepec s/n, Ciudad Universitaria, 50100, Toluca, México
oportillor@uaemex.mx

Abstract. Violence against women captured in videos and surveillance systems necessitates effective identification to enable appropriate reactions for controlling and mitigating of its effects in public spaces and the potential apprehension of aggressors. While several algorithms have been developed for violence detection, their evaluation has primarily focused on controlled scenarios with clear differentiation between violent and non-violent scenes, representing two-class identification problems. However, real-world situations often present challenges where specific actions, such as hugs or effusive greetings, fall into an ambiguous class that is difficult to classify. Consequently, this transforms into a multi-class identification problem. In this study, we assess the performance of three pre-trained models, namely VGG16, ResNet50, and Inception V3, to evaluate their efficacy in addressing the multi-class identification challenges. Furthermore, we compare their performance against datasets consisting of two-class classifications, where the models generally exhibit satisfactory results. Our analysis reveals that the models struggle to differentiate the ambiguous scenes effectively, with Inception V3 achieving a 0% correct detection rate for this class. Notably, VGG16 outperforms the other models, attaining accurate detections of 48% for the ambiguous class, 75% for non-violence scenes, and 54% for violence scenes. This research sheds light on the limitations of current classification models when confronted with the complexity of real-world scenarios. The findings emphasize the importance of developing improved algorithms capable of accurately distinguishing ambiguous situations and enhancing the performance of violence detection systems.

Keywords: Artificial Neural Networks · Deep Learning · Physical-violence against women · Transfer Learning · VGG16 · ResNet50 · InceptionV3

^{*} TecNM has partially supported this investigation with project 13816.22-P.

^{**} Corresponding author

1 Introduction

In Mexico, The National Institute of Statistics Geography and Informatics has reported that 34.7% of women experience physical violence. The states with violence reports are the State of Mexico (78.7%), Mexico City (76.2%), and Queretaro (75.2%). In those states, women aged 15 years and older have experienced some gender violence in their lifetime. According to data from the 2021 National Survey on the Dynamics of Relationships in Households, violence against women in Mexico remains a significant concern. The survey findings indicate that women residing in urban areas experience a higher prevalence of violence, accounting for 73% of reported cases. Additionally, women between the ages of 25 and 34 face a particularly heightened risk, with a prevalence rate of 75%. The study also reveals that women with lower education levels are disproportionately affected, with 78% reporting incidents of violence. Furthermore, women who are separated, divorced, or widowed are significantly impacted, comprising 74% of reported cases [1]. In many cases, women victims of violence are part of a group already in precarious or vulnerable conditions, such as indigenous women and inhabitants of rural communities. Physical violence is one of the most severe problems in our society, and the efforts to mitigate it should be focused on the worrying consequences that can lead to death or permanent physical and emotional injuries [2].

Additionally to these facts, already severe, another complex factor that makes their containment even more complicated is the amount of economic, technological, and human resources required to provide care and response to the victims of violence. However, with the new technologies that appear every day, there are more tools to help the victims; such is the case that arises [3], who presents an architecture developed for the detection of criminal activities based on real-time analysis using a deep learning technique called Faster R-CNN, which has the function of detecting robberies and thefts. This application could return erroneous data if the offender's *modus-operandi* changes or the event unfolds differently than those used to train the network. Another example is shown in [4], where different deep learning models were studied to identify violence in images using pre-trained models exhibiting success rates of approximately 90 % using the public datasets *MoviesFight* and *HockeyFight*.

In the same context, in [5] the authors present a comparison of pre-trained models (such as VGG16, ResNet, InceptionV3 and MobileNet) in the detection of physical violence in video. Their experiments were conducted using a public dataset with recorded videos of students acting out the scenes with violent and non-violent behaviors. The results of these models showed that the MobileNet architecture gave the best result in terms of accuracy, while InceptionV3 came in second place.

Also, in [6], a neural network model for violence detection based on separating frames into three stages is proposed. This method is known as an end-to-end deep learning framework. An outstanding feature of this method is that frames that could slow down the network's learning process are removed in the first stage, resulting in much faster training.

One way to identify violence is through the posture taken by the aggressor. In [7], an algorithm capable of identifying fights by estimating the posture that someone adopts just before starting a fight was presented. Its outcome was achieved mainly by using a deep neural network called Open Pose, which detects silhouettes or outlines of people

in video and images. This network comprises multiple layers of convolutional neural networks that process an image to obtain the body posture of the people in that image.

A different approach to treating the problem is presented in [8] to detect a person's state of mind through convolutional neural networks, which could anticipate a violent action. It is an antecedent for women's differentiation in video images. This uses image separation using the Stochastic Gradient Descent method that seeks to minimize an objective function written as the sum of differentiated functions and solves the problem of high computational cost using the Principal Component Analysis (PCA) technique. This work resulted in a robust algorithm capable of identifying human emotions through facial expressions with a high level of effectiveness.

In [9] is proposed a neural network capable of distinguishing attacks with bladed weapons and firearms through the labeling of images and objects that involves thousands of photo frames of different types of weapons, ranging from the most common (such as pistols) to the most destructive (such as shotguns or assault rifles), to carry out later training that gives the algorithm the ability to identify firearms in real-time. On the other hand, in [10] is presented a SELayer-C3D violence detection model addressing the problem primarily in local regions, obtaining an accuracy of 99%.

Other problems to consider in identifying violence or similar tasks are false positives and error rates, which could be the difference between life and death. In [11], a neural network model that significantly reduces false positives and the error rate is proposed by changing the neural network approach. The challenge is to detect falls in older adults at their homes; because the camera does not see some falls in some cases, instead of focusing on detecting falls, the authors seek to train the algorithm to detect when the person is on the floor. In this way, higher percentages of effectiveness were achieved.

Much of the work aimed at detecting violence on video presents the problem that once violence has been identified, it is impossible to know who has attacked whom, that is, who is the victim and who is the perpetrator. So, in [12], the authors focused on developing a neural network that can provide additional data about the detected aggression. It is achieved by an Acceleration Measure Vector (AMV) composed of direction and magnitude of motion where the temporal derivative of AMV can detect abrupt movements. In this sense, if it is a kick or a punch, the algorithm detects who has initiated the attack.

Usually, videos used to train neural network models for detecting violent events use collections of pre-fabricated or acted scenes that could undoubtedly represent an inconvenience when facing actual acts of violence. Therefore, [13] proposed a network model based on ResNet50, which was trained with real videos from video surveillance cameras using different environments to detect violence and to classify it into three categories, the first of which involves violence with firearms and the remaining two only violent or inappropriate behaviors. The result is a truly more efficient model since its training was more precise and punctual and, therefore, could give better results if applied to real surveillance videos.

Despite the significant advances that have been made in detecting violence on video, there are still efforts aimed at treating the problem in a general way, and their study has

not focused on violence against women. For this reason, in this work, we focus on physical violence against women and address the problem of false positives. For that, we built a dataset with three classes that contain images (obtained from video) related to violence against women: a) violent scenes, b) non-violence, and c) scenes that could be difficult to classify or cause false positives. The effectiveness of three popular deep learning neural network models: VGG16, ResNet50, and Inception V3, are studied in this multi-class context.

2 Theoretical foundations

One of the most common Neural Network architectures is Feed-Forward, comprised of at least three neuron layers (one input, one output, and one hidden) [14]. In deep learning versions, Feed-Forward Neural Networks have two or more hidden layers, reducing the number of nodes per layer and using fewer parameters. Although it implies a more complex optimization problem [15], the availability of modern frameworks like TensorFlow or Apache-Spark (characterized by new technological approaches like Graphical Processing Units or Cluster Computing), makes this drawback less restrictive. Similarly, recurrent neural networks allow more efficient processing of long and variable-length data sequences [16].

Convolutional Neural Networks (CNN) is a trendy architecture that combines a fully connected layer (similar to multilayer perceptron or another linear classifier like support vector machines) with one or more convolutional layers to build a map that has the function of extracting key features (see Fig. 1). This architecture implements pooling to reduce the dimensionality of the features vector and save only the relevant features [17]. Consequently, abstract features can be extracted by the convolutional layers, where convolutional kernels behave as local filters for unprocessed sequential data and produce non-variant local features. Succeeding pooling layers extract the essential features within fixed-length windows [18]. CNNs are powerful extractors of meaningful features [19] and are effective in problems related to modeling image data, summarization, and classification [20]. VGG16, ResNet50, and Inception V3 are very popular pre-trained CNN models in classification tasks with promising results.

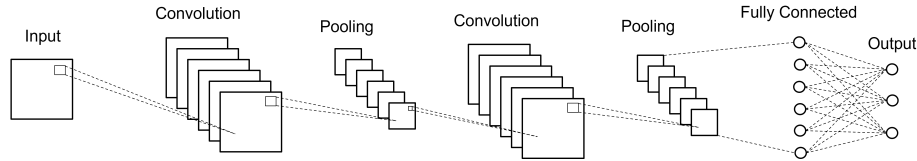


Fig. 1. Typical CNN is comprised of convolutional layers interleaved with pooling layers.

2.1 VGG16

It is the first deep neural architecture successor of the AlexNet model, also called VGGNet, which comprises multiple convolutions of fully connected layers, and many im-

age recognition parameters [21]. The main difference between the previous models and VGG16 is that its creators proposed using considerably small 3×3 receptive fields that can also be called filters and a stride of just one pixel. It is important to note that the receptive field or filter in the first layer in the AlexNet had an 11×11 with stride 4. It also involves fewer weight parameters and reduces the over-fitting in the training process.

At the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which is an annual competition where the new CNN models of image classification are presented, the VGG16 model with 16 layers achieved a test accuracy of 92.7%, leading the top 5 on ImageNet classification with an error of 7.3% [21]. Although the model did not win the event, its performance was so good that it caught the attention of researchers, scientists, and developers, enabling further innovations.

2.2 ResNet50

ResNet50 model is part of the ResNet family, whose architecture was introduced to avoid dead connections and reduce the effect of layers on performance. The number of deep layers depends on its version; then, ResNet50 has 50 deep layers. It is a residual CNN model that has produced ground-breaking performances in image recognition, object detection, face recognition, and image classification. The architecture has four stages that fix the degradation problem in deeper CNNs [22] and was designed with a convolution of 7×7 kernels, a max pooling with a stride of size 2, and the kernel increment in each layer. This configuration obtains very high accuracy and efficiency values. Therefore, this model can be used for artificial or non-artificial vision tasks. In addition, to perform a faster training process, it uses the "bottleneck" concept to reduce the number of parameters and the number of matrix operations.

2.3 Inception V3

Inception V3 is a CNN of the Inception family which aims to improve computational efficiency and reduce the number of parameters compared to its predecessor, using computation added in the most efficient way possible through factored convolutions and aggressive regularization [23]. The InceptionV3 version has 42 layers, with which it has demonstrated that it can achieve an accuracy of over 78.1%. This model consists of symmetric, non-symmetric, and convolution compilation blocks, as well as average reduction, maximum reduction, concatenations, and fully connected layers. Its most important feature is the consumption of fewer computer resources, which requires fewer optimization techniques, fewer parameters, and processes fewer convolutions, thus increasing speed and resources in training.

3 Experimental Set-Up

3.1 Dataset

The images used for this investigation are based on the videos in the public GitHub repository "airtlab"⁴. This repository contains 350 video clips labeled as non-violent

⁴ <https://github.com/airtlab/A-Dataset-for-Automatic-Violence-Detection-in-Videos>

and violent, which were simulated by non-professional actors, and are available for educational and research purposes [24].

Our modified dataset includes three classes of images related to physical violence against women: a) violence, b) no violence, and c) scenes where it is hard to distinguish between violence or non-violence (ambiguous), i.e., those situations where differentiating violence from a non-violence scene could be too difficult to an automatic classifier (for example, a strong hug or a very warm greeting, among others). Fig 2 shows an example of a scene with violence against the woman (left), a non-violence scene (center), and a scenario that could be difficult to classify or ambiguous (right). It is worth mentioning that for situations of violence we selected videos in which the male actor was the perpetrator and the female actress was the victim.



Fig. 2. Examples of images for the three classes involved in the classification of scenes: violence against a woman (left), a non-violence scene (center), and image hard to be classified as violence or non-violence or ambiguous(right).

Each frame of the selected videos was extracted using the Python programming language and the OpenCV library (open-source) [25]. Next, each video frame was manually labeled according to where violence was evident, where it was not, and where it was hard to identify (ambiguous). Finally, we created a dataset (D) of 2100 images with 700 violence (where the male actor was the perpetrator), 700 non-violence, and 700 ambiguous images. The images were normalized in size (224×224 pixels) to match the requested input shape of the pre-trained models selected as classifiers, and then the images were converted to grayscale.

To study the impact of images that are hard to be identified in the classifier performance, we set two datasets D_i , where D_1 contains only violence and non-violence images and D_2 violence, non-violence, and ambiguous images.

Dataset $D_i = \{x_1, x_2, \dots, x_Q\}$ was split randomly in three disjoint sets: TR_i to train the model, TV_i to validate it, and TE_i to test its generalized ability, $D_i = TR_i \cup TV_i \cup TE_i$; and $TR_i \cap TV_i \cap TE_i = \emptyset$, and x_q corresponds to q -th image in D_i . TR_i contains 50% of the samples of D_i , while TV_i and TE_i contains the remaining 50% (25% for each one).

3.2 Free parameters specification

The classification models selected for comparison are pre-trained on large data sets, so only fine tuning is required for their implementation in more specific tasks, such as ours in detecting physical violence against women. Considering this, to carry out the

fine tuning for each model we freeze the last 3 layers of each one, to later connect it to the layers of a new model. The model structure to connect with the output of the pre-trained model was found by testing different configurations and selecting the best one for each model based on the performance shown.

For the VGG16 and ResNet50 models, their outputs were connected to a flattened layer, two dense layers with 128 nodes, and the ReLu (rectified linear unit) activation function. The lost function used was Categorical Cross-entropy, which classifies the data by predicting the probability of whether the data belongs to one class or the other class. It is also worth mentioning that categorical cross-entropy is a measurement tool commonly used to evaluate the precision of probabilistic predictions. Adam optimizer was used to train the models due to its utility as a stochastic gradient descent method. The size of the batch was set to 25, and the number of epochs to 50.

Finally, for the InceptionV3 model, a flattened layer was also considered, but this model differs from the others because its dense layer has 4096 nodes. InceptionV3 uses a ReLu activation function, a 0.2 dropout layer, and a Softmax layer with n neurons. As in the previous models, n changes when the number of classes changes. In addition, the trial-and-error method found all the parameters selected for fine-tuning, where these values gave the best results for the classification tasks using the pre-trained weights of ImageNet.

We used the Linux operating system (Ubuntu distribution), where Anaconda software was installed as a development environment for the Python programming language. The TensorFlow framework was used with Keras as a programming interface to design, configure, and train the studied models. All tests were carried out on a computer that had a 4 GB NVIDIA GeForce GTX 1650TI graphics card with CUDA (Unified Computing Device Architecture) support.

3.3 Performance of the DL models

The confusion matrix is one way to visualize the performance of any classifier. It is generally used in binary classification problems [26]; nevertheless, it also could be applied to multi-class problems [27]. The confusion matrix is a cross table that records the number of occurrences between two raters, the true/actual classification, and the predicted classification, as shown in Table 1. In this paper, we use a confusion matrix with three columns for the predicted category and three rows for the actual occurrences (which correspond to three classes: violence, non-violence, and ambiguous) to test the performance of the studied models.

Table 1 presents three classes (A, B, and C), with 16, 22, and 18 samples in each one, of which only 5, 6, and 7 (respectively) are classified correctly, the rest are misclassified. In addition, we use the *accuracy* (Eq. 1) to know the global performance of each classifier, which is obtained directly from the confusion matrix as follows:

$$accuracy = \frac{\sum_i^N c_{i,i}}{\sum_i^N \sum_j^N c_{i,j}}, \quad (1)$$

where N is the total number of samples, i corresponds to the number of classes and $c_{i,j}$ corresponds to the value at position (i, j) of the confusion matrix.

Table 1. Example of confusion matrix for a multi-class problem

		Predicted class			
		Class A	Class B	Class C	Total
True class	Class A	5	7	4	16
	Class B	9	6	7	22
	Class C	8	5	5	18
	Total	22	18	16	56

4 Results and discussion

This section presents the experimental results of the pre-trained models used to detect physical violence against women in videos. VGG16, ResNet50, and InceptionV3 were tested on two different datasets: a two-class dataset, and a three-class dataset with ambiguous images. To evaluate the performance of such models, the confusion matrix was used as a metric to visualize how many images in the dataset were correctly classified based on their actual labels. The number of correctly classified images is in bold in all matrices presented.

4.1 VGG16 model

Regarding this model, Table 2 shows a high level of classification for the non-violence class, and the number of violence images correctly classified that the VGG16 model performed well in the binary classification test (two-classes dataset).

Table 2. Confusion matrix for binary classification with VGG16 model.

		Predicted class	
		Non-violence	Violence
True class	Non-violence	174	1
	Violence	40	135

In contrast, Table 3 shows that in the multi-class classification task, this classifier still distinguishes the non-violence class best, followed by the violence class, while

it shows the worst classification performance for the ambiguous class. However, the model could not perform well for all the classes involved. This could be explained due to the complexity of the third class, where the movements in the images used can be mistaken as violence against women when it could be a greeting, a hug, or something related.

Table 3. Confusion matrix for multi-class classification with VGG16 model.

		Predicted class		
		Ambiguous	Non-violence	Violence
True class	Ambiguous	84	79	12
	Non violence	36	131	8
	Violence	60	20	95

4.2 ResNet50 model

To evaluate the performance of this model, Table 4 presents the number of correctly classified and misclassified images for the binary classification task, where the violence class showed the best classification, with 82% of the 175 violence images used for the test data. In contrast, only 24% of the non-violence images were correctly classified. This shows that the ResNet50 model could not learn how to differentiate violence from non-violence scenes.

Table 4. Confusion matrix for binary classification with ResNet50 model.

		Predicted class	
		Non-violence	Violence
True class	Non-violence	42	133
	Violence	31	144

In the multi-class scenario test for ResNet50, Table 5 shows a better classification performance for the non-violence class, and most images are classified in this class. Thus, ResNet50 exhibited a low overall performance in a multi-class task, obtaining the lowest number of images classified for the violence class, which is the area of interest.

Moreover, as can be observed, both VGG16 and ResNet obtained the same number of correctly classified non-violence images.

Table 5. Confusion matrix for multi-class classification with ResNet50 model.

		Predicted class		
		Ambiguous	Non-violence	Violence
True class	Ambiguous	33	135	7
	Non-violence	30	131	14
	Violence	49	107	19

4.3 InceptionV3 model

Finally, the last pre-trained model used to detect physical violence against women in video images showed a low performance for the class of interest in the two-class test. It showed similar behavior to the other two models for the binary classification, tending to classify better the non-violence images, as can be observed in Table 6, where only 91 images (about 50%) of the violence class were well classified. In contrast, 93% of the non-violence images were correctly classified.

Table 6. Confusion matrix for binary classification with InceptionV3 model.

		Predicted class	
		Non-violence	Violence
True class	Non-violence	163	12
	Violence	84	91

Table 7 evidences a low performance of the InceptionV3 model for the multi-class task. The network could not classify images of the ambiguous class in this scenario; instead, most of the images were classified as non-violent. This could be explained by the fact that most of the images were extracted from non-violent video clips, where the behavior shown in the scene appears to be confusing in determining whether raising the hand will be a punch or simply a greeting. On the other hand, InceptionV3 only classified 39 images (about 20%) for the interest class, while the non-violence class performed well with 171 images correctly classified.

Table 7. Confusion matrix for multi-class classification with InceptionV3 model.

		Predicted class		
		Ambiguous	Non-violence	Violence
True class	Ambiguous	0	167	8
	Non-violence	0	171	4
	Violence	0	136	39

Furthermore, to support the results of the confusion matrices, accuracy values were calculated for each test. These values are shown in Table 8, where it is observed how VGG16 continues to rank as the best classification algorithm in both approaches, with 88% accuracy 59% accuracy in the multiclass approach. However, these values indicate that the algorithms' learning and the classification level are poor. The latter could be associated with the number of images used for this task and the complexity of determining when an act of violence against women occurs.

Table 8. Accuracy values of the selected pre-trained models. Data in bold are the best precision values for each scenario.

	VGG16	ResNet50	InceptionV3
Binary	0.8829	0.5314	0.7257
Multiclass	0.5905	0.3486	0.4000

Based on the results obtained with the selected hyperparameter configuration, VGG16 emerged as the most effective model in both: binary and multi-class scenarios, consistently demonstrating relatively strong performance, as evident from the previously presented confusion matrices. InceptionV3 followed as the second-best model, exhibiting satisfactory results in binary classification but struggling significantly in the multi-class setting. On the other hand, ResNet50 initially performed poorly in binary classification but surpassed InceptionV3 in multi-class classification. Although InceptionV3 achieved higher accuracy values than ResNet50, it failed to classify any images from the ambiguous class correctly.

Notably, the accuracy values were obtained considering only two categories, underscoring the model's limited performance in the multi-class approach. Consequently, the results highlight the challenges faced by the classifiers in learning classes other than non-violence, suggesting that non-violence class was learned more effectively by all models tested. These findings underscore the need for further research and model de-

velopment to improve the classifiers' ability to classify complex classes, particularly in multi-class scenarios, accurately. By addressing these limitations, enhanced models can contribute to more robust violence detection systems that effectively identify and distinguish between various classes, ultimately fostering safer environments for women.

5 Conclusions and future research

This study examined the performance of popular pre-trained models for binary image classification in multi-class scenarios, specifically in identifying violence against women. The presence of an ambiguous class, which is challenging to distinguish from the typical binary classes of violence or non-violence scenes, complicates the accurate classification of violence in real video captures.

The findings highlight the limitations of models that perform well in controlled scenarios and two-class classification problems but struggle to respond effectively when faced with the ambiguity inherent in real-life situations, such as frames depicting hugs or effusive greetings that may be misclassified as violence.

Analyzing the confusion matrices, it was observed that the ResNet50 model achieved an 82% correct identification rate for violence scenes (the most significant class) in a two-class dataset. However, when the dataset included ambiguous scenes as a third class, this model could only identify 10% of the violence images and 18% of the ambiguous class.

Similarly, the InceptionV3 model demonstrated a 52% correct detection rate for violence images in the two-class dataset but surprisingly failed to identify any scenes from the ambiguous class (0%) in the three-class approach. Furthermore, it was considered that since it is a deeper architecture, the gradients may not flow efficiently to the shallower layers in the established epochs, and therefore their learning capacity could be affected.

On the other hand, the VGG16 model exhibited the best performance in identifying ambiguous images. It correctly recognized 77% of violence images in the two-class problem, and it achieved 54% correct identification for violence scenes, 75% for non-violence scenes, and 48% for the ambiguous class in the three-class context. These results can be attributed to the architecture of this model, because by having more parameters the network has a greater capacity to easily fit the new data, therefore it learns the features better.

Future research should focus on developing or fine-tuning models capable of accurately predicting classes that can potentially confuse automatic classifiers. This is essential to avoid false positives and prevent misinterpretation of potentially dangerous scenes. Additionally, exploring more realistic multi-class problems instead of binary classification would facilitate a more comprehensive evaluation of proposed models. These efforts will contribute to advancing violence detection systems and improve their effectiveness in real-world scenarios.

References

1. INEGI. Violencia contra las mujeres en México. <https://www.inegi.org.mx/tablerosestadisticos/vcmm/>, 2022.

2. David. Moctezuma-Navarro, José. Narro-Robles, and Lourdes. Orozco-Hernandez. La mujer en México: inequidad, pobreza y violencia. *Rev. mex. cienc. polít. soc [online]*, pages 117–146, 2014.
3. Julio Ernesto Suárez Páez. *Arquitectura de detección de actividades criminales basada en análisis de vídeo en tiempo real*. PhD thesis, Universitat Politècnica de València, 2020.
4. Ernesto Luis Bisbé. Detección de escenas de violencia con modelos deep learning. B.S. thesis, Universidad Autónoma de Madrid, 2020.
5. E De S Victor, Tiago B Lacerda, Pérciles BC Miranda, André CA Nascimento, and Ana Paula C Furtado. Federated learning for physical violence detection in videos. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.
6. Fath U Min Ullah, Amin Ullah, Khan Muhammad, Ijaz Ul Haq, and Sung Wook Baik. Violence detection using spatiotemporal features with 3d convolutional neural network. *Sensors*, 19(11):2472, 2019.
7. Jairo Egberto Powell González et al. Detección de peleas en videos usando estimación de postura y bi-lstm. Master’s thesis, Benemérita Universidad Autónoma de Puebla, 2021.
8. Patricia Sereno Rodriguez. Reconocimiento de expresiones faciales mediante el uso de redes neuronales convolucionales. B.S. thesis, Universitat Politècnica de Catalunya, 2017.
9. Jacqueline Sánchez and Miguel A Campos. Red neuronal artificial para detección de armas de fuego y armas blancas en video vigilancia. *Revista de Iniciación Científica*, 7(2):83–88, 2021.
10. Bin Jiang, Fangqiang Xu, Wenxuan Tu, and Chao Yang. Channel-wise attention in 3d convolutional networks for violence detection. In *2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA)*, pages 59–64. IEEE, 2019.
11. Amal El Kaid, Karim Ba’ina, and Jamal Ba’ina. Reduce false positive alerts for elderly person fall video-detection algorithm by convolutional neural network model. *Procedia computer science*, 148:2–11, 2019.
12. Ankur Datta, Mubarak Shah, and N Da Vitoria Lobo. Person-on-person violence detection in video data. In *2002 International Conference on Pattern Recognition*, volume 1, pages 433–438. IEEE, 2002.
13. Soheil Vosta and Kin-Choong Yow. A cnn-rnn combined structure for real-world violence detection in surveillance cameras. *Applied Sciences*, 12(3):1021, 2022.
14. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
15. I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016.
16. Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, 31(7):1235–1270, 07 2019.
17. B. Bengfort, R. Bilbro, and T. Ojeda. *Applied Text Analysis with Python*. O’Reilly Media, Inc., jun 2018.
18. Rui Zhao, Ruqiang Yan, Jinjiang Wang, and Kezhi Mao. Learning to monitor machine health with convolutional bi-directional lstm networks. *Sensors*, 17(2), 2017.
19. Jheng-Long Wu, Yuanye He, Liang-Chih Yu, and K. Robert Lai. Identifying emotion labels from psychiatric social texts using a bi-directional lstm-cnn model. *IEEE Access*, 8:66638–66646, 2020.
20. Teja Kattenborn, Jens Leitloff, Felix Schiefer, and Stefan Hinz. Review on convolutional neural networks (cnn) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173:24–49, 2021.
21. H. Yang, J. Ni, J. Gao, Z. Han, and T. Luan. A novel method for peanut variety identification and classification by improved vgg16. *Sci Rep*, 11:15756, 2021.

22. Dhananjay Theckedath and RR Sedamkar. Detecting affect states using vgg16, resnet50 and se-resnet50 networks. *SN Computer Science*, 1(2):1–7, 2020.
23. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.
24. Miriana Bianculli, Nicola Falcionelli, Paolo Sernani, Selene Tomassini, Paolo Contardo, Mara Lombardi, and Aldo Franco Dragoni. A dataset for automatic violence detection in videos. *Data in Brief*, 33:106587, 2020.
25. G. Bradski. The opencv library. *Dr. Dobbs’ Journal: Software Tools for the Professional Programmer*, 25:120–123, 2000.
26. A. Luque, A. Carrasco, A. Martin, and A. De-Las-Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231, 2019.
27. Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview, 2020.