Universidad Nacional Federico Villarreal

VRIN | VICERRECTORADO DE INVESTIGACIÓN

ESCUELA UNIVERSITARIA DE POSGRADO

SISTEMA WEB BASADO EN REDES NEURONALES CONVOLUCIONALES PARA RECONOCER LA VIOLENCIA FÍSICA EN ZONAS URBANAS

Línea de investigación:
Sistemas inteligentes, robótica, domótica

Tesis para optar el grado académico de Doctor en Ingeniería de Sistemas

Autor

García Díaz, José Edgar

Asesor

Rodríguez Rodríguez, Ciro

ORCID: 0000-0003-2112-1349

Jurado

Flores Masías, Edward José

Flores Vidal, Higinio Exequiel

Soto Soto, Luis

Lima - Perú

2025

SISTEMA WEB BASADO EN REDES NEURONALES CONVOLUCIONALES PARA RECONOCER LA VIOLENCIA FÍSICA EN ZONAS URBANAS

INFORME DE ORIGINALIDAD

21%	18%	7%	11%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	hdl.handle.net Fuente de Internet	2%
2	repositorio.ucv.edu.pe Fuente de Internet	2%
3	www.scielo.org.mx Fuente de Internet	1%
4	sedici.unlp.edu.ar Fuente de Internet	1%
5	www.coursehero.com Fuente de Internet	1%
6	conceptodefinicion.de Fuente de Internet	1%
7	repositorio.ucsg.edu.ec Fuente de Internet	1%
8	Submitted to Universidad Nacional Federico Villarreal Trabajo del estudiante	<1%

Universidad Nacional Federico Villarreal

VRIN | VICERRECTORADO DE INVESTIGACIÓN

ESCUELA UNIVERSITARIA DE POSGRADO

SISTEMA WEB BASADO EN REDES NEURONALES CONVOLUCIONALES PARA RECONOCER LA VIOLENCIA FÍSICA EN ZONAS URBANAS

Línea de investigación:

Sistemas inteligentes, robótica, domótica

Tesis para optar el grado académico de Doctor en Ingeniería de Sistemas

Autor

García Díaz, José Edgar

Asesor

Rodríguez Rodríguez, Ciro

ORCID: 0000-0003-2112-1349

Jurado

Flores Masías, Edward José

Flores Vidal, Higinio Exequiel

Soto Soto, Luis

Lima – Perú

2025

Dedicatoria

A mi familia, que junto a mi esposa Zarela y mis tres hermosos hijos, Sofia, Almendra y Rodrigo, me brindaron su tiempo y apoyo incondicional, en esta etapa de mi vida académica, haciendo grandes esfuerzos para comprender mi deseo de superación para ellos, espero pueda recompensarlos con este trabajo.

De manera especial a mis queridos padres, Edgar y Fátima, y a mi hermana Jessica, por brindarme su amor familiar hoy, mañana y siempre.

Agradecimientos

Agradezco a nuestro padre celestial, Dios todopoderoso, por darme la fuerza y concederme la gracia de llegar hasta aquí.

A todos mis familiares y amigos que de alguna manera me brindaron su apoyo y ánimo para continuar.

A la plana docente del doctorado de Ingeniería de Sistemas de la Universidad Nacional Federico Villareal por sus grandes enseñanzas académicas, que estoy seguro nunca olvidare.

De manera especial a mi asesor, el Dr. Ciro Rodríguez Rodríguez, por haber compartido sus conocimientos, por haberme guiado y apoyado durante el desarrollo de la tesis.

Muchas gracias.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Resumen

Esta investigación tuvo como objetivo desarrollar un Sistema Web basado en Redes Neuronales Convolucionales (CNN) para el reconocimiento de la violencia física en zonas urbanas, fue del tipo aplicada, nivel descriptivo – predictivo y diseño cuasi experimental; dentro de los objetivos específicos se desarrolló las siguientes actividades: primero, se realizó pruebas de rendimiento a los algoritmos CNN pre entrenados (VGG16, MobileNetV2 y YoloV8); para evaluar sus resultados y elegir al mejor de ellos; segundo, se sometió al sistema web a pruebas para verificar su eficiencia en el reconocimiento de acciones de violencia (patada, trompada, forcejeo y estrangulación); y tercero, se midió el tiempo de respuesta del sistema web al detectar una acción violenta, comparándolo con un método tradicional de videovigilancia. Los resultados de estas actividades mostraron que, YoloV8 fue elegido con un accuracy del 89%, y por tener una diferencia significativa con respecto a los otros algoritmos *(p – valor = 0.001)*, asimismo, se logró conseguir un resultado de "buena concordancia" entre el funcionamiento del sistema web y el método tradicional de videovigilancia *(kappa = 0.667)*; finalmente, se obtuvo un promedio del tiempo de respuesta del sistema web de *0.19 min* con respecto al método tradicional de *2.94 min*, con una diferencia significativa entre ambos *(p – valor = 1,846E-14)*. En consecuencia, se concluye que el Sistema Web basado en CNN optimiza el reconocimiento de la violencia física en zonas urbanas; esta afirmación se sustenta con las pruebas de comparación, evaluación y validación realizadas, cumpliendo así con los objetivos e hipótesis de estudio.

Palabras claves: Deep learning, clasificación, detección de objetos, transfer learning, violencia física.

ABSTRACT

The objective of this research was to develop a Web System based on Convolutional Neural Networks (CNN) for the recognition of physical violence in urban areas. It was applied research, descriptive-predictive in nature, and had a quasi-experimental design. The specific objectives included the following activities: first, performance tests were conducted on the pre-trained CNN algorithms (VGG16, MobileNetV2, and YoloV8), to evaluate their results and choose the best one; second, the web system was tested to verify its efficiency in recognizing acts of violence (kicking, punching, struggling, and strangling). Third, the web system's response time in detecting a violent act was measured and compared to a traditional video surveillance method. The results of these activities showed that YoloV8 was chosen with an accuracy of 89%, and because it showed a significant difference compared to the other algorithms *(p-value = 0.001)*. Furthermore, a "good agreement" result was achieved between the web system's performance and the traditional video surveillance method *(kappa = 0.667)*. Finally, the average response time of the web system was 0.19 minutes, compared to 2.94 minutes for the traditional method, with a significant difference between the two *(p-value = 1.846E-14)*. Consequently, it is concluded that the CNN-based web system optimizes the recognition of physical violence in urban areas. This assertion is supported by the comparison, evaluation, and validation tests performed, thus fulfilling the study's objectives and hypotheses.

Keywords: Deep learning, classification, object detection, transfer learning, physical violence.

I. INTRODUCCIÓN

1.1. Planteamiento del problema

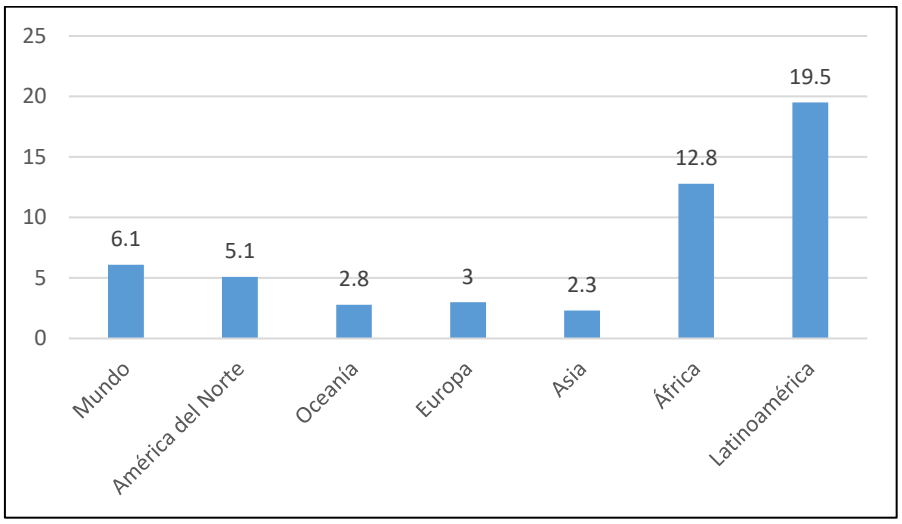
En consideración a la Organización Panamericana de la Salud (OPS) en sociedad con la Organización Mundial de la Salud (OMS), en su artículo sobre "Prevención de la Violencia", definen a la violencia como el "uso intencional de fuerza o poder físico real o como amenaza contra uno mismo, una persona, grupo o comunidad que tiene como resultado la probabilidad de daño psicológico, lesiones, la muerte, privación o mal desarrollo". En la página web de la OPS/OMS se da a conocer que a nivel mundial al menos 470.000 personas al año han fallecido por homicidio lo cual es muy alarmante para la zona de las Américas ya que presenta una de las tasas más altas de homicidios en el mundo, tres veces más del promedio global, la cual indica que casi 500 personas mueren al día como resultado de la violencia interpersonal, donde se tiene que al menos 1 de cada 3 mujeres han sufrido violencia ya sea físicamente o sexualmente de parte de su pareja (OPS, 2021). Diversos estudios relacionados a la violencia (Programa de las Naciones Unidas para el Desarrollo [PNUD], 2020; Rettberg, 2020; Centro de Investigación en Política Pública, 2019) también indican que en América Latina se tiene altos niveles de violencia desde ya hace varios años, la cual es considerada como una de las regiones más violentas, y aunque las políticas gubernamentales en muchos casos hacen sus esfuerzos y estrategias, durante el periodo de la Covid-19, esto ha sido un factor preponderante que al parecer incrementó la violencia en sus múltiples manifestaciones (urbana, secuestro, justicia por la propia mano, entre otros), que alcanzó una tasa promedio anual de al menos 17.2 homicidios por cada 100.000 habitantes. En la Figura 1 se observa un resumen de las tasas de homicidios por región, donde se confirma a Latinoamérica como una de las regiones con más violencias en el mundo. Cabe destacar que dentro de los objetivos de desarrollo sostenible al 2030 se consideran las metas para reducir las distintas formas de violencia, así como sus correspondientes tasas de mortalidad, a través del fortaleciendo de las instituciones

involucradas a todo nivel, teniendo en cuenta a los países en vías de desarrollo, para que obtengan la capacidad necesaria en todos sus aspectos.

Es importante señalar que los índices de violencia en América Latina y el Caribe han generado una sensación de inseguridad entre su misma población, la cual es un problema que está relacionado con la inestabilidad política en los gobiernos de turno, así como el bajo desarrollo económico, lo que hace difícil tener un análisis claro y poder combatir sobre los múltiples factores asociados (Latinoamérica21, 2021).

Figura 1

Tasa de víctimas de homicidios por región en el año 2017



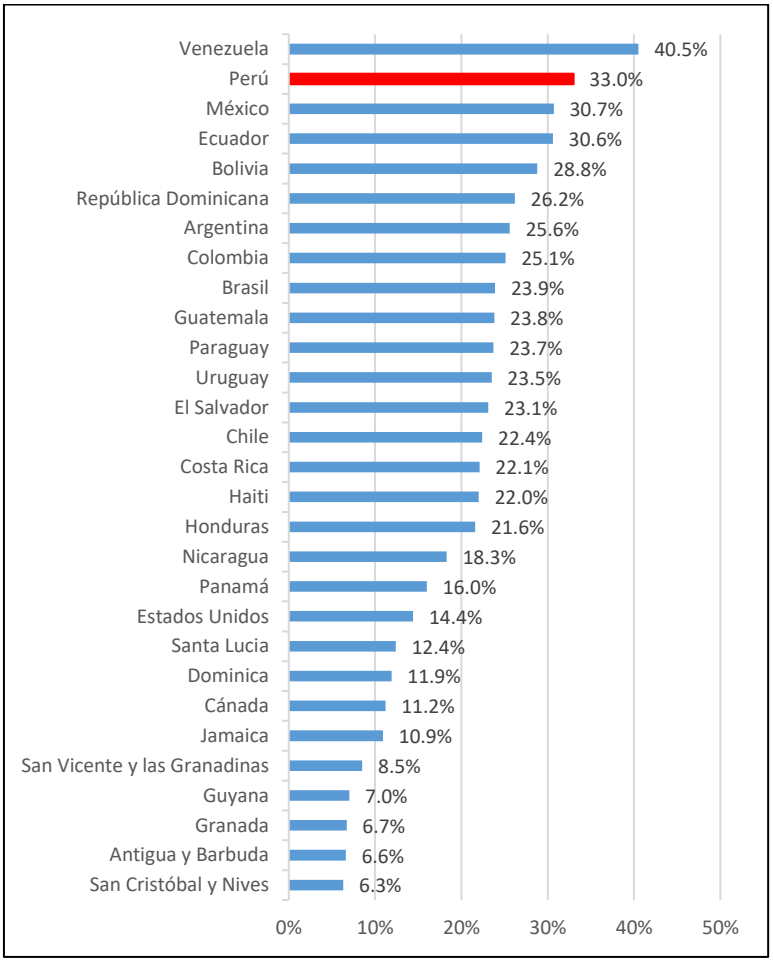
Nota: Resumen de las víctimas de homicidios a nivel mundial, siendo los datos de África al año 2015. Tomado de "Violencia en América Latina hoy: manifestaciones e impacto" por Rettberg, 2020, Revista de Estudios Sociales, 73.

Asimismo, en el Perú a través de la Ley No. 30364 (El Peruano, 2015), se busca tomar medidas correctivas para la violencia contra la mujer e integrantes de la familia, donde clasifica a la violencia en cuatro tipos: física, psicológica, sexual y económica o patrimonial. En tanto, el Instituto Nacional de Estadística e Informática (INEI, 2018), en su anuario sobre criminalidad y seguridad ciudadana, indica que los casos más conocidos de violencia son: la

violencia psicológica y la física, y que en el 2017, el 65,4% de mujeres entre 15 a 49 años, fueron violentadas alguna vez por parte de su pareja. También, el INEI indica que las mujeres en su mayoría fueron víctimas de violencia psicológica (61,5%), así como también de violencia física (30,6%) seguido de la violencia sexual (6,5%). Ahora bien, en el portal web estadístico del Ministerio de la Mujer y Poblaciones Vulnerables (MIMP, 2021), indica que, en el Perú, se registraron más de 2,400 casos de violencia familiar, sexual, así como otros de alto riesgo. Dentro de este marco, el Barómetro de las Américas (2017), que corresponde al Proyecto de Opinión Pública de América Latina (LAPOP) ubica al Perú como uno de los primeros países del ranking con las tasas más altas de víctimas por delincuencia, superando solamente a Venezuela (Instituto de Estudios Peruanos, 2018), tal como se observa en la Figura 2.

Figura 2

Víctimas por delincuencia en el Perú en relación con los países de América



Nota: Resumen de las víctimas por delincuencia en América, siendo los datos entre los años 2016 y 2017. Tomado de "Victimización por delincuencia en las Américas, 2016/17" de Barómetro de las Américas por LAPOP, 2017, Cultura política de la democracia en Perú y en las Américas, 2016/17.

Esta investigación se realizó en la ciudad de Iquitos, que cuenta con aproximadamente 600 mil habitantes, y que se ubica en el departamento de Loreto, considerado como el más grande del Perú, con un área de más de 368 mil km2 (Odicio, 1992), sin embargo, es uno de los departamentos que por años ocupa una de las posiciones más bajas del ranking en sectores como salud y educación, con puntajes bajos a nivel nacional según el Índice de Competitividad Regional (INCORE, 2023), lo cual se refleja en sus estadísticas de victimización por hechos delictivos; y que a pesar de los supuestos esfuerzos de los distintos gobiernos regionales y municipales de turno, al tratar de establecer políticas y campañas para implementar sistemas de vigilancia y seguridad ciudadana, hasta la fecha no hay un control satisfactorio, por el contrario durante los últimos años en la ciudad de Iquitos existe una sensación del incremento de la violencia urbana y del sicariato causando aún más inseguridad, esto obligó, hace ocho años, a la autoridad municipal a instalar más de 60 cámaras de videovigilancia en puntos críticos de la ciudad, pero actualmente muchas de ellas ya se encuentran inoperativas.

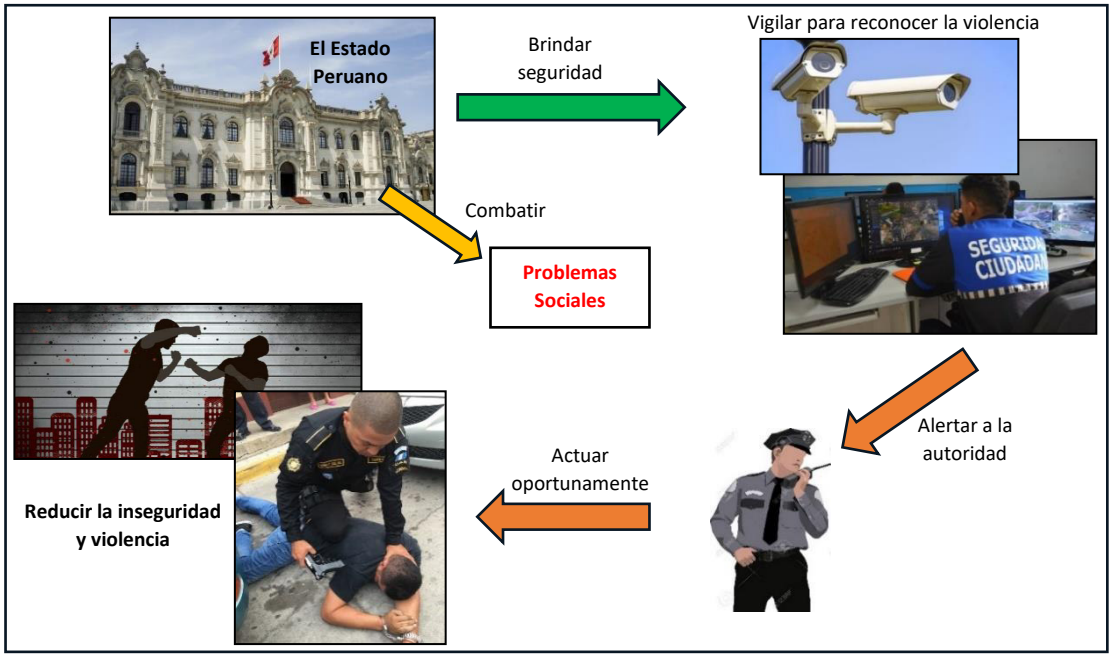
1.2. Descripción del problema

Entonces, uno de los principales problemas sociales a combatir es la inseguridad ciudadana, que da como resultado la violencia física con altas tasas de víctimas inocentes; en el Perú las más afectadas son las mujeres, que muchas veces no pueden denunciar por falta de pruebas, y aunque el gobierno disponga de algunos programas de apoyo social (MIMP, 2020) creando puestos de control de videovigilancia municipal, implantar seguridad vecinal con

patrullaje integrado junto a la policía nacional entre otros para contrarrestar esta problemática; estas estrategias resultan limitadas debido a la necesidad y/o dependencia del factor humano para identificar, reconocer y alertar las acciones de violencia de manera oportuna, es decir, debe existir el personal capacitado para el monitoreo constante en las calles y en múltiples pantallas de videovigilancia para reconocer acciones violentas en la ciudad y así actuar oportunamente, con un proceso similar a la Figura 3 que muchas veces es demorado o simplemente no es eficaz.

Figura 3

Esquema actual para combatir el problema social de la inseguridad y la violencia en el Perú



Nota: Flujo de un sistema tradicional de videovigilancia para poder identificar un acto de violencia y dar a conocer a la autoridad competente para que tome acciones oportunamente.

A la actualidad, y de acuerdo al proceso visto en la figura anterior, básicamente se tienen los siguientes problemas a resolver: Primero, determinar si existe o no existe la acción de violencia física y segundo, mejorar el tiempo de respuesta para alertar la acción de violencia física; cuyos indicadores y valores de la situación actual se muestran en la Tabla 1:

Tabla 1

Indicadores y valores de la situación actual

Indicadores	Valores
Existencia de la acción de violencia física.	Existe, no existe (se reconoce con dificultad)
Tiempo de respuesta para alertar la violencia física.	De 2 a 5 minutos (el tiempo es demorado)

Entonces, se plantea el desarrollo de un sistema web basado en la técnica del aprendizaje supervisado con redes neuronales convolucionales, de manera que se pueda detectar acciones de violencia a partir de imágenes de cámaras de videovigilancia, reduciendo así errores al tratar de reconocer y determinar las acciones violentas a través de la mejora de la precisión en la predicción, a la vez que permita reducir los tiempos de respuestas para alertar sobre dichas acciones una vez reconocidas y sin la dependencia del factor humano.

Asimismo, en la Tabla 2 se compara la Situación Actual (AS – IS) en consideración a la Situación Propuesta (TO – BE) luego de la implementación de la solución.

Tabla 2

Situación Actual (AS – IS) frente a la Situación Propuesta (TO – BE)

Situación Actual (AS – IS)	Situación Propuesta (TO – BE)
Existen ocasiones en que no se puede determinar si existe o no existe la violencia física, debido a errores involuntarios del factor humano o método tradicional.	Determinar de manera eficiente la existencia de la violencia física.
Los tiempos de respuestas para alertar sobre la violencia detectada son demorados, debido a las dudas en las tomas de decisiones del factor humano.	Reducir el tiempo de respuesta para alertar la violencia física detectada.

1.3. Formulación del problema

1.3.1. Problema general

¿En qué medida el desarrollo de un Sistema Web basado en Redes Neuronales Convolucionales optimiza el reconocimiento de la violencia física en zonas urbanas?

1.3.2. Problemas específicos

PE1: ¿Cómo comparar las métricas de evaluación de los modelos pre entrenados de Redes Neuronales Convolucionales para seleccionar el modelo con el mejor rendimiento en el reconocimiento de la violencia física?

PE2: ¿De qué manera evaluar el rendimiento del Sistema Web basado en Redes Neuronales Convolucionales para determinar la existencia de la violencia física dentro de una zona urbana?

PE3: ¿Cómo validar el rendimiento del Sistema Web basado en Redes Neuronales Convolucionales con la finalidad de medir el tiempo de respuesta para alertar la violencia física?

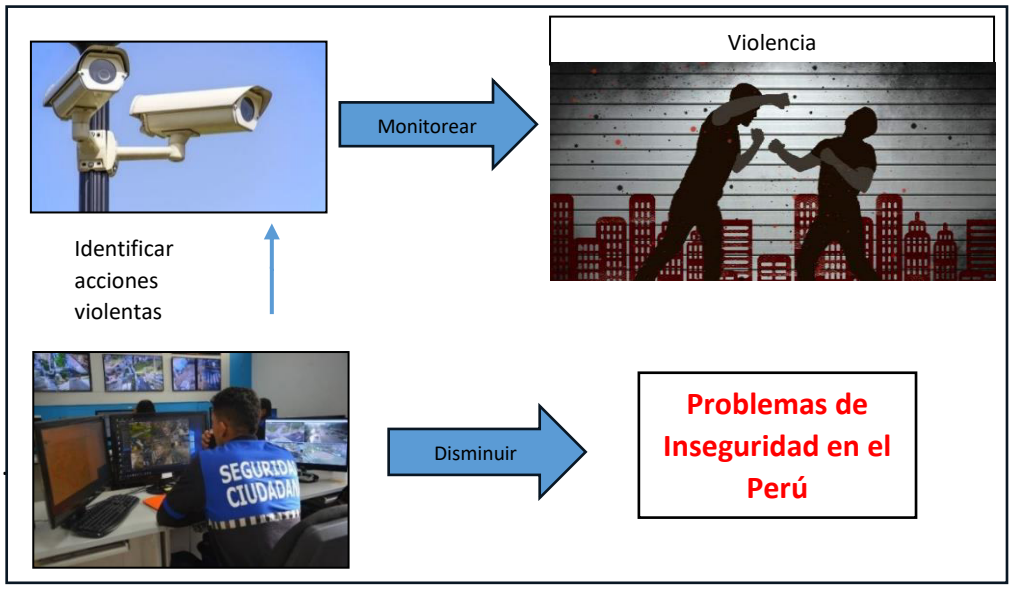
1.4. Antecedentes

Según, Lumba et al. (2019) en su investigación "Solución Computacional para el Reconocimiento de Acciones Básicas de Violencia en Tiempo Real, a partir del uso de Redes Neuronales Convolucionales, Secuencias de Video y Computación de Alto Rendimiento" manifiesta que actualmente la inseguridad y la violencia son grandes problemas sociales en el Perú y que según el INEI en el 2017, indica que el 65,4% de las mujeres de 15 a 49 años, fueron víctimas de violencia por parte de su pareja íntima, donde, mayormente fueron víctimas de violencia psicológica (61,5%), violencia física (30,6%) seguido de la violencia sexual (6,5%). Asimismo, indica que el Barómetro de las Américas 2017, Proyecto de Opinión Pública de

América Latina (LAPOP) ubica al Perú como uno de los primeros países del ranking con una de las tasas más altas de víctimas por delincuencia, superando solamente a Venezuela. También indica que en el informe técnico del INEI sobre Seguridad Ciudadana, 26 de cada 100 peruanos con más de 15 años, fue víctima de algún acto violento entre marzo y agosto del 2018, dicho informe también indica que a nivel nacional dentro del sector urbano, 13 de cada 100 personas han sufrido el robo de su dinero, billetera o celular y a nivel urbano, que en lugares con más de 20,000 habitantes, 14 de cada 100 personas sufrieron los mismos actos delictivos; que en el sector de Lima Metropolitana, 15 de cada 100 personas de igual manera han sufrido lo mismo y que en los centros poblados de entre 2,000 a 20,000 habitantes se han visto afectados 9 de cada 100. En líneas generales, afirma que el Perú afronta cifras muy alarmantes de violencia, sobre todo en contra de las mujeres. El investigador, también menciona a la ciudad de Iquitos, y comenta que en los últimos 10 años se incrementó considerablemente los casos de violencia urbana, donde la autoridad edil pensando en mejorar la seguridad instalaron algo más de 60 cámaras de videovigilancia en puntos estratégicos de los distritos de Belén, Punchana, San Juan y en el mismo Iquitos, con un esquema de control similar a la Figura 4; finalmente, el autor indica que un alto porcentaje de acciones violentas fueron registradas gracias a estas cámaras, pero sin embargo, el personal técnico, encargado del monitoreo no se abastece, existiendo casos de violencia no identificadas, por lo que recomienda pensar en nuevos mecanismos de detección automática de violencia a través del uso de la Inteligencia Artificial y la Computación de Alto Rendimiento.

Figura 4

El rol de la videovigilancia en el control de la seguridad ciudadana y la violencia física.

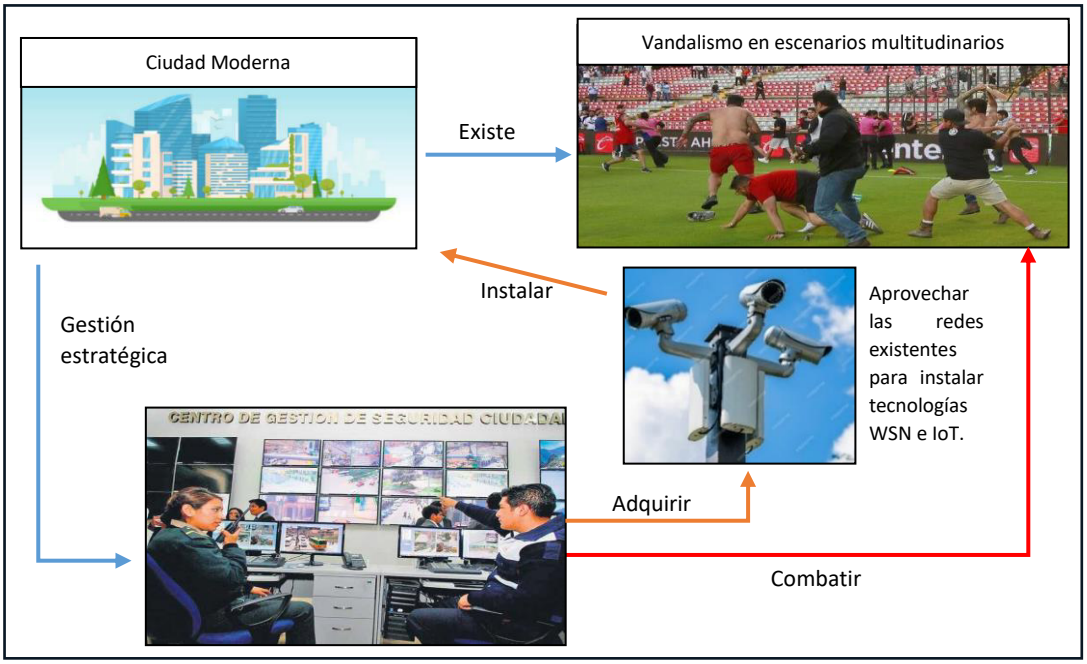


Según, Baba et al. (2019) en su investigación "Enfoque de aprendizaje profundo para la detección de violencia en áreas urbanas" comenta que los lugares con multitud de gente son características de las ciudades modernas y que estas a veces son difíciles de gestionar; una parte importante de esta tarea está relacionada con la detección de la violencia especialmente en entornos multitudinarios como conciertos de música, eventos deportivos o reuniones públicas. Entonces, para superar este problema, las ciudades modernas deben aumentar la capacidad de su seguridad o desarrollar algún sistema de monitoreo eficiente como plan estratégico, si bien la primera solución requiere de un alto presupuesto, la segunda opción podría beneficiarse de la tecnología, como Wireless Sensor Networks (WSN) e Internet Of Things (IoT) de bajo costo, donde se puede aprovechar la infraestructura ya implementada y que podría utilizarse para futuras aplicaciones aminorando los costos de instalación. Pues en general, Baba et al., indican que solo se requiere de una conexión de red y una configuración mínima; asimismo, dan a conocer que estas ventajas permitieron a otras ciudades la difusión de otros sistemas similares, como el de monitoreo de redes de sensores, captar tráfico, detectar situaciones anormales de

peatones además de vandalismo, peleas y robos. La Figura 5 resume la importancia que Baba da al plan estratégico en las ciudades modernas para combatir la violencia.

Figura 5

La importancia del plan estratégico de la seguridad ciudadana en las ciudades modernas.



Según, Becerra et al. (2019), en su investigación “Determinantes de la adopción de la inteligencia artificial en la prevención de violencia en eventos deportivos masivos de fútbol en la ciudad de Lima”, comenta que en el año 2018 ocurrieron dos eventos deportivos muy importantes: el Mundial realizado en Rusia y la Copa Libertadores, donde, en dichos eventos, la seguridad fue un factor clave para que ambos torneos terminaran con éxito. Pero, en el caso de Rusia, pese a los fuertes contingentes de seguridad, tres personas, irrumpieron en el campo de juego en el partido entre Francia contra Croacia, donde la "invasión" fue captada por las cámaras y obligó a detener el juego. En el caso de la final de la Copa Libertadores entre River Plate contra Boca Juniors, también suscitó un hecho negativo, que suspendió el encuentro ya que los hinchas de River Plate atacaron el bus que llevaba al club de Boca Juniors, lo que desencadenó acciones violentas contra la policía y, finalmente, se tuvo que jugar el

compromiso deportivo en España. Ahora, con respecto al Perú, el autor indica que el país no es ajeno a esta problemática, pues los llamados "clásicos" entre los clubes más populares como Universitario, Alianza Lima, Sporting Cristal, Sport Boys, entre otros, también generan inseguridad y que en muchos casos no asisten las dos hinchadas al mismo estadio. Se sabe que la violencia en el fútbol peruano, no solo se da durante el partido, sino antes y después del compromiso. Esto afecta los alrededores de la zona urbana de los estadios durante dichos eventos deportivos, lo que también genera un impacto económico por los desmanes a reparar, dejando en juicio la reputación de los clubes y de los organizadores, esto se puede resumir en la Figura 6. Todo esto hace que se genere una sensación insegura hacia los clubes de fútbol, hacia la Federación Peruana de Fútbol – FPF y hacia los responsables directos que no son capaces de organizar de manera segura los eventos de fútbol y prever sus consecuencias.

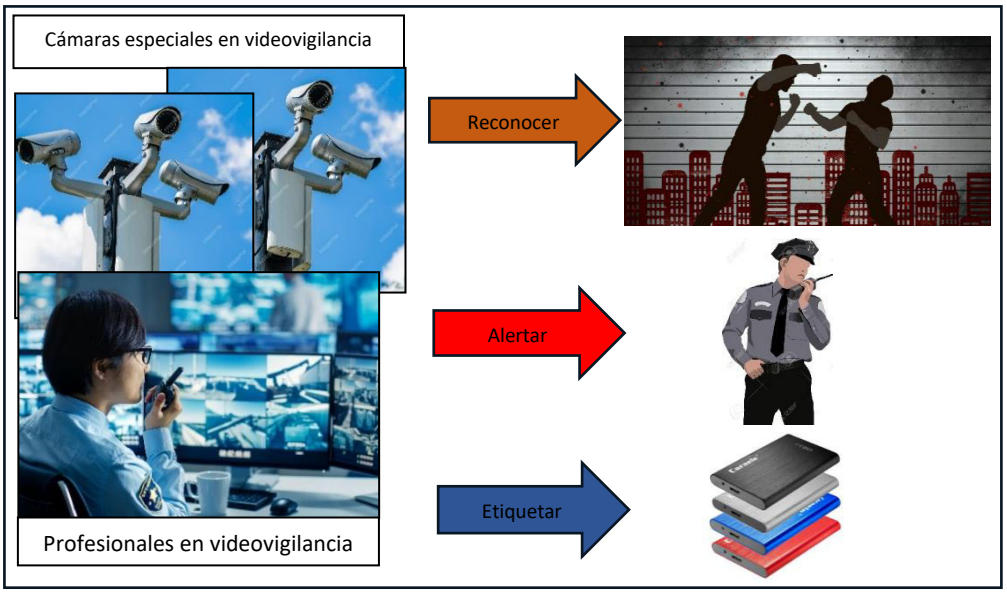
Figura 6

Los eventos deportivos de futbol y su influencia para incitar a la violencia urbana

Según, Machaca (2019) en su investigación “Reconocimiento de eventos anómalos en videos obtenidos de cámaras de vigilancia, usando redes convolucionales”, también hace énfasis de su preocupación para mantener la seguridad e indica que a la fecha se han instalado cámaras masivamente en diferentes puntos estratégicos de las ciudades, básicamente para apoyar en la detección de actos delincuenciales y alertar oportunamente a los efectivos policiales, sobre casos anómalos, como son: asaltos, peleas, etc.; sin embargo, debido al trabajo manual que realizan las personas encargadas en la supervisión de las pantallas, esto se torna insuficiente para tener una respuesta rápida y efectiva, es decir, el problema de contar con gran cantidad de cámaras se convierte en una tarea casi imposible para detectar acciones de violencia; a esto se suma la tarea de realizar el etiquetado, anotación e indexación manual de los videos para su conservación (ver Figura 7). Asimismo, Machaca menciona que se debe considerar todas las tareas que conllevan a poder detectar una acción humana, la cual resulta ser muy engorroso y complejo porque existe una gran diversidad de personas; empezando de su aspecto físico (ropa, rasgos gestuales, etc.), seguido del estilo que tienen al realizar alguna acción o movimiento, por ejemplo, no todos corren ni caminan de igual manera. Al final, Machaca menciona que existen estadísticas recientes que demuestran que el Perú es un país con altas tasas de violencia y actos criminales que se dan más en la vía pública, y que a pesar que los centros de videovigilancia van en aumento en los últimos años, surgen básicamente dos problemas fundamentales para ellos como son: primero, el personal insuficiente para la supervisión de las cámaras y segundo, el inviable control de todas ellas en paralelo siendo muy difícil detectar acciones violentas en tiempo real con un sistema tradicional.

Figura 7

La importancia de los profesionales en los centros de videovigilancia tradicional.

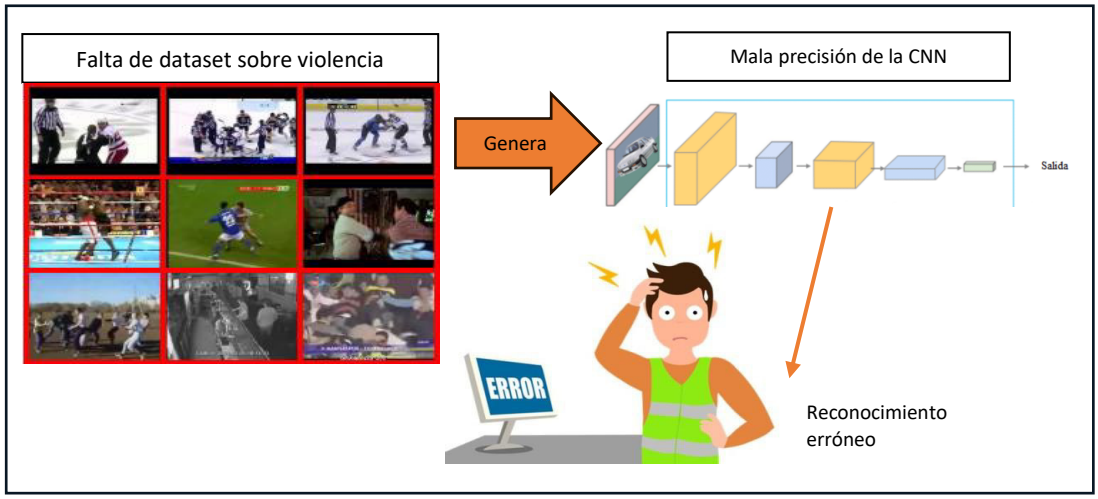


Según, Imran et al. (2020) en su investigación "Reconocimiento de actividad violenta en videos robusto, eficiente y que preserva la privacidad" comenta que hoy en día, las cámaras de videovigilancia se pueden encontrar en todos los lugares como centros comerciales, aeropuertos, escuelas, prisiones, etc. para registrar actividades en marcha, estos sistemas de vigilancia deben funcionar las 24 horas del día, los 7 días de la semana, y recopilar una gran cantidad de información en vídeos. Sin embargo, es inviable que un operador humano monitoree todas las cámaras continuamente y pueda alertar a las autoridades si se produce alguna situación peligrosa (como peleas, robos o ataques terroristas); esto conlleva a una fuerte demanda de tener un sistema de vigilancia inteligente, capaz de detectar automáticamente actividades anómalas o violentas sin ninguna intervención humana. Además, existen casos donde hay la necesidad para calificar y etiquetar automáticamente gran cantidad de videos que a veces son subidos diariamente a las redes sociales o sitios web. Estos sistemas no sólo tienen que ser precisos, sino también deben ser rápido y eficiente en memoria para poder implementarse en dispositivos móviles e integrados. Sin embargo, Imran indica que existen

dos razones por el cual no se han realizado muchas investigaciones de este tipo, primero porque existe un número muy limitado de dataset de actividades violentas que sean públicas para poder entrenar las CNN y segundo es el uso de funciones o algoritmos diseñados con características muy complejas, que no se adaptan bien a las variaciones de los acontecimientos violentos, un esquema didáctico del este problema se muestra en la Figura 8. Aunque ahora se han propuesto técnicas de última generación basadas en DNN logrando buenos resultados para la detección de la violencia, pero hay que tener en cuenta que estas técnicas son computacionalmente muy costosas para satisfacer los requisitos de un sistema de vigilancia en tiempo real.

Figura 8

La importancia del dataset y algoritmos adecuados para entrenar las CNN.



Según, Sakiba et al. (2023) en su investigación "Detección de delitos en tiempo real mediante LSTM convolucional y YOLOv7" motivada por la ausencia de las medidas preventivas para detener a los perpetradores de la violencia, que va en aumento en conjunto con las altas tasas de criminalidad en ciudades grandes con áreas metropolitanas, donde a pesar de contar con sistemas de vigilancia, que recopilan información continua, estas no son suficientes por ser tareas demasiadas complicadas, que requieren de muchos recursos y de la vigilancia humana para su detección; se sabe que hoy los Sistemas de Circuito Cerrado de televisión (CCTV) se utilizan muy comúnmente para monitorear áreas peligrosas en busca de

actividad criminal, sin embargo, las tasas de criminalidad no han disminuido a pesar de su uso generalizado en distintas zonas urbanas, puesto que, los sistemas de video vigilancia requieren de la constante supervisión humana, lo que puede provocar algunos errores como la omisión de hechos delictivos importantes mientras se monitorean a la vez numerosas pantallas del CCTV, tal como se aprecia en la Figura 9. Pero Sakiba et al. también indica que ahora es posible automatizar estos procesos que antes eran complicados, a través de la búsqueda de patrones subyacentes que pudieran indicar sobre la presencia de la intención criminal, comenta que actualmente existe una amplia variedad de algoritmos y modelos disponibles para la identificación criminal, pero, sin embargo, el estudio del reconocimiento de acciones violentas aún está en sus primeras etapas. Finalmente, la investigación de Sakiba et al. intenta identificar dos grandes categorías de conducta delictiva que son: las posturas violentas y el uso de armas letales, para lo cual realiza un análisis de imágenes y videos en tiempo real, donde, primero utilizan un ConvLSTM, que es una red neuronal recurrente con MobileNet v2 para detectar las posturas violentas y luego de manera separada prueba la versiones de YOLO v4 y v7 para identificar objetos de armas letales, como una pistola o un cuchillo, con imágenes donde una persona amenaza a otra.

Figura 9

Los Sistemas de CCTV y su falta de eficiencia para el control de la violencia.

1.5. Justificación de la investigación

1.5.1. Conveniencia

Esta investigación es conveniente porque aborda con tecnología una de las principales problemáticas sociales que actualmente enfrenta el Perú y el Mundo, como es la delincuencia, donde se planteó controlar parte del problema a través de la creación de un sistema web de apoyo al servicio de videovigilancia de una zona urbana, el cual se desarrolló a través de la técnica del reconocimiento de imágenes de acciones de violencia con el uso de la Inteligencia Artificial y Deep Learning, finalmente éste sistema es muy conveniente porque es fácil de usar e instalar.

1.5.2. Relevancia Social

Se conoce que los distintos niveles de gobierno siempre están tratando de brindar seguridad ciudadana a la población, a través de estrategias y decretos de urgencias en el marco de los estados de emergencia por inseguridad en varias ciudades del país, desplazando grandes delegaciones militares para combatirla; sin embargo, los índices de inseguridad y de violencia no han disminuido, al contrario, la actual crisis política que vive el país se convierte en un factor que incrementan el mismo. En ese sentido, es necesario contar con herramientas con IA, como el Sistema Web basado en CNN que se desarrolló, ya que se complementa con el sistema tradicional de videovigilancia para el bienestar de la sociedad, pues brinda confianza y tranquilidad el saber que se cuenta con un sistema automatizado. En un futuro, a mediano plazo, se espera que exista una red interconectada de videovigilancia automatizada entre las instituciones públicas involucradas a la seguridad ciudadana, tales como: serenazgo, la policía nacional, la defensoría del pueblo, la defensoría de la mujer entre otras, donde, estos tipos de sistemas web podrían formar parte de sus planes estratégicos.

1.5.3. Implicaciones prácticas

En efecto, con la presente investigación, se creó un datasets propio o particular de acciones de violencia, el mismo que fue desarrollado con la ayuda de los estudiantes de la facultad de ingeniería de sistemas e informática de la UNAP, el cual estuvo basado en imágenes de los videos realizados dentro de una zona urbana de la ciudad de Iquitos, y que se utilizaron para entrenar los algoritmos basados en Redes Neuronales Convolucionales (CNN) y evaluar sus indicadores de rendimiento relacionados al reconocimiento de imágenes, básicamente, con dicho dataset se entrenó a tres modelos o algoritmos CNN utilizando la técnica de Transfer Learning, entonces, se puede decir que, esto implicó de manera práctica en los alumnos de la FISI – UNAP al aprender cómo crear un dataset personalizado para entrenar algoritmos CNN.

1.5.4. Aportes

Luego de la revisión sistemática de la literatura, artículos y tesis relacionadas a esta investigación para conocer el estado del arte, se encontraron obviamente, los aportes de los autores, que sirvió para contrastar con los aportes de la presente tesis; donde, por ejemplo, se encontró el trabajo de Imran et al. (2020), quien utilizó Imágenes Dinámicas Aproximadas (ADI) en lugar de las densas representaciones del flujo óptico para reconocer actividades violentas, y así existen muchos otros aportes; sin embargo, no se encontró trabajos donde se trate de integrar las CNN con los sistemas web o apps que faciliten su uso para el usuario final; por lo que, el aporte más importante de este estudio es justamente el desarrollo de un sistema web que permite interactuar y reconocer las acciones de violencia además de alertar oportunamente a través de una interfaz web, entonces, se podría considerar a esta innovación como el principal aporte de esta investigación.

1.6. Limitaciones de la investigación

Antes de enumerar las limitaciones de esta investigación, es importante mencionar que, el desarrollo de soluciones tecnológicas en ciudades aisladas, como Iquitos, siempre presenta

un conjunto de limitantes propias del contexto geográfico y social, pero a pesar de ello se pudo superarlas en cada actividad. Entonces, algunas limitaciones encontradas fueron las siguientes:

- Débil infraestructura computacional para procesar modelos o algoritmos basados en Redes Neuronales Convolucionales, puesto que estos tipos de algoritmos son muy complejos y necesitan de la computación de alto rendimiento (HPC) para un mejor tiempo de respuesta.
- Deficiente servicio de internet por no contar con conexión de fibra óptica en varios sectores de Iquitos y aunque existe conexión satelital muchas veces es inestable debido a las precipitaciones, que interrumpen la estabilidad a los servicios de la nube, como Google Colab.
- Falta de soluciones tecnológicas similares a esta investigación, y que ayuden a la seguridad o a otros servicios dirigidos a la ciudadanía de Iquitos, pues de alguna manera hubiesen servido como base del conocimiento preliminar.

1.7. Objetivos

1.7.1. Objetivo general

Desarrollar un Sistema Web basado en Redes Neuronales Convolucionales para optimizar el reconocimiento de la violencia física en zonas urbanas.

1.7.2. Objetivos específicos

OE1: Comparar las métricas de evaluación de los modelos pre entrenados de Redes Neuronales Convolucionales para seleccionar el modelo con mejor rendimiento en el reconocimiento de la violencia física.

OE2: Evaluar el rendimiento del Sistema Web basado en Redes Neuronales Convolucionales para determinar la existencia de la violencia física dentro de una zona urbana.

OE3: Validar el rendimiento del Sistema Web basado en Redes Neuronales Convolucionales con la finalidad de medir el tiempo de respuesta para alertar la violencia física.

1.8. Hipótesis

1.8.1. Hipótesis general

El desarrollo de un Sistema Web basado en Redes Neuronales Convolucionales optimiza significativamente el reconocimiento de la violencia física en zonas urbanas.

1.8.2. Hipótesis específicas

HE1: La comparación de las métricas de evaluación de los modelos pre entrenados de Redes Neuronales Convolucionales permite obtener diferencias significativas en la selección del modelo con mejor rendimiento en el reconocimiento de la violencia física.

HE2: Con la evaluación del rendimiento del Sistema Web basado en Redes Neuronales Convolucionales se determina una concordancia significativa con el método tradicional para detectar la existencia de la violencia física dentro de una zona urbana.

HE3: Con la validación del funcionamiento del Sistema Web basado en Redes Neuronales Convolucionales se estima una diferencia significativa del tiempo de respuesta para alertar la violencia física.

II. MARCO TEÓRICO

2.1. Marco Conceptual

2.1.1. Violencia

Según, Pérez (2023) indica que la Organización Mundial de la Salud (OMS), define a la violencia como *el uso intencional del poder físico o la fuerza, como una amenaza o hechos de daños contra otras personas o hacia uno mismo*, siendo violento contra grupos de personas o comunidades, donde, los daños pueden ser físicos, psicológicos y hasta pueden causar la muerte. Asimismo, clasifica a la violencia en tres grupos, de acuerdo a las manifestaciones de quienes la cometen, siendo estos los siguientes:

- *Violencia interpersonal*: esta categoría está comprendida por la violencia entre familiares, de pareja y ancianos, también la violencia contra menores y personas sin parentesco.
- *Violencia autoinfligida:* este grupo se refiere al comportamiento suicida de las personas y las autolesiones que se producen.
- *Violencia colectiva:* en este grupo encontramos a la violencia política, social y económica entre varias personas.

2.1.1.1. Tipos de Violencia. Torres (2016) y Pérez (2023) indican que se pueden distinguir los tipos de violencia con respecto al modo en el que se intenta dañar o perjudicar, por lo que, se debe observar la naturaleza y el contenido agresivo sobre las personas más vulnerables, entre ellas tenemos:

A. Violencia familiar. Se produce por algún integrante del círculo familiar que puede ser ocasionado por alguna lesión no accidental, ya sea de manera física o psicológica. Se puede decir que este tipo de violencia está penado por la ley, sin embargo, muchas veces es un delito que no se denuncia, ya que la víctima, que normalmente es mujer, siente temor o vergüenza de denunciar a su pareja o a algún miembro de su familia.

B. Violencia de género. Llamado también como violencia sexista, es un tipo de agresión que daña física, psicológica o relacionalmente a una persona debido a su género o identidad. Es una agresión intencional, ya sea por la fuerza o con la intención de causar daño, coaccionar, limitar o manipular a la víctima, la cual puede tener efectos devastadores que pueden llevar a la discapacidad, el coma, inclusive a la muerte. Normalmente las victimas que sufren violencia de género no denuncien por temor a las represalias a él o sus seres queridos o porque creen que no recibirán apoyo adecuado.

C. Violencia verbal. Es la que pretende dañar a la otra persona con un mensaje de palabras, que puede o no contener insultos o palabras soeces, puesto que, para producir malestar psicológico, ansiedad, dañar la autoestima o la imagen pública de las personas inclusive no es necesario utilizar estas clases de recursos.

D. Violencia sexual. Es ocasionado por comportamientos y tipos de contacto físico que dañan a la persona por la causa de su alto contenido sexual, que darse inclusive a través de violaciones, previa acciones violentas física, donde el componente sexual no es un simple complemento, sino que además es una forma violenta de dañar psicológicamente a la otra persona.

E. Violencia económica. Es una acción de violencia que daña la capacidad de una o más personas para arrebatarles de alguna manera el dinero que ganan, por ejemplo, la ciberdelincuencia y la utilización indebida de cuentas bancarias, así como los engaños para realizar falsos negocios que terminan por ser una estafa.

F. Negligencia. Es una forma de violencia que se da por omisión, acá la acción consiste en no hacer nada frente a las obligaciones profesionales que deberían garantizar el bienestar del otro, por ejemplo, cuando un personal de salud se niega a brindar ayuda a algún herido por temas personales lo que conlleva a una negligencia médica.

G. Violencia religiosa. Se refiere al aprovechamiento del poder para manipular a las personas con diversas creencias y promesas del plano espiritual, por ejemplo, las sectas que realizan agresiones y amenazas hacia las personas que dan dinero para el mantenimiento de una falsa institución.

H. Violencia cultural. Es otra forma de violencia, que se manifiesta en base a una cultura de identidad sobre las creencias de los antepasados o algún legado cultural milenario, por ejemplo, normalizar la circuncisión en varones o la ablación de los genitales femeninos como parte de una costumbre.

I. Ciberbullying. El ciberbullying utiliza las redes sociales para publicar información malintencionada sobre alguna persona o grupo de personas con el fin de ridiculizar, humillar o dañar la imagen. En realidad, es un tipo de violencia de largo alcance difícil de precisar, puesto que, no se puede estimar el gran número de personas que ven el contenido en todo el mundo.

J. Violencia física. De acuerdo a lo visto anteriormente, la violencia física es parte del grupo de la violencia interpersonal, asimismo, Torres (2016) y Pérez (2023) indican que la violencia física es la más común y la más fácil de identificar, la cual consiste en cualquier acción acompañado de un daño apropósito, a través de la fuerza física, arma u objeto que pueda dañar o no el cuerpo de la otra persona. En realidad, existe un sin número de acciones consideradas como violentas, por ejemplo, los castigos corporales, puñetes, patadas, ahorcamiento, forcejeos, azotes, palmadas y lesiones penales que hasta podrían causar la muerte, de igual manera, el encerrar o inmovilizar a una persona a través de amarres, son consideradas como acciones de secuestro.

También, se considera violencia física cuando existe una invasión al espacio de la otra persona, ya sea través del contacto directo por medio de golpes o empujones u otra forma de restringir sus movimientos, por ejemplo, amenazando con arma blanca o de fuego, peor aún si es forzando a tener relaciones sexuales, entre otros las cuales pueden tener consecuencias que

van desde lo simple a lo muy grave, algunos de ellos causando lesiones o enfermedades, incapacidad para laborar, suicidios, homicidios, miedo, etc.

Es preciso indicar que no solo las mujeres son víctimas de hechos violentos, cualquiera puede ser víctima de actos de violencia sin importar su edad, raza, sexo o religión, inclusive los animales. Según, Pérez (2023) indica que la violencia es el resultado de la evolución cultural, y para combatirla sería necesario también cambiar los aspectos culturales; además, según estudios realizados en la Universidad de Wisconsin de EEUU y publicados en la revista Science, indican que: *"el cerebro humano está conectado con revisores naturales y equilibradores que controlan las emociones negativas, pero ciertas desconexiones parecen aumentar el riesgo de comportamiento violento e impulsivo"* y que estos tipos de acciones están relacionados con un componente del cerebro llamada *serotonina* que en las personas violentas al parecer están disminuida.

2.1.2. Inteligencia Artificial (IA)

Sobre la IA el Grupo Iberdrola (2023) sostiene que "Es la combinación de algoritmos diseñados para crear máquinas con las mismas características de los humanos. Una tecnología que aún resulta lejana y misteriosa, pero que desde hace unos años siempre está presente en nuestro día a día".

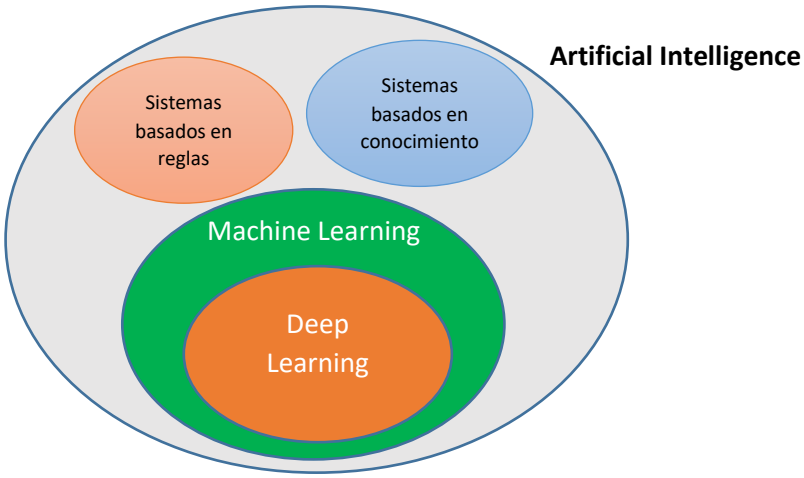
Asimismo, Boden (2016) indica que la IA "Tiene por objeto que los ordenadores hagan las mismas cosas que hace la mente, con inteligencia y visión; pero todas son capacidades psicológicas de percepción, asociación, predicción, planificación y control, que permiten a los humanos y animales alcanzar sus objetivos" (p. 11).

2.1.2.1. Artificial Intelligence, Machine Learning y Deep Learning. Según, Ketkar y Moolayil (2021) indican que la Inteligencia Artificial ha evolucionado hasta nuestros días, y se puede considerar que se ha dividido en cuatro partes tal como se muestra en la Figura 10 que es una representación clásica de la relación entre el aprendizaje automático, el aprendizaje

profundo y que ahora se suman los sistemas basados en reglas y los sistemas basados en conocimiento, la cual se refiere a los sistemas que pueden experimentar, razonar y reaccionar como seres humanos.

Figura 10

Relación entre AI, ML y DL



Nota: Esto representa la relación que existe entre la Inteligencia Artificial, el Machine Learning y el Deep Learning. Tomado de "The AI landscape" de Ketkar y Moolayil, 2021, "Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch".

2.1.2.2. Machine Learning (ML). Machine Learning técnicamente forma parte de la Inteligencia Artificial cuyo objetivo es la aplicación de algoritmos para predecir, a lo que adicionalmente se señala:

El ML es una disciplina del campo de la IA que sirve para crear sistemas que aprendan automáticamente, en ese contexto quiere decir identificar patrones de millones de datos, donde la máquina realmente aprende a partir de un algoritmo que estudia los datos y que finalmente es capaz de predecir un comportamiento futuro; asimismo, hace que los sistemas mejoren de forma autónoma y sin intervención humana (González, 2023).

El ML a menudo se asocia con términos como big data e inteligencia artificial. Sin embargo, ambos son muy diferentes al ML, entonces, para comprender mejor qué es el ML y por qué es útil, es importante comprender sobre big data y cómo es su aplicación dentro del ML. En ese sentido, big data es un término utilizado para describir enormes datasets que se crean como resultado de grandes aumentos en los datos que se recopilan y almacenan, por ejemplo, puede ser a través de cámaras, sensores o redes sociales (Vasilev et al., 2019, p. 7).

2.1.2.3. Tipos de aprendizaje de Machine Learning. Con respecto a los tipos de aprendizajes el Grupo Iberdrola (2023) indica que técnicamente los algoritmos de ML por lo menos se dividen en tres categorías de las cuales las dos primeras que se describen a continuación son las más conocidas:

A. Aprendizaje supervisado. Estos algoritmos de aprendizaje son los más comunes y estudiados, se basan en un sistema de etiquetado previo y que están asociados a los datos que les permitirán tomar decisiones o hacer predicciones; entonces, dependiendo de lo que se quiera predecir, puede usarse para resolver dos tipos de problemas: regresión o clasificación. Por ejemplo, si hay un dataset de imágenes etiquetados como “flores”, entonces el algoritmo será capaz de encontrar esas etiquetas, de manera que si se selecciona la etiqueta “flores” podrá también encontrar imágenes similares (IAT, 2023).

B. Aprendizaje no supervisado. En este caso los algoritmos no cuentan con conocimiento previo o etiquetado, se enfrentan a la incertidumbre con la finalidad de encontrar patrones que permitan ayudarlo a organizarse de alguna manera; dependiendo de lo que se desee agrupar, se puede agrupar los datos por: agrupación o asociación. Siguiendo con el ejemplo de las flores, en lugar de buscar un tipo específico de datos, se buscan patrones con similitudes que se puedan agrupar, a este tipo de aprendizaje se conoce como también Deep Learning (IAT, 2023).

C. ***Aprendizaje por refuerzo.*** En este tipo de aprendizaje la máquina aprende continuamente observando por sí misma sobre su ambiente, la información la obtiene del mundo exterior y aprende de acuerdo a ello, basado en sistema de prueba y error, es decir no se indica al software cómo debe comportarse, sino que este aprende de su entorno y modula su comportamiento siendo capaz de tomar la mejor decisión ante diferentes situaciones en el que se recompensan las decisiones correctas (IAT, 2023). Por ejemplo, se viene utilizando en el reconocimiento facial, diagnósticos médicos o clasificar secuencias de ADN.

2.1.2.4. Modelos de Machine Learning. Con respecto a los modelos de Machine Learning en el Sitiobigdata (2019) se indica que técnicamente no es fácil clasificar los algoritmos de aprendizaje por lo que la idea básica es dividir el espacio de instancias utilizando una de las siguientes formas de modelos:

A. Modelos lógicos. Estos modelos utilizan una expresión lógica para dividir el espacio de instancia en segmentos para construir modelos de agrupación, la misma que devuelve un valor booleano (verdadero o falso). Por ejemplo, los modelos de árboles de decisión, en este modelo los datos se transforman en probabilidades de actuación según una u otra regla (IAT, 2023).

B. Modelos geométricos. En modelos geométricos, las características podrían describirse como puntos en espacios de instancias en dos dimensiones (x, y) o un espacio tridimensional (x, y, z). Por ejemplo, estos modelos pueden usar líneas o planos para determinar el espacio en instancia (lineales) o usar la distancia para determinar la similitud o diferencia (IAT, 2023).

C. Modelos probabilísticos. Estos modelos ven características y variables objetivo como las variables aleatorias, este proceso representa y manipula el nivel de incertidumbre con respecto a las variables, donde normalmente se emplean las estadísticas bayesianas para definir la distribución de las probabilidades (IAT, 2023).

2.1.2.5. Técnicas de Machine Learning. Entre las técnicas de Machine Learning más destacadas por los grupos IAT (2023) y APD (2019) por ser las más comunes y populares, indican a los siguientes:

A. Árboles de decisiones. Como su nombre lo dice, esta técnica presenta una estructura de árbol para definir la toma de decisiones, los cuales lo conforman un nodo interno, una serie de ramas que contienen las reglas de decisión y los nodos hoja que representan a los resultados, es una de las técnicas más empleadas para la elaboración de modelos predictivos. Un árbol de decisión es parecido a un diagrama de flujo que usa un método de bifurcación para mostrar cada resultado posible de una decisión. Cada nodo dentro del árbol representa una prueba de una variable específica, y cada rama viene a ser el resultado de esa prueba (APD, 2019).

B. Reglas de asociación. Con esta técnica se busca desarrollar reglas de asociación que permitan definir los factores de relaciones entre variables, la cual indica la correlación como medida estadística de qué tan fuertes son las relaciones entre los atributos en un conjunto de datos (IAT, 2023).

C. Algoritmos genéticos. Estos algoritmos toman patrones y reglas basándose en cómo funciona la biología o la genética. Por ejemplo, la mutación o el cruce de datos para tener nuevas clases que permitan solucionar un problema, así como vida natural que hace evolucionar adaptándose a cualquier entorno (IAT, 2023).

D. Redes neuronales artificiales. Se basan en las neuronas del sistema nervioso humano o de algunos animales, donde cada nodo funciona como una neurona conectada con otras neuronas, creándose una red en la que todas colaboran entre sí, las cuales son capaces de convertir un estímulo de entrada en uno de salida. En esencia, es una gran cantidad de elementos interconectados, que trabajan como uno solo para desarrollar problemas específicos, aprendiendo del ejemplo y la experiencia, y finalmente sirven para modelar relaciones no

lineales con datos de alta dimensión, a veces con variables de entrada difíciles de comprender (APD, 2019).

E. Vectores de soporte. Representan a los modelos supervisados de ML, las cuales se entrenan con elementos clasificados en dos categorías, cuyo objetivo es producir un modelo que sea capaz de determinar si un nuevo ejemplo pertenece a una u otra categoría (IAT, 2023).

F. Clustering. Conocido como algoritmo de agrupamiento o clustering trabaja con modelos de aprendizaje automático no supervisado o no etiquetados donde la información se clasifica en distintos subgrupos o clusters y los elementos se integran en base a determinados criterios. El algoritmo busca el número de grupos representados por la variable K y de manera iterativa asigna a cada punto de datos uno de los K grupos según sus características proporcionadas (APD, 2019).

G. Redes bayesianas. Está representada por una serie de variables al azar y sus condicionales descritas a través de un grafo acíclico. Es un tipo de algoritmo de clasificación basado en el teorema de Bayes que clasifica a cada valor de manera independiente uno del otro, lo que permite predecir una clase o categoría teniendo en cuenta el conjunto de características, a través de la probabilidad. Este clasificador a pesar de ser simple funciona muy bien superando a los métodos de clasificación más sofisticados. (APD, 2019)

H. Algoritmos de Aprendizaje Profundo. Se puede decir que, estos algoritmos ejecutan sus datos en base a varias capas de redes neuronales, las cuales pasan una representación simplificada de los datos a la siguiente capa, funcionan muy bien con datasets que tienen cientos de características o columnas. Sin embargo, un dataset no estructurado, como por ejemplo una imagen, tiene una gran cantidad de características lo que hace que el proceso a veces se vuelve engorroso o inviable (APD, 2019).

2.1.2.6. Deep Learning (DL). El DL es un subcampo dentro del ML que se ocupa de los algoritmos que se parecen mucho a una versión muy simplificada del sistema del cerebro

humano y que resuelve una vasta categoría de la inteligencia automática moderna. Se pueden encontrar muchos ejemplos comunes dentro del ecosistema de aplicaciones de los teléfonos inteligentes (iOS y Android): detección de rostros en la cámara, autocorrección y texto predictivo en teclados, aplicaciones de embellecimiento mejoradas por IA, asistentes inteligentes como Siri/Alexa/Google Assistant, Face – ID (desbloqueo facial en iPhones), sugerencias de vídeos en YouTube, sugerencias de amigos en Facebook, filtros para gatos en Snapchat, todos productos hechos con la última tecnología solo para aprendizaje profundo. Esencialmente, el aprendizaje profundo es omnipresente en la vida digital actual. (Ketkar y Moolayil, 2021, p. 2)

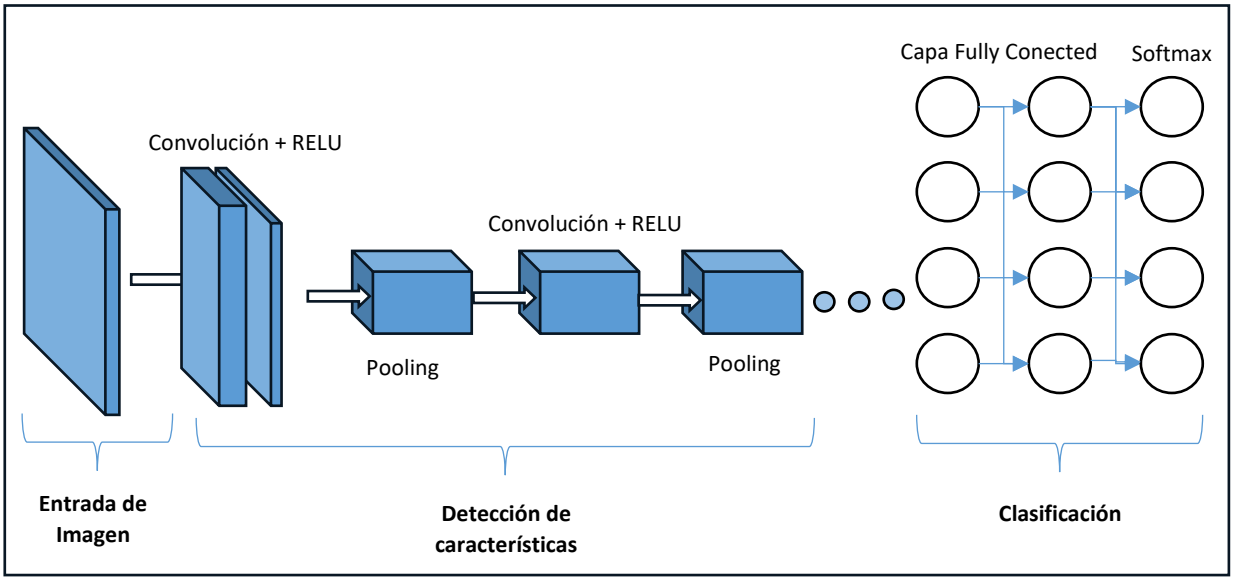
Algunos de los algoritmos de Machine Learning son muy complejos e intentan resolver o modelar los problemas, como por ejemplo los problemas de la visión por computador, donde una imagen a color es una matriz de tres dimensiones y que una señal de video, genera en promedio 30 imágenes por segundo, generándose gran cantidad de información que deben ser procesadas rápidamente. Entonces, debido a lo complejo de estos algoritmos, se llama le aprendizaje profundo o Deep Learning donde las Redes Neuronales Convolucionales son los más representativos (Suárez, 2020, p. 45).

2.1.2.7. Redes Neuronales Convolucionales (CNN). Esta estructura de red fue mencionada por primera vez en Japón y propuesta por primera vez por Kunihiko Fukushima en 1988, las cuales se presentaron como técnicas de aprendizaje automático que trata problemas muy complejos por lo que forman parte de las técnicas de aprendizaje profundo o Deep Learning (Fukushima, 1988). Son conocidos como CNN por sus siglas en inglés, y fueron diseñadas para trabajar con visión artificial que hoy en día aprovecha el poder del hardware para el procesamiento en paralelo de los GPU (Graphics Processing Unit) los mismos que ayudan mucho a implementar los modelos complejos para obtener buenos resultados.

Asimismo, las CNN se basan en las redes neuronales artificiales, sin embargo, éstas últimas solo utilizan de 2 a 3 capas mientras que los CNN pueden tener cientos de capas, las cuales se combinan de una manera no lineal, justamente inspirados en el proceso del sistema nervioso biológico, donde las neuronas se interconectan y forman redes, que a partir de un entrenamiento repetitivo aprenden características de un dataset y una vez aprendido, estas redes ya saben identificar las características similares para los nuevos casos sometidos (Suárez, 2020, 46). Los CNN básicamente está compuesta por tres partes o capas, como se puede observar en la Figura 11:

Figura 11

Partes fundamentales de las Redes Neuronales Convolucionales



Nota: Relación de las partes que conforman las CNN para el procesamiento de imágenes. Tomado de "Red Neuronal Convolucional" de Suárez, 2020, "Arquitectura de detección de actividades criminales basada en análisis de vídeo en tiempo real".

A. Entrada de imagen. En esta parte se deben ingresar las imágenes, que son los datos a ser procesados por la CNN, además acá se definen algunas características importantes del mismo como son el tamaño y color de las imágenes.

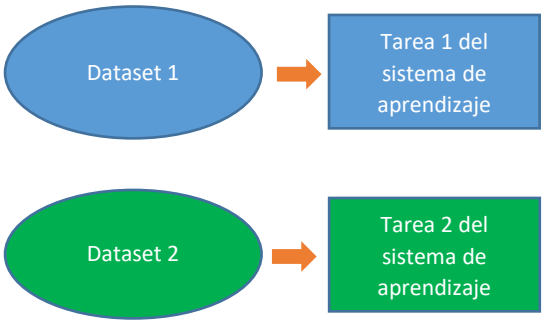
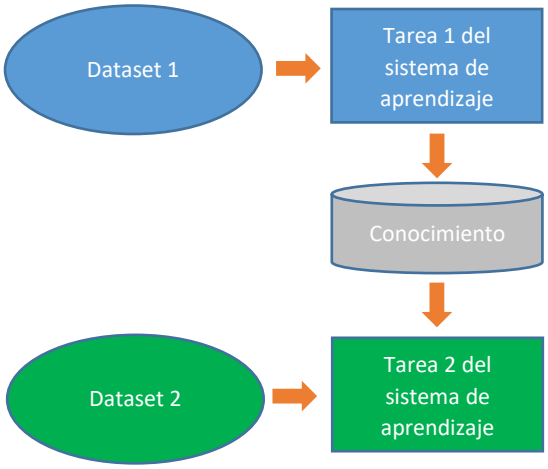
B. Detección de características. En esta parte se identifican las características de los datos que son procesados, es decir se debe aprender a identificar los principales rasgos físicos de los seres humanos; como son el torso, las extremidades (piernas y brazos) y la cabeza, este proceso se compone de tres operaciones fundamentales: la Convolución: se realiza el filtro convolucional de las imágenes para la identificar características; el Pooling: lo cual simplifica la salida a través de operaciones no lineales de downsampling, la cual reduce la cantidad de parámetros que la red debe aprender; Rectified Linear Unit (ReLU): esto ayuda a tener un entrenamiento más rápido y efectivo. Asimismo, según la arquitectura de red, varía la cantidad, posición y características sobre estas operaciones, las cuales repercuten directamente en la detección de dichas características.

C. Clasificación. En la última capa de la CNN se procede a clasificar los objetos de acuerdo a las características aprendidas y para eso existe una capa Fully Connected, la cual muestra como salida un vector de dimisión k, donde k representa al número de clases de la red a clasificar, con este vector se tienen las probabilidades de cada clase en cada objeto clasificado. Finalmente, en la arquitectura del CNN se usa una función softmax (función exponencial normalizado), la cual entrega la salida de la clasificación.

2.1.2.8. Transfer Learning. El transfer learning no es un concepto nuevo y específico del machine learning. Existe una gran diferencia entre el enfoque tradicional de construcción y capacitación de modelos de machine learning y el uso de una metodología que sigue los principios de transfer learning, esto se puedo apreciar en la Figura 12 (De Luca y Irigoitia, 2021).

Figura 12

Comparación del modelo tradicional de ML vs Transfer Learning

Machine Learning Tradicional	Transfer Learning
Aprendizaje aislado en una sola tarea: - El conocimiento no se retiene ni se acumula. El aprendizaje se realiza sin considerar conocimientos adquiridos en el pasado en otras tareas.	El aprendizaje de nuevas tareas se basa en tareas aprendidas previamente: - El proceso de aprendizaje puede ser más rápido, más certero con menor cantidad de datos de entrenamiento.
	

Nota: Relación el procesamiento de ML vs TL. Tomado de "Análisis de la técnica de Transfer Learning en Machine Learning a través de un caso de estudio: La clasificación de productos en el Banco Alimentario de La Plata." de De Luca y Irigoitia, 2021, "Traditional Machine Learning vs Transfer Learning Adaptado de: Sarkar (2018)".

El aprendizaje tradicional está aislado y ocurre únicamente en función de tareas específicas, conjuntos de datos y capacitación de modelos aislados separados entre ellos. No se retiene ningún conocimiento que pueda transferirse de un modelo a otro. Con transfer learning, se puede aprovechar el conocimiento (características, pesos, etc.) de modelos previamente entrenados para entrenar modelos más nuevos e incluso abordar problemas como tener menos datos para la tarea más nueva.

2.1.2.9. Modelo pre entrenado. Para poder definir este término, primero veamos una analogía de lo que es Transfer Learning o Aprendizaje por Transferencia, una buena analogía es la relación alumno – maestro, donde el profesor dicta un curso después de recolectar conocimientos detallados del tema, entonces, la información se transmite a través de una serie de conferencias a lo largo del tiempo, donde es el maestro es un experto y transfiere información o conocimiento a los estudiantes. Lo mismo ocurre cuando una red se entrena con gran cantidad de datos y en el proceso de entrenamiento el modelo aprende los pesos y sesgos, estos pesos se transfieren a otras redes para probar o volver a entrenar un nuevo modelo similar, es decir la red puede comenzar con pesos pre entrenadas en lugar de entrenarlos desde cero.

Una vez conocido esto, decimos que los modelos pre entrenados actúan sobre los mismos dominios que el dominio previsto, por ejemplo, para una tarea de reconocimiento de imágenes, se puede descargar un modelo Inception que ya está entrenado en ImageNet, luego, el modelo Inception se puede usar para una tarea de reconocimiento diferente y, en lugar de entrenarlo desde cero, los pesos se pueden dejar como están con algunas características aprendidas. Este método de entrenamiento es útil cuando faltan datos de muestra, a la actualidad existen muchos modelos pre entrenados disponibles (como VGG, ResNet e Inception Net en diferentes datasets).

Hay muchas razones para utilizar modelos pre entrenados, en primer lugar, se requiere mucha potencia de cálculo costosa para entrenar grandes modelos en grandes datasets. En segundo lugar, puede llevar varias semanas entrenar modelos grandes. Entrenar nuevos modelos con pesos pre entrenados puede acelerar la convergencia y ayudar a la generalización de la red. Para su uso necesitamos considerar los siguientes criterios con los respectivos dominios de aplicación y el tamaño del dataset cuando utilizamos los pesos pre entrenados que se muestran en Tabla 3.

Tabla 3

Criterios a considerar para aplicar Transfer Learning en modelos pre entrenados

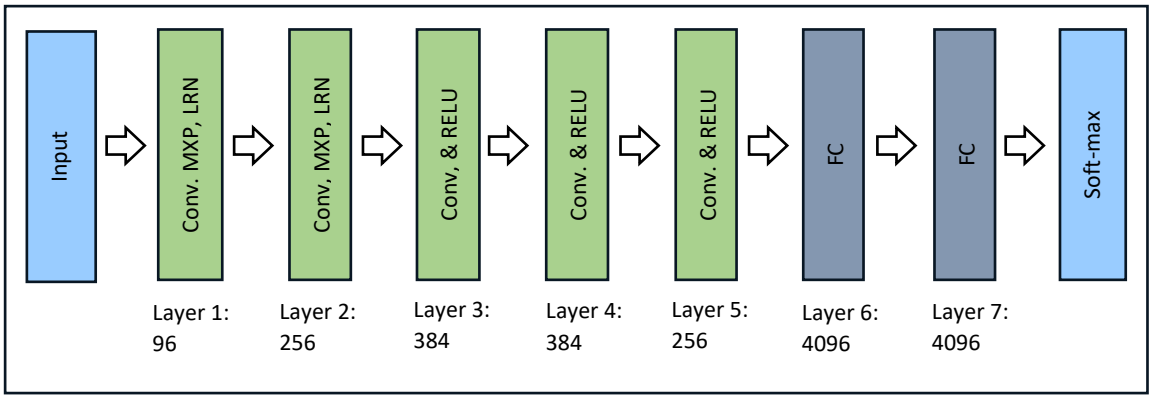
	Nuevo dataset, pero pequeña	**Nuevo dataset, pero grande**
Modelo pre entrenado en un dataset similar, pero nuevo	Congele pesos y entrene un clasificador lineal a partir de funciones de nivel superior	Ajuste todas las capas (entrenamiento previo para una convergencia más rápida y una mejor generalización)
Modelo pre entrenado en un dataset diferente, pero nuevo.	Congele pesos y entrene un clasificador lineal a partir de funciones que no sean de nivel superior	Ajuste todas las capas (entrenamiento previo para mejorar la velocidad de convergencia)

2.1.2.10. Modelos pre entrenados de Redes Neuronales Convolucionales. Como se sabe en los últimos años se han venido desarrollando distintas arquitecturas o modelos de redes neuronales convolucionales, las cuales cada vez son más profundas llegando a los cientos de capas. Entre los modelos pre entrenados destacan:

A. AlexNet. Es el primer modelo CNN presentado, propuesto en el 2012 por Alex Krizhevsky junto con su director de tesis Geoffrey Hinton, en el concurso de reconocimiento de imágenes ImageNet Large Scale Visual Recognition Challenge (ILSVRC) donde logró superar uno de los parámetros de evaluación: "error de las 5 primeras (top-5, error del 25%)" con una gran mejora de hasta el 15.3% en comparación con otras técnicas de aprendizaje. AlexNet solo está formado por 5 capas de convolución, dos capas MaxPool, dos fases de normalización, dos capas fully connected y una capa Softmax, como se observa en la Figura 13 (Krizhevsky et al., 2012); AlexNet fue muy popular, sin embargo, posteriormente presentó modificaciones como ZFNet con mejora en los hiperparámetros de AlexNet con la cual se obtuvo mejores resultados para la detección.

Figura 13

Arquitectura de la Red Neuronal Convolucional AlexNet



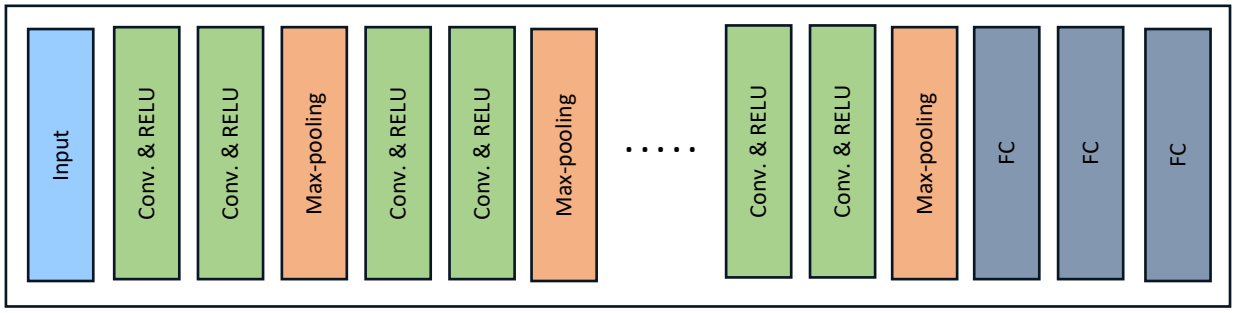
Nota: Esquema de la estructura interior de la CNN AlexNet. Tomado de "Architecture of AlexNet: Convolution, max-pooling, LRN and fully connected (FC) layer" de Alom et al., 2018, "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches".

B. VGGNET. Propuesto en el 2014 por Visual Geometry Group (VGG) de Karen Simonyan y Andrew Zisserman pertenecientes a la Universidad de Oxford, donde presentaron los modelos denominados "muy profundos" VGG-16 y VGG-19 siendo finalista en el concurso de ILSVRC logrando reducir el margen de error al 7.3%. La principal contribución de este trabajo es que muestra que la profundidad de una red es un componente crítico para lograr una mejor precisión de reconocimiento o clasificación en CNN (Simonyan y Zisserman, 2014). La arquitectura VGG consta de dos capas convolucionales, las cuales utilizan la función de activación ReLU, después de la función de activación hay una única capa de agrupación máxima y varias capas fully connected que también utilizan una función de activación ReLU, como se muestra en la Figura 14. La capa final del modelo es una capa Softmax para clasificación. En VGG, el tamaño del filtro de convolución se cambia a un filtro de 3x3 con un

paso de 2. A los modelos VGG como: VGG-11, VGG-16 y VGG-19 se propusieron que tuvieran 11, 16 y 19 capas respectivamente (Alom et al., 2018).

Figura 14

Esquema básico de la arquitectura de los modelos VGG



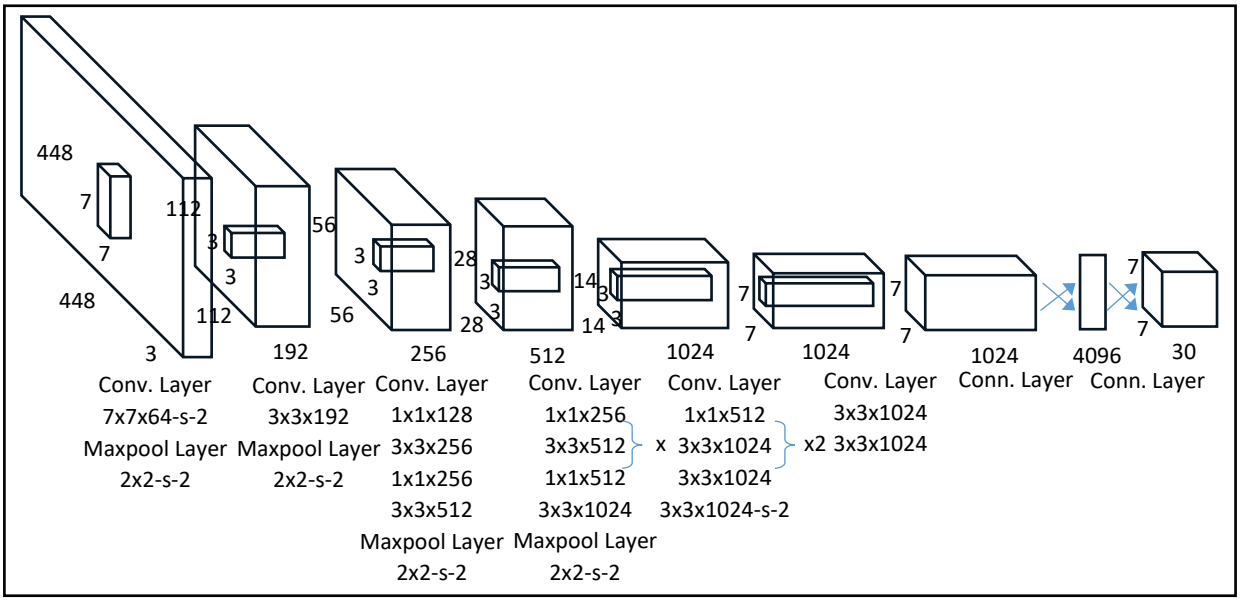
Nota: Todas las versiones de los modelos VGG terminan igual con tres capas fully connected. Sin embargo, el número de capas convolucionales varía VGG-11 contiene 8 capas convolucionales, VGG-16 tiene 13 capas convolucionales y VGG-19 tiene 16 capas convolucionales. El modelo VGG-19 es computacional más caro, contiene 138 millones de pesos y 15,5 millones de MAC. Tomado de "Basic building block of VGG network: Convolution (Conv) and FC for fully connected layers" de Alom et al., 2018, "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches".

C. YOLO (You Only Look Once). Propuesto en el 2015, por Joseph Redmon y otros investigadores como un nuevo enfoque para la detección de objetos que es más rápido en comparación con otros modelos como los Modelo de Partes Deformables (DPM) y los modelos de análisis de imágenes basadas por regiones (R – CNN) donde se replantea la detección de objetos como un problema de regresión simple, directamente desde los píxeles de la imagen hasta las coordenadas del cuadro delimitador y probabilidades de clase por eso se denomina "***You Only Look Once"*** (usted sólo mira una vez) una imagen para predecir qué objetos están presentes y dónde se encuentran; es decir, a diferencia de los procesos descrito anteriormente, YOLO realiza en una sola todas las fases del proceso de detección (Redmon et al., 2016).

YOLO tiene como base a GoogLeNet, que es un modelo para clasificación de imágenes, sin embargo, esta red YOLO tiene 24 capas convolucionales seguidas de 2 capas fully connected y en lugar de los módulos iniciales utilizados por GoogLeNet, YOLO utiliza una reducción de 1 x 1 capas de reducción seguidas de 3 x 3 capas convolucionales, tal como se muestra en la Figura 15 (Peñafiel, 2022).

Figura 15

Arquitectura de la Red Neuronal Convolucional YOLO



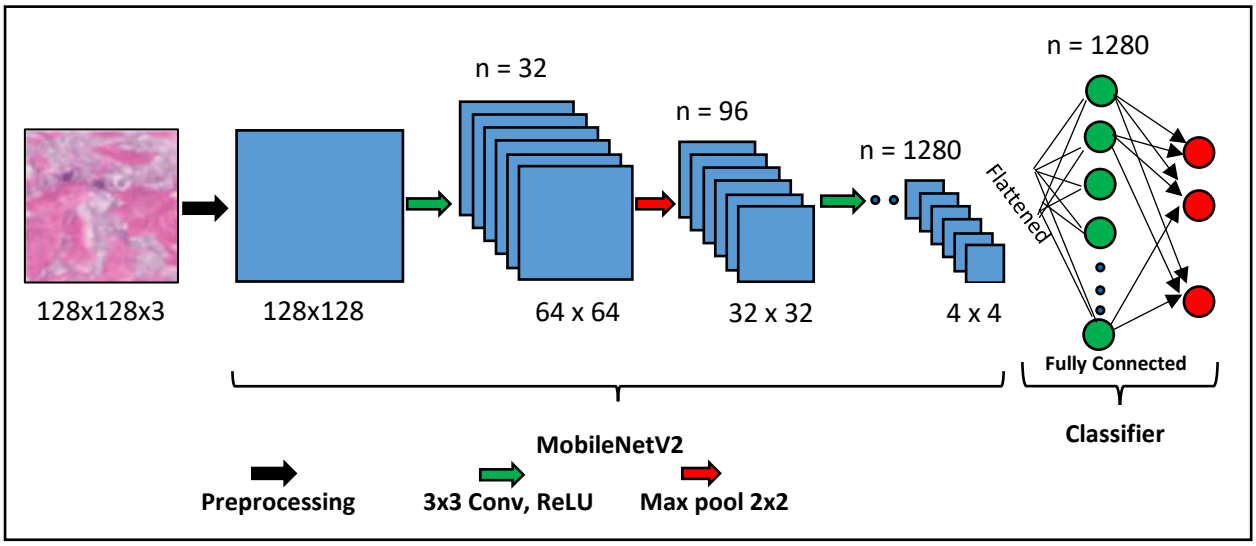
Nota: Esquema de la arquitectura de YOLO. Tomado de "The Architecture" de Redmon et al., 2016, "You Only Look Once: Unified, Real-Time Object Detection".

D. MobileNetV2. Propuesto en el 2017, es una arquitectura ligera de red neuronal convolucional (CNN), está diseñada específicamente para aplicaciones de visión integradas y móviles. Los investigadores de Google lo desarrollaron como una mejora del modelo MobileNet original. Otro aspecto notable de este modelo es su capacidad para lograr un buen equilibrio entre el tamaño y la precisión del modelo, lo que lo hace ideal para dispositivos con recursos limitados. MobileNetV2 incorpora varias características clave que contribuyen a su eficiencia y eficacia en las tareas de clasificación de imágenes, cuya arquitectura se observa en

la Figura 16. Estas características incluyen convolución separable en profundidad, residuos invertidos, diseño de cuellos de botella, cuellos de botella lineales y bloques de compresión y excitación (SE). Cada una de estas características juega un papel crucial en la reducción de la complejidad computacional del modelo manteniendo una alta precisión (Sharma, 2023).

Figura 16

Esquema básico de la arquitectura de MobileNetV2



2.1.2.11. Métricas de evaluación de desempeño

A. Matriz de confusión. Luego de implementar el modelo de clasificación, se obtienen los valores predichos versus los valores reales. Con estos valores se construye la matriz de confusión, que es una tabla cruzada entre las clases predichas y reales (Konasani y Kadre, 2021), como se observa en la Tabla 4.

Tabla 4

Criterios a considerar en una Matriz de confusión

		Predicción	
		Positivo	Negativo
Clase Real	Positivo	Verdaderos Positivos (TP)	Falsos Positivos (FP)
	Negativo	Falsos Negativos (FN)	Verdaderos Negativos (TN)

La matriz de confusión es una herramienta muy útil para evaluar el desempeño de los modelos de clasificación, donde podemos obtener cuatro valores importantes como son:

- *Verdaderos Positivos (TP),* es el número de casos predichos como pertenecientes a una clase, y que efectivamente pertenecen a esa clase, es decir, son las predicciones correctas.
- *Verdaderos Negativo (TN),* indica el número de casos predichos como no pertenecientes a una clase, y que verídicamente no pertenecen a esa clase.
- *Falsos Positivos (FP),* son los casos predichos como pertenecientes a una clase y que realmente no pertenecen a la clase.
- *Falsos Negativos (FN),* son los elementos predichos como no pertenecientes a una clase, cuando en realidad estos sí pertenecen a esa clase (Akosa, 2017). Entonces, en base a esto términos de la matriz de confusión podemos determinar las siguientes métricas de evaluación:

B. Exactitud. La exactitud o accuracy es una métrica que mide el valor verdadero que se puede encontrar en la clasificación de imágenes, y se representa mediante la siguiente fórmula (1).

$$\frac{TP + TN}{TP + TN + FP + FN} \dots\dots\dots\dots (1)$$

C. Precisión. Mide la habilidad del clasificador para determinar la clase correctamente, y se representa mediante la fórmula (2).

$$\frac{TP}{TP + FP} \dots\dots\dots\dots (2)$$

D. Sensibilidad / Recall. Se refiere a la habilidad del clasificador en predecir las muestras de interés o muestras positivas a clasificar y se representa mediante la fórmula (3).

$$\frac{TP}{TP + FN} \dots\dots\dots\dots (3)$$

E. Especificidad. Estima el porcentaje del número de casos negativos entre un grupo de casos a clasificar y se representa mediante la fórmula (4).

$$\frac{TN}{TN + FP} \dots\dots\dots\dots (4)$$

F. F1-score. Indica la relación del clasificador en alcanzar mejores resultados en las clases de interés o prioritarias, y se representa mediante la fórmula (5).

$$\frac{2 * Precisión * Sensibilidad}{Precisión + Sensibilidad} \dots\dots\dots\dots (5)$$

G. G_mean. Mide el balance de desempeño sobre las clases con mayor y menor número de elementos, y se representa mediante la fórmula (6).

$$\sqrt{(Sensibilidad * Especificidad)} \dots\dots\dots\dots (6)$$

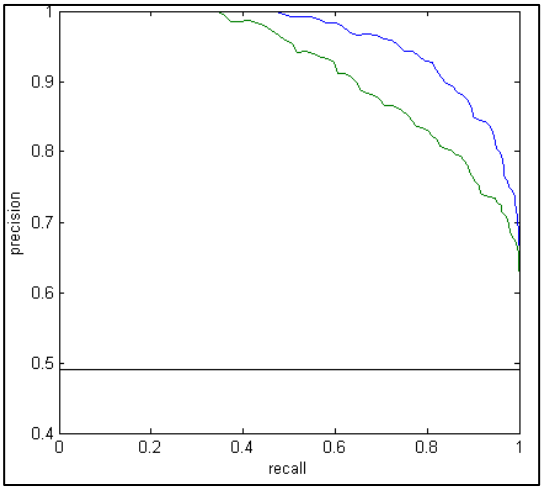
H. Index Balanced Accuracy (IBA). Calcula el promedio de la exactitud obtenida entre las clases con mayor y menor número de casos predichos correctamente, y se representa mediante la fórmula (7).

$$[1 + (0.1)(Sensibilidad - Especificidad)][G\_mean^2] \dots\dots\dots\dots (7)$$

I. Curva de Precisión – Recall (PRC). Una PRC es simplemente un gráfico con valores de precisión en el eje Y y valores de recall en el eje X, como por ejemplo se observa en la Figura 17.

Figura 17

Ejemplo de gráfico de la Curva Precisión – Recall



J. Mean Average Precision (mAP). La precisión promedio media es una métrica que se utiliza para medir el rendimiento de un modelo en tareas como la detección de objetos, se calcula encontrando la precisión promedio (AP) para cada clase y promediando entre todas las clases, su fórmula es (8).

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k \quad \ldots\ldots\ldots\ldots (8)$$

Donde:

n = número de clases

AP_k = precisión promedio de las clases

2.1.3. Sistemas web

Son un tipo de aplicación a la que se puede acceder a través de Hypertext Transfer Protocol (HTTP), el término “basado en web” normalmente se utiliza para describir aplicaciones que se ejecutan dentro de un navegador web. Sin embargo, también se puede utilizar para hablar sobre aplicaciones que tienen un componente muy pequeño de la solución

dentro de la PC del cliente, donde el servidor principal de un sistema basado en web podría ser un servidor local a la que también se podría acceder a él utilizando Internet (Aezion, 2018).

Asimismo, se denomina sistema web a las aplicaciones que pueden utilizarse a través del acceso a un servidor web de manera remota o intranet, mediante un browser, hoy en día se usan más en las empresas, por las múltiples bondades que ofrecen, entre ellas: multiplataforma, seguridad de la información, flexibilidad, económico, disponibilidad, y que últimamente no requiere del aprendizaje adicional de nuevos programas. Por otro lado, facilitan la accesibilidad al trabajo colaborativo y en línea (Mayorga et al., 2022).

2.1.3.1. Sistema web inteligente. La Web 4.0, se conoce también como "Web inteligente", es la siguiente etapa evolutiva en el desarrollo de Internet. Se diferencia por un cambio de las páginas web estáticas tradicionales a aplicaciones web más inteligentes, dinámicas e interactivas. La Web 4.0 representa un importante paso adelante en la evolución de Internet y la manera en que interactuamos con ella (GeeksforGeeks, 2025).

Los sistemas de información inteligentes tienen componentes de software capaz de simular la Inteligencia humana. Estos tipos de sistemas tienen la capacidad de observar y rastrear los cambios que suceden en el medio ambiente. Dependiendo de lo implementado, conocimiento y modelos heurísticos, pueden responder a estos cambios hasta cierto punto. La disciplina de la inteligencia artificial (IA) se ocupa de la construcción de inteligencia de estos ordenadores capaces de resolver tareas de alta complejidad, cuyas habilidades pueden ser incluso competitivas con las de un humano (Achkoski et al., 2012).

2.1.3.2. Tipos de sistemas web. En esta sección se describen los diferentes tipos de sistemas web según Kienle y Distante (2014) con respecto a su evolución, a los cuales clasifican como: sitios web estáticos, aplicaciones web, servicios web, aplicaciones enriquecidas de internet basadas en Ajax, computación en la nube y en HTML5.

A. Sitios web estáticos. Corresponde a la primera ola tecnológica de la web, que consistió en sitios web estáticos codificados principalmente en HTML, tanto así que en un artículo se generó conciencia y popularizó la noción de que la web estaba predispuesta a convertirse en "la próxima montaña de mantenimiento" como punto de partida para futuras investigaciones, se reconoció que las características de los sitios web podrían ser considerados como software y, por lo tanto, que era necesario realizar investigaciones sobre su evolución en áreas del proceso de desarrollo, gestión de versiones, pruebas y estructura de despliegue (Kienle y Distante, 2014).

B. Aplicaciones web. A lo largo de los años, surgieron nuevos sitios web evolucionados con un comportamiento dinámico tanto en el lado del cliente (a través de JavaScript) como en el lado del servidor (vía CGI y PHP). Esta nueva generación se denominó aplicaciones web, adaptándose a la creciente sofisticación de las aplicaciones web de esos años, que es un reflejo de la Web 2.0, tomando también el término de Aplicaciones Ricas de Internet (RIA) para distinguir estas aplicaciones web técnicamente complejas de las más primitivas. Una aplicación web suele estar basada en eventos que activan JavaScript y que de hecho provocan un cambio de estado en el otro extremo, así como también existen las páginas basadas en Ajax que dan como resultado un modelo de navegación basada por formularios de consulta interactivos y que acceden a una base de datos del lado del servidor (Kienle y Distante, 2014).

C. Servicios web. En la época en que se establecieron las aplicaciones web, el concepto de los servicios web comenzaron a ser más prominentes, este concepto fue impulsado principalmente desde una perspectiva empresarial que preveía ahorros de costos y mayor flexibilidad. Los servicios web están estrechamente relacionados con la Arquitectura Orientada a Servicios (SOA) en el sentido de que los servicios web son una tecnología que permite realizar un sistema que se adhiere al estilo SOA, donde la migración hacia servicios web puede verse como una arquitectura de evolución que implica desafíos importantes como:

componentes distribuidos con múltiples propietarios, ejecución distribuida, archivos generados por máquinas (por ejemplo, WSDL, XSD y BPEL) y mensajes (por ejemplo, mensajes SOAP). Comprender un servicio web y el lenguaje de descripción (WSDL) puede ser compleja porque contiene otros conceptos (tipos, mensajes, tipos de puertos, enlaces y servicios) que pueden estar altamente interrelacionados (Kienle y Distante, 2014).

D. Sistemas web basados en Ajax. Otra gran evolución de la web viene marcada por la posibilidad de una aplicación web con conexión asincrónica para obtener datos e información de presentación, donde se emplean frameworks y tecnologías Ajax, dando como resultado aplicaciones web altamente sofisticadas cuya funcionalidad supera a las aplicaciones de escritorio. En comparación con los sitios web y aplicaciones web tradicionales en los que los contenidos y funcionalidades se distribuyen entre varias páginas, la web basada en Ajax consiste en una sola web cuyo contenido y funcionalidades de usuario cambian dinámicamente y sin recargas páginas; donde, además, se utiliza JavaScript como principal lenguaje para implementar funciones, tanto del lado del cliente como del lado del servidor (Kienle y Distante, 2014).

E. Sistemas web basados en computación en la nube. Podría decirse que la computación en la nube es otro paso importante en la evolución de la web, sin embargo, cabe destacar que la computación en la nube es un principio independiente que se aplica a los sistemas de software en general. La “ejecución de aplicaciones dentro de un servidor de red o la descarga del software de la red cada vez que se ejecuta el software” es una de sus características destacadas, entonces, la computación en la nube se puede definir como "un modelo para permitir un acceso conveniente a la red bajo demanda de un grupo compartido de recursos informáticos configurables (redes, servidores, almacenamiento, aplicaciones, y servicios) que pueden ser rápidamente aprovisionados y liberados con un mínimo esfuerzo de gestión o interacción con el proveedor del servicio”, es preciso indicar, que el acceso a estos

recursos no tiene porqué realizarse desde una interfaz de usuario de un navegador ni con tecnologías basadas en la web (Kienle y Distante, 2014).

F. Sistemas web basados en HTML5. Hoy en día el estándar HTML5 es una realidad, y ya está en camino de convertirse en una plataforma omnipresente para construir todo tipo de sistemas web. Es decir, HTML5 también toma un papel importante en la evolución de la web. Sin embargo, hay que tener en cuenta en este punto, ya que existen dos enfoques principales en competencia: los sistemas web basados en HTML5, que son independientes del proveedor y representan la open web, y los sistemas web nativos, que son específicos del proveedor. Ejemplos de esto último son las aplicaciones para iPhone y iPad de Apple o en dispositivos Android de Google. Si bien en estas aplicaciones se pueden utilizar protocolos y principios basados en web (HTML y REST) y mientras su apariencia puede ser similar a la de los sistemas basados en navegador, están construidos con plataformas específicas del proveedor y bibliotecas de gráficos nativas, y generalmente se distribuyen en forma binaria, entonces, los que queda ver es si ambos enfoques coexistirán o si uno extinguirá al otro (Kienle y Distante, 2014).

2.1.3.3. Sistema Web de reconocimiento. Tiene como objetivo identificar elementos dentro de una escena, que pueden pertenecer a distintas clases o categorías a partir de una misma imagen, con el apoyo de las redes neuronales; las persona realizan esta tarea de manera natural, sin embargo, para ser implementada en un computador resulta muy complejo, por sus diversas técnicas algorítmicas. Este proceso es conocido también como visión por computadora, la misma que se basa en redes neuronales convolucionales y hoy en día es una de los técnicas más utilizadas para implementar estos tipos de sistemas, para lo cual se debe tener como base una serie de entrenamientos y ajuste de los parámetros internos de los algoritmos para la obtención de las características propias, apariencia, conocimiento geométricos, etc. con la finalidad de lograr reconocer con éxito los objetos de la imagen, para

lo cual se requiere de alto rendimiento computacional y gran capacidad de datos para el entrenamiento (Ruiz Sarmiento et al., 2020).

2.1.4. Zona urbana

Según en la definición de los Censos de 1972, 1981 y 1993 realizados en Perú, una zona urbana es aquella donde vive la aglomeración de gente cuyas viviendas se ubican contiguamente, y que normalmente habitan en todas las capitales de los distritos (Comisión Económica para América Latina y el Caribe [CEPAL], 2013).

La diferencia entre zona urbana y rural se refiere a la ciudad y al campo respectivamente, son dos tipos de espacios de hábitat de los seres humanos, donde se concentra la mayor parde de la población y que por siglos han sido motivo de disputa por reconocerse como tal, pero que, sin embargo, ambos espacios son complementarios y la tendencia es que poco a poco sean considerados homólogos por el rápido crecimiento urbano a nivel mundial (Raffino, 2023).

2.1.4.1. Características de las zonas urbanas

- Tiene espacios urbanizados, con edificaciones y obras públicas de material noble o concreto.
- Concentra la mayor parte de la población, según el Banco Mundial el 56% de la población vive en zonas urbanas.
- Es donde se realiza la mayoría de las actividades económicas, culturales y tecnológicas de un país.
- Posee un alto nivel de contaminación por concentrar la mayor parte del parque automotor e industrial.
- Sin importar sus dimensiones a la población considerada como ciudad también se debe considerarla como zona urbana.

III. MÉTODO

3.1. Tipo de investigación

El tipo de investigación es aplicada, porque se desarrolla un Sistema Web basado en Redes Neuronales Convolucionales, para solucionar el problema del reconocimiento automatizado de la violencia física en una zona urbana.

3.1.1. Nivel de la investigación

La Investigación es de nivel descriptivo y predictivo. Es descriptivo porque se describe los resultados obtenidos de las fichas de evaluación y las observaciones realizadas, estableciendo así la relación entre las variables y permitiendo tener un conocimiento del fenómeno que se presenta. Es predictivo porque se predice el reconocimiento de la violencia física a través del procesamiento automatizado de las imágenes utilizando un modelo de Red Neuronal Convolucional.

3.1.2. Diseño de investigación

El diseño es cuasiexperimental con post – test y con grupo de control:

G_e	X	O_1
G_c	--	O_2

Donde:

G_e = Grupo Experimental: Grupo de estudio al que se le aplicará el estímulo (Sistema Web basado en Redes Neuronales Convolucionales)

G_c = Grupo Experimental: Grupo de estudio al que no se le aplicará el estímulo.

O_1 = Datos de la Post Test para los indicadores de la VD: Mediciones en el grupo experimental.

O_2 = Datos de la Post Test para los indicadores de la VD: Mediciones en el grupo de control.

X = Sistema Móvil con Machine Learning: Estímulo o condición experimental.

-- = Falta de estímulo o condición experimental.

Se trata de la confrontación de forma intencional de un grupo G_e, cuyos elementos son escogidos por conveniencia, conformada por acciones de violencia física en espacios urbanos, al que se le aplicó un estímulo de Sistema Web basado en Redes Neuronales Convolucionales (X), luego del cual se le aplica una prueba posterior a dicho tratamiento (O_1). A un segundo grupo G_c, con elementos también escogidos por conveniencia, conformada por acciones de violencia física en espacios urbanos, al que no se le administra ningún estímulo, sirviendo sólo como grupo de control; al cual también se le aplica una prueba posterior (O_2), se espera que los valores de O_1 sean más óptimos que los valores de O_2.

3.2. Población y muestra

Tabla 5

Población y muestra

Unidad Muestral:	Proceso de reconocimiento de la violencia física en espacios urbanos de las ciudades del mundo. Limitaciones: - Ciudades con cámaras de videovigilancia. - Ciudades con equipo servidor de alta gama (workstation). - Nivel mundial.
Universo	Todos los procesos de reconocimiento de la violencia física en espacios urbanos de las ciudades que cuenten con cámaras de videovigilancia instaladas y equipo servidor de alta gama (workstation) a nivel mundial. Debido a que no se puede conocer ni determinar la cantidad de procesos antes mencionado, por lo tanto, se tiene: N = indeterminado
Muestra	Proceso de reconocimiento de la violencia física en espacios urbanos de la ciudad de Iquitos. n = 30

Acerca del tamaño de la muestra, según (Pande et al., 2004) en su libro “Las claves prácticas de SIX SIGMA”, se tomó una muestra de 30 procesos.

3.2.1. Tipo de muestreo

El tipo de muestreo está definido de manera **aleatoria**

3.3. Operacionalización de variables

Figura 18

Relación entre variables, dimensiones e indicadores

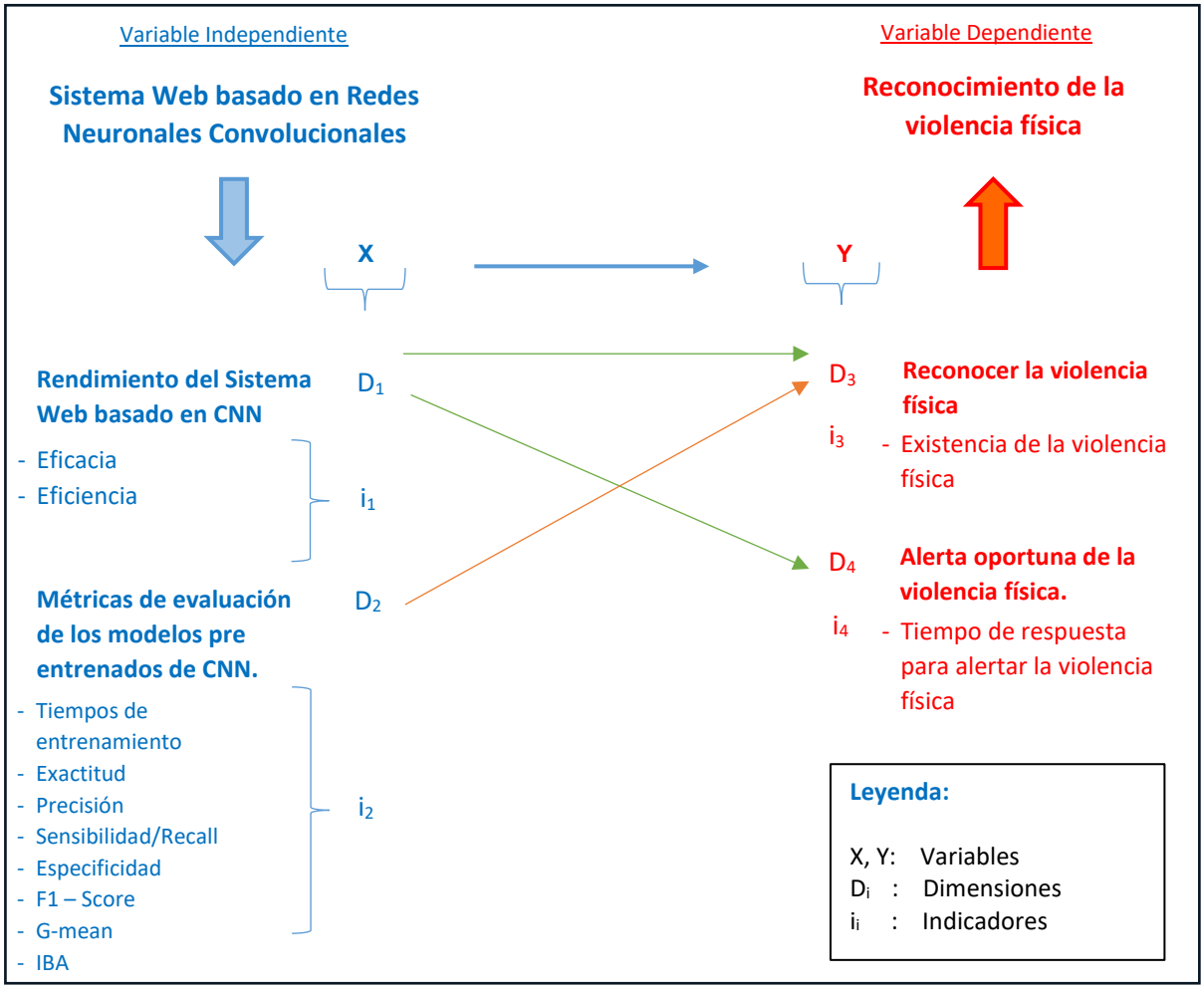


Tabla 6

Operacionalización de variables

Tipo	Variables	Definición Conceptual	Definición Operacional	Dimensiones	Indicadores	Descripción
Independiente	Sistema Web basado en Redes Neuronales Convolucionales.	Perspectivas para que el usuario perciba la calidad del producto cuando es usado en un ambiente y en un contexto de uso específico.	Se evalúa la eficacia del sistema utilizando una ficha de observación para determinar la concordancia del sistema con respecto al método tradicional con la ayuda de los usuarios expertos del proceso desarrollado.	Rendimiento del Sistema Web basado en CNN	- Eficacia - Eficiencia	La evaluación de estos indicadores busca determinar el nivel de calidad del del producto en relación al funcionamiento juntamente con las características más relevantes para los usuarios.
		Son componentes integrales para cualquier proyecto de CNN, cuyo objetivo es estimar el aprendizaje de un modelo para asignar alguna observación a ciertas clases sobre los datos futuros	Se evalúa cada una de las métricas que tienen su propia fórmula enmarcada en los resultados de la matriz de confusión, lo cual, generalmente se presentan en un rango de 0 a 1, donde la puntuación 1 especifica como un buen rendimiento.	Métricas de evaluación de los modelos pre entrenados de CNN.	- Tiempos de entrenamiento. - Exactitud. - Precisión. - Sensib/Recall - Especificidad - F1-Score - G-mean - IBA	Estos indicadores miden el rendimiento de los modelos pre entrenados de CNN para el pre procesamiento y posterior reconocimiento de las acciones de violencia física, las cuales permitirán una elección adecuada y oportuna para la utilización del modelo.
Dependiente	Reconocimiento de la violencia física.	Medida aplicada para la predicción y/o el reconocimiento de la violencia física o interpersonal, vale decir entre dos personas.	Se estima una escala aplicada sobre la existencia de la violencia entre los involucrados en una acción violenta física, bajo un puntaje que puede estar en un rango de 0 a 1 punto, donde una aproximación a 1 significa que existe violencia física.	Reconocer la violencia física.	- Existencia de la violencia física.	Este indicador da a conocer la existencia de la violencia física, además del tipo que puede ser: patada, puñete, forcejeo o estrangulamiento; con la cual también se podría estimar el nivel de peligrosidad de la violencia.
		Intervalo de tiempo desde la detección o reconocimiento de la acción de violencia física hasta la comunicación o emisión de alerta a los usuarios involucrados.	Mide el tiempo que transcurre desde que se reconoce la violencia física hasta que el sistema emite una alerta a los usuarios de acuerdo al nivel de la violencia.	Alertar oportunamente la violencia física.	- Tiempo de respuesta para alertar la violencia física.	Un indicador importante para determinar la eficiencia del funcionamiento del sistema web con el modelo CNN, estimando así el tiempo de respuesta para alertar la violencia frente al nivel de su peligrosidad.

3.4. Instrumentos

Los instrumentos para la recolección de la información documental fueron:

- PC, Memoria USB, HD
- Filmaciones públicas de violencia física.
- Fotografías públicas de violencia física.

Los instrumentos para la recolección de la información experimental fueron:

- Fichas de observación
- Cámara Fotográficas
- Celular con cámara de alta gama
- Internet: Google Colab

Los instrumentos para la recolección de la información de campo fueron:

- Ficha de observación.
- Filmaciones propias de violencia física.
- Fotografías propias de violencia física.

3.5. Procedimientos

Para el procedimiento documental se revisó lo siguiente:

- Libros
- Tesis
- Internet: Kaggle, UCI ML
- Artículos científicos
- Revistas científicas

Para el procedimiento experimental se realizó lo siguiente:

- Ejecución de experimentos con distintos algoritmos pre entrenados de CNN.
- Seguimiento de la evolución del rendimiento de los algoritmos de CNN en el reconocimiento de las acciones de violencia física.
- Se desarrolló un sistema web que sirva como plataforma tecnológica para interactuar con el algoritmo de YOLOv8.

Para el procedimiento de campo se realizó lo siguiente:

- Observación directa: Individual, de los resultados obtenidos.

3.6. Análisis de datos

Las actividades que se realizaron para el análisis de datos fueron:

- Coordinación sobre los requerimientos necesarios para la implementación del sistema web para el reconocimiento automatizado de violencia física en zonas urbanas.
- Diseñar las fichas de observación.
- Aplicación de los instrumentos de recolección de datos para obtener información.
- Procesamiento de la información obtenida.
- Representación de la información mediante tablas de frecuencia y gráficos.
- Análisis e interpretación de la información.

Finalmente, la información obtenida fue procesada de manera automatizada y/o computarizada utilizando Microsoft Excel y el paquete estadístico computacional SPSS, para demostrar la estadística descriptiva e inferencial respectivamente; para las pruebas de contrastación de hipótesis con datos paramétrico se utilizó las pruebas estadísticas de ANOVA de un factor, el índice de Kappa de Cohen y t – Student para muestras independientes.

3.7. Consideraciones éticas

Durante el desarrollo de este proyecto de investigación se tuvo en cuenta las siguientes consideraciones éticas:

- Se respetó los derechos de autor de las fuentes utilizadas, toda fuente de información fue citada.
- Todas las fuentes de información fueron referenciadas al final de la investigación.
- La información recolectada es auténtica y veraz.
- Se cuidó la privacidad de las personas que forman parte de las imágenes recolectadas.
- Se practicó la preservación de confidencialidad.
- Se cuidó la revelación de información.
- Se aplicó el Código de Núremberg para preservar la ética profesional.

IV. RESULTADOS

4.1. Comparación de las métricas de evaluación de los modelos pre entrenados de Redes Neuronales Convolucionales.

4.1.1. Selección de modelos CNN pre entrenados

Para la etapa de selección de modelos, se realizó el estado del arte de la investigación a través de la revisión sistemática de la literatura, sobre temas relacionados a la tesis y que hayan utilizado algoritmos de IA con técnicas de *Transfer Learning (T.L.)* para clasificar y detectar acciones de violencia, de los cuales también se consideró sus hiperparámetros y métricas finales de rendimiento para el análisis comparativo, estos resultados se detallan en las Tablas 7 y 8:

Tabla 7

Revisión de otros autores que utilizan modelos de CNN pre entrenados

Autores de referencia	Modelos utilizados por el autor	Dataset de pre entrenamiento
Sakiba et al. (2023)	MobileNetV2	ImageNet
	YOLOv7	
Gaytán et al. (2022)	VGG16	ImageNet
	ResNet50	
	MobileNet	
Mehmood (2021)	InceptionV1	ImageNet + Kinetics
Patel (2021)	InceptionV3	ImageNet
	ResNet50	
Haque et al. (2021)	InceptionV3	ImageNet
	NASNetMobile	
	VGG19	
	MobileNet	
Traoré & Akhloufi (2020)	VGG16	INRIA Person
Soliman et al. (2019)	VGG16	ImageNet
Laureano (2019)	InceptionV3	COCO
	MobileNet	
	YOLOv2	

Tabla 8

Comparación de métricas de rendimiento entre los modelos de CNN propuestos por los autores

Autores de referencia	Modelos propuestos por el autor	Dataset train	Epoch train	Hiperparámetros del modelo						Métricas de rendimiento del modelo			
				Input size	Loss Func.	Learning rate	Batch size	Func. Act.	Opt.	Precisión (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Sakiba et al. (2023)	ConvLSTM	RLVS	20	64 x 64	Cat. cross entropy	0.001	16	ReLU	SGD	91.00	91.50	91.00	91.00
	YOLOv7	Img. de Google	500	224 x 224	-	-	-	-	-	96.40	71.30	81.97	mAP@0.5 75.90
Gaytán et al. (2022)	VGG16	Propio	50	-	-	-	-	-	-	79.92	79.28	79.24	79.33
	ResNet50			-	-	-	-	-	-	89.44	88.67	88.61	88.67
	MobileNet			-	-	-	-	-	-	90.55	89.34	89.25	89.33
Mehmood (2021)	2-Stream 3D-CNN	UFLV-DS	70	224 x 224	Cat. cross entropy	10^{-3} – 10^{-5}	-	-	Adam	97.00	97.00	97.00	98.00
Patel (2021)	CNN + LSTM	Hockey	50	-	Bin. cross entropy	0.0001	-	ReLU	RMSprop	-	-	-	89.50
		Movie V		-		0.00005	-			-	-	-	100.00
		V. Flow		-		0.00005	-			-	-	-	91.40
Haque et al. (2021)	MobileNet	Propio	100	-	-	-	-	-	-	-	-	-	95.41
Traoré & Akhloufi (2020)	2D BiGRU-CNN	Hockey	250	128 x 176	-	0.0008	10	Sigmoid	SGD	-	-	-	98.00
		RLVS		128 x 128	-					-	-	-	90.25
		V. Flow			-	0.0006				-	-	-	95.50
Soliman et al. (2019)	VGG16 + LSTM	Hockey	2000	224 x 224	Cat. cross entropy	0.06	100	Tanh	SGD	-	-	-	95.10
		Movie V								-	-	-	99.00
		V. Flow								-	-	-	90.01
Laureano (2019)	YOLOv2	YouTube y propio.	300	854 x 480 640 x 360	-	-	-	-	-	84.00	91.00	87.00	88.00

De acuerdo al análisis de los datos observados en las tablas anteriores, se estimó que los modelos CNN pre entrenados con técnicas de Transfer Learning a evaluar en esta investigación fueron: VGG16, MobileNet y YOLO por ser los más utilizados y presentar los mejores resultados de rendimiento, entonces, a continuación, se procedió a crear un dataset con imágenes de acciones violentas, para ser aplicadas con modelos escogidos y posteriormente comparar sus rendimientos, para seleccionar al que presente los mejores resultados.

4.1.2. Preparación de un dataset de prueba

En esta etapa de preparación y recopilación de los datos, se optó por crear un dataset propio con imágenes de acciones de violencia simuladas, realizando grabaciones de videos dentro de una zona urbana de la ciudad de Iquitos, ubicada exactamente en la Calle Pevas 2da cuadra, (más detalles se puede ver en el punto 4.2.2 de la tesis) de los cuales se escogió los frames relacionados a las clases estudiadas haciendo uso del software "VLC media player".

Al final, el total de frames capturados y seleccionados fue de dos mil (2000) imágenes con dimensiones de 1920 x 1088 pixeles, los cuales se clasificaron en cuatro (4) tipos o clases de acciones de violencia, basándose en movimientos de: "estrangulación", "forcejeo", "patada" y "trompada" cuya distribución de las cantidades y sus carpetas de ubicación, se detallan en la Tabla 9 y la Figura 19:

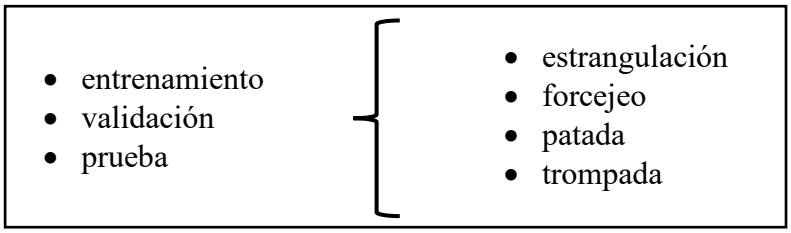
Tabla 9

Distribución de las cantidades de imágenes seleccionadas para el dataset

Clases	Entrenamiento 60%	Validación 30%	Prueba 10%	Total
Estrangulación	300	150	50	500
Forcejeo	300	150	50	500
Patada	300	150	50	500
Trompada	300	150	50	500
TOTAL	**1200**	**600**	**200**	**2000**

Figura 19

Distribución de las carpetas de clasificación para el dataset



Nota: Cada carpeta de la izquierda llevan dentro las cuatro sub carpetas de la derecha.

Es preciso indicar que, la distribución anterior de carpetas fue utilizada para el entrenamiento, validación y prueba de los modelos VGG16 y MobileNetV2 en vista que esta configuración es la más adecuada para ambos modelos y puedan realizar el proceso de clasificación. Algunas de las imágenes de muestra del dataset se observan en la Figura 20.

Figura 20

Ejemplo de las imágenes seleccionadas para el dataset

Asimismo, se estima que la cantidad de imágenes generadas fueron lo suficiente para poder experimentar, es decir, este número de imágenes se encuentra dentro del rango permitido para poder desarrollar una buena inferencia durante el entrenamiento y validación de los modelos mencionados, sin embargo, también se aplicó la técnica de Data Augmentation (ver código de la Figura 21) para obtener derivaciones de las mismas imágenes y así evitar problemas de sobreajuste u overfitting con respecto a los dos primeros modelos.

Ahora, para realizar las pruebas con YOLOv8 se preparó imágenes de entrenamiento y validación a través de la técnica de etiquetado con "Bounding Box" teniendo en cuenta las clases antes mencionadas, para lo cual se utilizó la herramienta de LabelImg, tal como se puede apreciar en la Figura 22, con este modelo no se aplicó Data Augmentation.

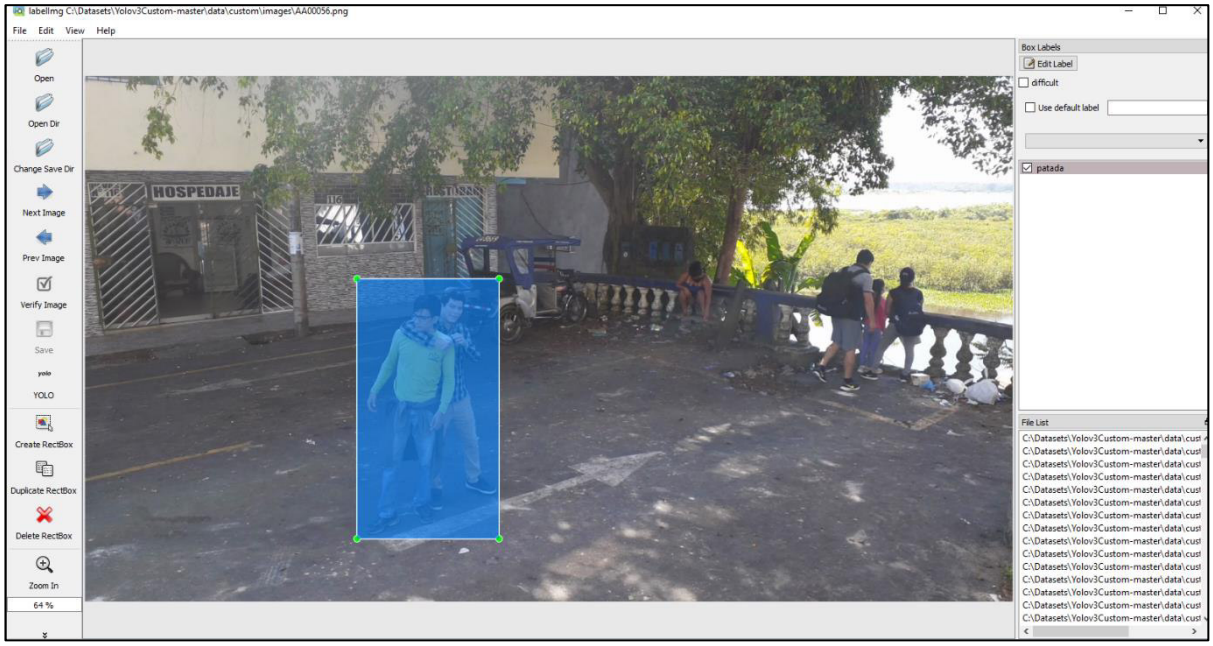
Figura 21

Código Data Augmentation para las imágenes seleccionadas

```
train_datagen = ImageDataGenerator(
    rotation_range=20,
    zoom_range=0.2,
    width_shift_range=0.1,
    height_shift_range=0.1,
    horizontal_flip=True,
    vertical_flip=False,
    preprocessing_function=preprocess_input)

valid_datagen = ImageDataGenerator(
    rotation_range=20,
    zoom_range=0.2,
    width_shift_range=0.1,
    height_shift_range=0.1,
    horizontal_flip=True,
    vertical_flip=False,
    preprocessing_function=preprocess_input)

train_generator = train_datagen.flow_from_directory(
    train_data_dir,
    target_size=(width_shape, height_shape),
    batch_size=batch_size,
    class_mode='categorical')

validation_generator = valid_datagen.flow_from_directory(
    validation_data_dir,
    target_size=(width_shape, height_shape),
    batch_size=batch_size,
    class_mode='categorical')
```

Figura 22

LabelImg para realizar el etiquetado de las imágenes utilizadas con YOLOv8



4.1.3. Evaluación de los modelos seleccionados utilizando el dataset creado

Los modelos VGG16 y MobileNetV2 fueron pre entrenados utilizando el dataset de ImageNet que contiene hasta mil (1000) tipos de clases, de los cuales se obtuvieron sus pesos originales para aplicar Transfer Learning, con la finalidad de no volver a entrenar todo desde cero, asimismo, para el entrenamiento se redimensionó las imágenes de entrada a 224 x 244 pixeles y se configuró los hiperparámetros de sus últimas capas para obtener una salida de solo cuatro (4) clases de acciones de violencia física; dicha configuración se pueden observar en los códigos de las Figuras 23 y 24. Por otro lado, para el caso de YOLOv8 se utilizó directamente el peso de "yolov8x" disponible en la página de https://github.com/ultralytics/ultralytics y se aplicó en el código de entrenamiento, como se puede observar en la Figura 26.

Los modelos seleccionados fueron evaluados utilizando el dataset desarrollado, los mismos que se entrenaron de manera local en una Laptop con tarjeta gráfica o GPU Nvidia modelo GeForce GTX 1660 Ti de 8 Gb de memoria, asimismo, a través del software Anaconda y Jupyter Notebook cada modelo fue sometido a ciento cincuenta (150) épocas de iteración; al

finalizar cada proceso se evaluaron los datos obtenidos y los gráficos de entrenamiento y validación, tal como se pueden observar en las Figuras 27, 28 y 29 donde también se consideró el tiempo que se tomó cada algoritmo para procesar la información.

Figura 23

Código Transfer Learning para el modelo VGG16

```
from keras.applications.vgg16 import VGG16
image_input = Input(shape=(224, 224, 3))

model = VGG16(input_tensor=image_input, include_top=True, weights='imagenet')

model.summary()

last_layer = model.get_layer('block5_pool').output
x=Flatten(name='flatten')(last_layer)
x= Dense(128,activation='relu', name='fc1')(x)
x= Dense(128,activation='relu', name='fc2')(x)
out = Dense(num_classes, activation='softmax', name='output')(x)
custom_model = Model(image_input, out)
custom_model.summary()

for layer in custom_model.layers[:-3]:
    layer.trainable = False

custom_model.summary()
 custom_model.compile(loss='categorical_crossentropy',optimizer='adadelta',metrics=['ac
                                        curacy'])
```

Figura 24

Código Transfer Learning para el modelo MobileNetV2

```
from keras.applications.mobilenet import MobileNet
image_input = Input(shape=(224, 224, 3))

model = MobileNet(input_tensor=image_input, include_top=False, weights='imagenet' )

model.summary()

last_layer = model.layers[-1].output
x=Flatten(name='flatten')(last_layer)
x= Dense(128,activation='relu', name='fc1')(x)
x=Dropout(0.3)(x)
x= Dense(128,activation='relu', name='fc2')(x)
x=Dropout(0.3)(x)
out = Dense(num_classes, activation='softmax', name='output')(x)
custom_model = Model(image_input, out)

custom_model.summary()
 custom_model.compile(loss='categorical_crossentropy',optimizer='adadelta',metrics=['ac
                                        curacy'])
```

En la Figura 25, las mismas líneas de código permiten ejecutar los entrenamientos para los modelos VGG16 y MobileNetV2 a través de “custom_model”, es decir, al compilar este código se envía a entrenar al modelo correspondiente, dependiendo con cual se esté trabajando.

Figura 25

Código para ejecutar el entrenamiento de los modelos VGG16 y MobileNetV2

```
tiempo_inicial=time()
model_history = custom_model.fit_generator(
    train_generator,
    epochs=epochs,
    validation_data=validation_generator,
    steps_per_epoch=nb_train_samples//batch_size,
    validation_steps=nb_validation_samples//batch_size
    )

tiempo_final = time()
tiempo_ejecucion = tiempo_final - tiempo_inicial
print (tiempo_ejecucion)

classes = train_generator.class_indices
print(classes)
```

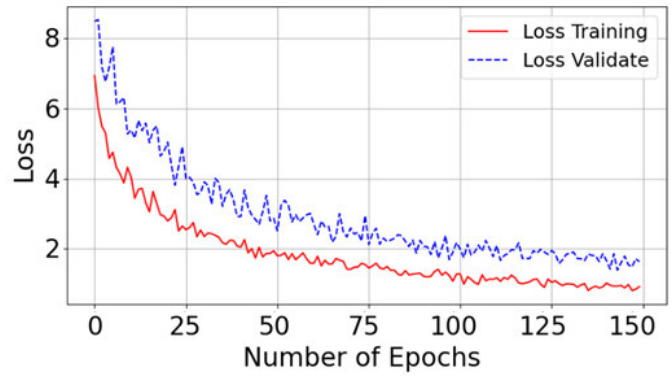
Figura 26

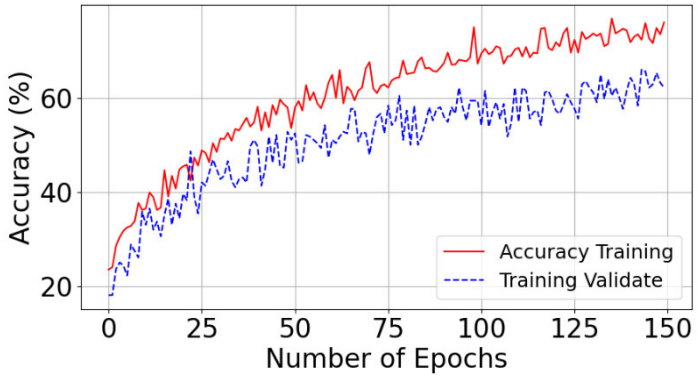
Código de entrenamiento para YOLOv8

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

Figura 27

Resultados del entrenamiento y validación – modelo VGG16

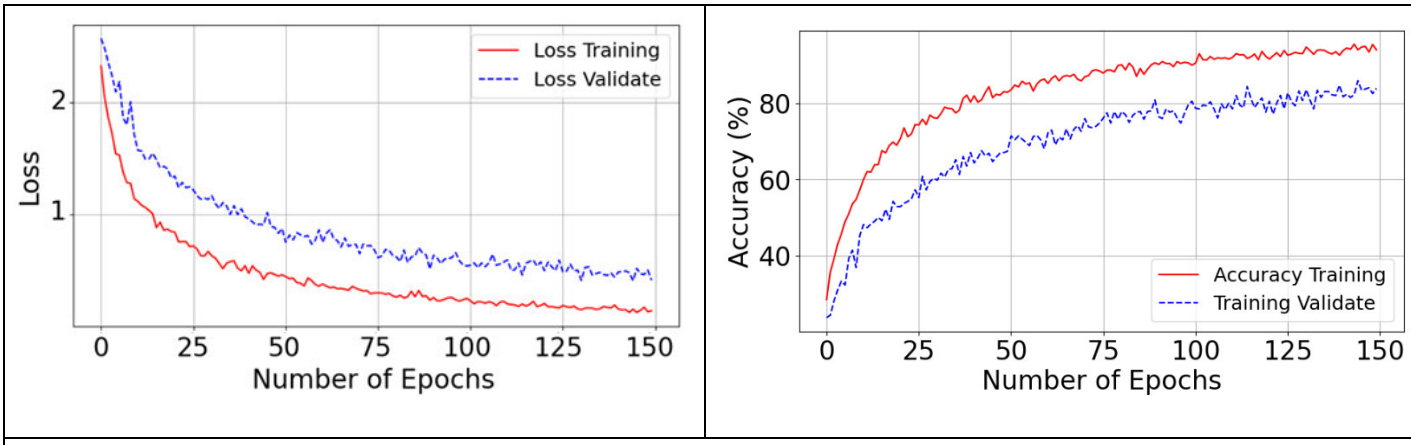




Tiempo = 4.5 h

Figura 28

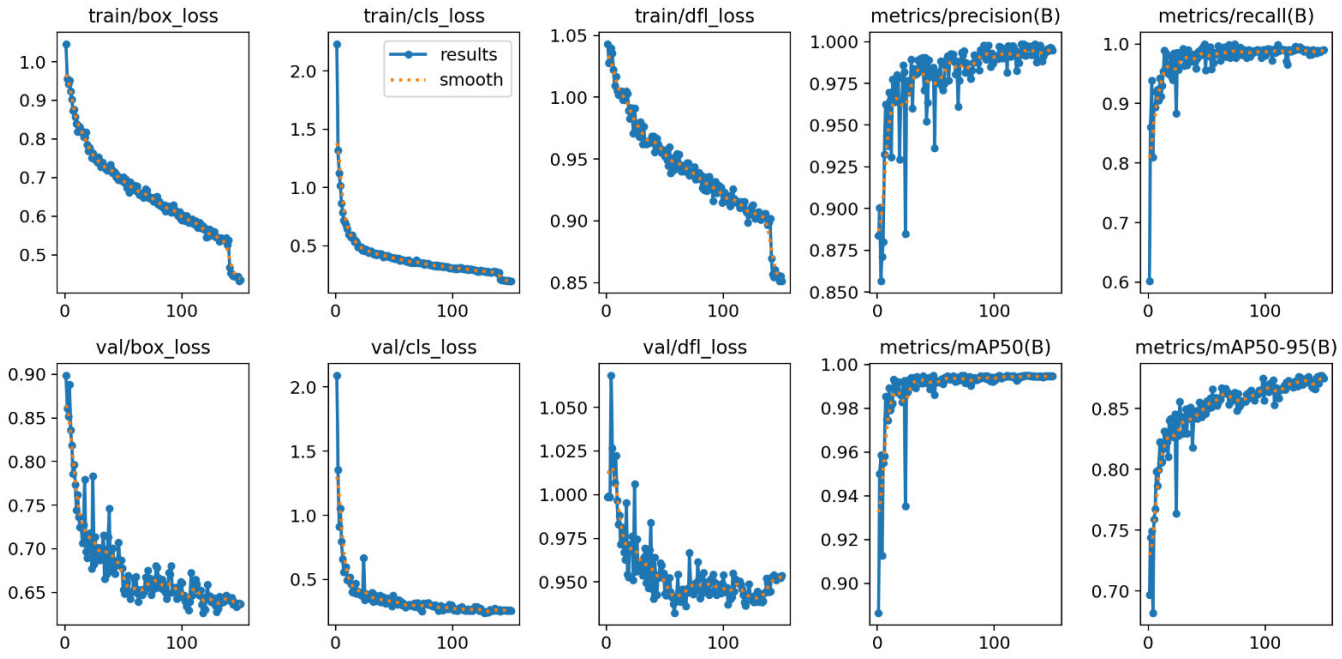
Resultados del entrenamiento y validación – modelo MobileNetV2



Tiempo = 6 h

Figura 29

Resultados del entrenamiento y validación – modelo YOLOv8



Tiempo = 2.8 h

4.1.4. Análisis comparativo de los modelos

Una vez obtenido los pesos (en formato ".h5" para VGG16 y MobileNet y formato ".pt" para YOLOv8) de los modelos entrenados, descritos en el paso anterior, se procedió a ejecutar el código de predicción para cada uno de ellos utilizando las imágenes de prueba (test) pre establecidas, tal como se detallan en los códigos de las imágenes de las Figuras 30, 31 y 32.

Figura 30

Código para ejecutar pruebas de predicción con el modelo VGG16

```
test_data_dir = 'dataset/test'

test_datagen = ImageDataGenerator()

test_generator = test_datagen.flow_from_directory(
    test_data_dir,
    target_size=(width_shape, height_shape),
    batch_size = batch_size,
    class_mode='categorical',
    shuffle=False)


custom_Model= load_model("model_VGG16_150_new.h5")

     predictions = custom_Model.predict_generator(generator=test_generator)
```

Figura 31

Código para ejecutar predicción con el modelo MobileNetV2

```
test_data_dir = 'dataset/test'

test_datagen = ImageDataGenerator()

test_generator = test_datagen.flow_from_directory(
    test_data_dir,
    target_size=(width_shape, height_shape),
    batch_size = batch_size,
    class_mode='categorical',
    shuffle=False)


custom_Model= load_model("model_MobileNet_150_new.h5")

  predictions = custom_Model.predict_generator(generator=test_generator)
```

Figura 32

Código para ejecutar predicción con el modelo YOLOv8

```
!yolo task=detect mode=predict model=/content/train/weights/best.pt conf=0.5
source={dataset.location}/test.yaml
```

Asimismo, para cada modelo se analizó los resultados de su matriz de confusión para obtener los puntajes de sus indicadores de desempeño (ver Figuras del 33 al 38), además se consideró como otro indicador importante al tiempo que se tomó cada modelo para llevar a cabo su entrenamiento, finalmente, el consolidado de estos resultados se detalla en la Tabla 10.

Figura 33

Matriz de confusión del modelo VGG16

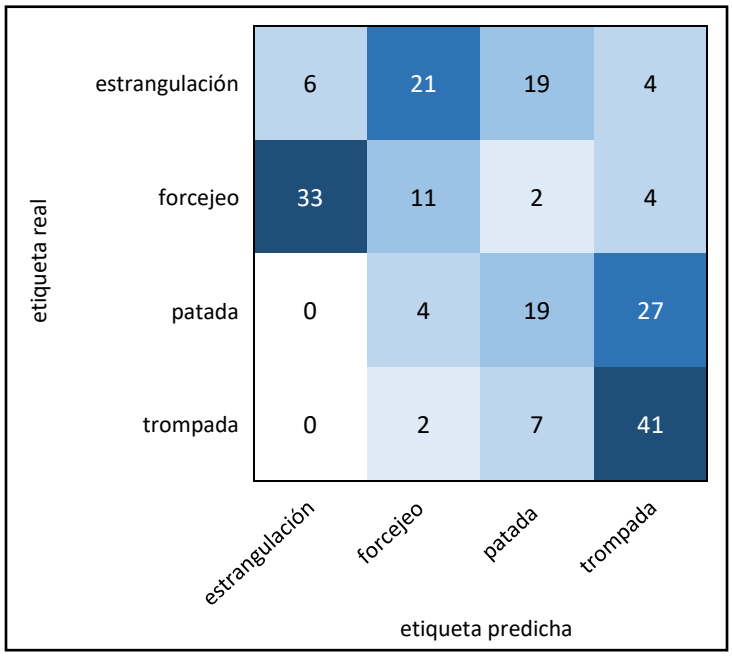


Figura 34

Resultados de los indicadores de desempeño del modelo VGG16

	precision	recall	f1-score	support
0	0.1538	0.1200	0.1348	50
1	0.2895	0.2200	0.2500	50
2	0.4043	0.3800	0.3918	50
3	0.5395	0.8200	0.6508	50
accuracy			0.3850	200
macro avg	0.3468	0.3850	0.3568	200
weighted avg	0.3468	0.3850	0.3568	200

Figura 35

Matriz de confusión del modelo MobileNetV2

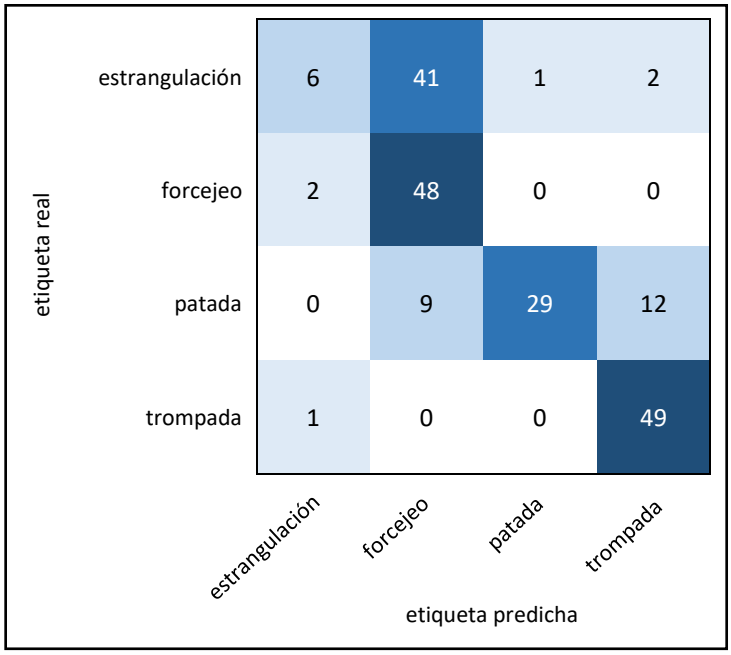


Figura 36

Resultados de los indicadores de desempeño del modelo MobileNetV2

	precision	recall	f1-score	support
0	0.6667	0.1200	0.2034	50
1	0.4898	0.9600	0.6486	50
2	0.9667	0.5800	0.7250	50
3	0.7778	0.9800	0.8673	50
accuracy			0.6600	200
macro avg	0.7252	0.6600	0.6111	200
weighted avg	0.7252	0.6600	0.6111	200

Figura 37

Matriz de confusión del modelo YOLOv8

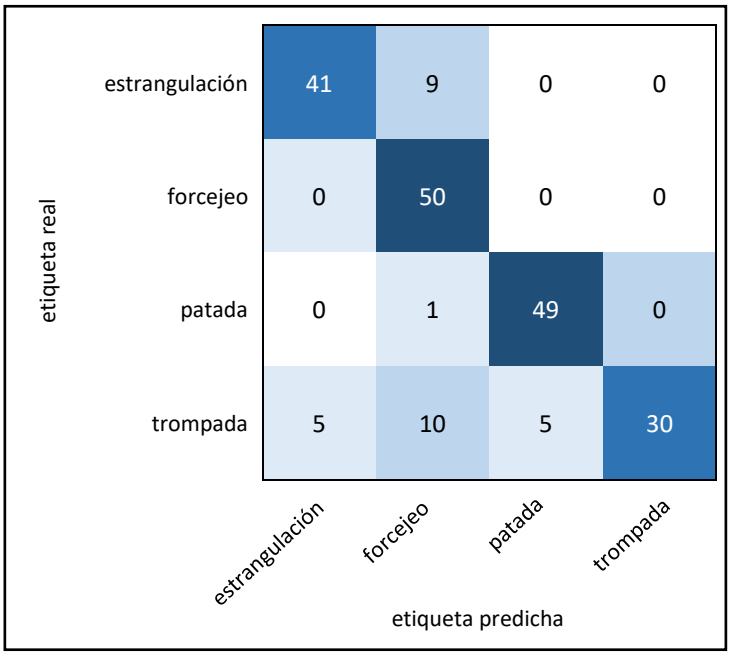


Figura 38

Resultados de los indicadores de desempeño del modelo YOLOv8

	precision	recall	f1-score	support
0	0.9317	0.7811	0.8498	50
1	0.9218	1	0.9593	50
2	0.9538	0.9935	0.9732	50
3	0.9628	0.7638	0.8518	50
accuracy			0.8900	200
macro avg	0.7725	0.9146	0.7685	200
weighted avg	0.9425	0.8900	0.9185	200

Tabla 10

Resultados de las métricas de rendimiento de los modelos seleccionados

Modelos	Clases	Métricas de rendimiento							
		Prec. (%)	Sens. (%)	Espe. (%)	F1-score (%)	G-mean (%)	IBA (%)	Acc.	Tiempo (horas)
VGG16	Estrangulación	15.38	12.00	68.26	13.48	28.62	7.73	**38.50**	**4.5**
	Forcejeo	28.95	22.00	70.97	25.00	39.51	14.85		
	Patada	40.43	38.00	67.44	39.18	50.62	24.88		
	Trompada	53.95	82.00	95.88	65.08	88.66	77.51		
	Promedio	**34.68**	**38.50**	**75.63**	**35.69**	**51.85**	**31.24**		
MobileNet V2	Estrangulación	66.67	12.00	97.67	20.34	34.23	10.71	**66.00**	**6.0**
	Forcejeo	48.98	96.00	62.69	64.86	78.19	63.15		
	Patada	96.67	58.00	99.04	72.50	75.79	55.07		
	Trompada	77.78	98.00	85.57	86.73	91.57	84.89		
	Promedio	**72.53**	**66.00**	**86.31**	**61.11**	**69.45**	**53.46**		
YOLOv8	Estrangulación	93.17	78.11	97.94	84.98	87.46	74.98	**89.00**	**2.8**
	Forcejeo	92.18	100.00	96.92	95.93	98.45	97.22		
	Patada	95.38	99.35	97.98	97.32	98.66	97.48		
	Trompada	96.28	76.38	99.06	85.18	86.98	73.95		
	Promedio	**94.25**	**88.46**	**94.75**	**90.85**	**92.89**	**85.91**		

Tabla 11

Resumen de los datos descriptivos de los modelos evaluados

Modelos pre entrenados de CNN	Media	Desv. Desviación	Desv. Error	Mínimo	Máximo
VGG16	44.59	16.78	6.85	31.24	75.63
MobileNetV2	68.14	11.14	4.54	53.43	86.31
YOLOv8	81.78	9.47	3.86	69.40	94.75

En los datos descriptivos de la Tabla 11, se observa que YOLOv8 tiene una media de 81.78 siendo la puntuación más alta en comparación con los otros dos, asimismo el puntaje promedio mínimo de uno de sus indicadores de desempeño fue de 69.40 y el máximo de 94.75, donde, el modelo VGG16 fue el que obtuvo la media más baja (44.59) entre los tres.

4.2. Evaluación del rendimiento del Sistema Web basado en Redes Neuronales Convolucionales

4.2.1. Integración del Sistema Web con YOLOv8

Para la etapa de desarrollo del sistema web e integración con el modelo de YOLOv8, se utilizó el lenguaje Python y las librerías de Tensorflow, Keras y OpenCV, los mismos que permiten crear aplicaciones de manera rápida y con un mínimo de código (ver Figuras 40 y 41); todo el proceso de desarrollo, así como los comandos de programación se puede observar dentro del Anexo E, en ello también se detalla el código de configuración para la correcta instalación del sistema. Por otro lado, en la Tabla 12 se muestran las recomendaciones técnicas para que el sistema web funcione de manera adecuada.

Tabla 12

Recomendaciones para el normal funcionamiento del sistema.

Hardware	Recomendaciones
- Cámara web con un mínimo de 12 megapíxeles. - PC o laptop con GPU NVIDIA de 8Gb de memoria mínimo. - Parlantes para el audio de la alarma.	- Captura de las imágenes con una inclinación de cámara de 15° hacia abajo. - Tener en cuenta que la cámara web este a una altura mínima de 2 metros y máxima de 3 metros. - La distancia máxima entre la cámara web y la zona de violencia para la detección es de 4 metros. - No exponer la cámara a condiciones de lluvia ya que esto dificulta el reconocimiento. - No realizar capturas en zonas con multitud de gente que obstaculizan el panorama de la cámara web. - Utilizar el sistema web en horas del día, y con bajo reflejo de la luz solar (sin mucho brillo). - Utilizar el navegador (browser) de Google Chrome.

Figura 39

Prototipo del esquema de implementación del Sistema Web basado en CNN

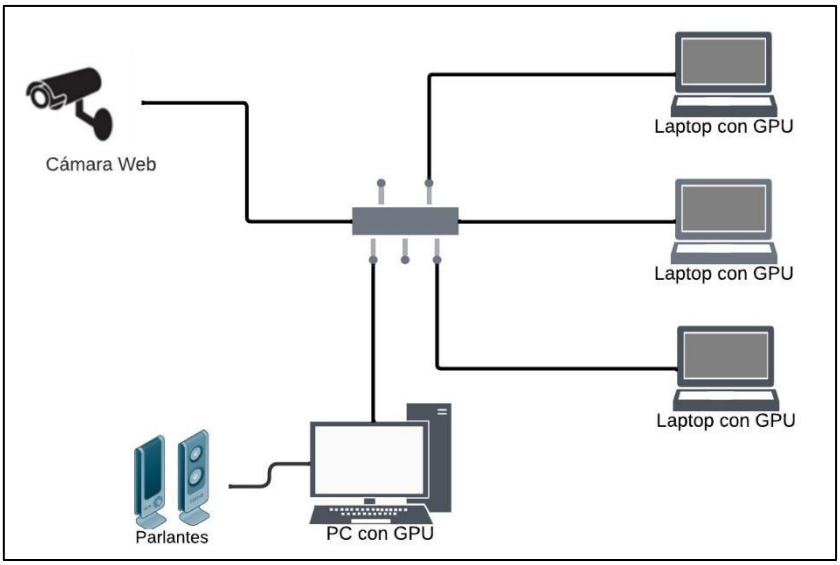


Figura 40

Página de inicio del Sistema Web basado en CNN

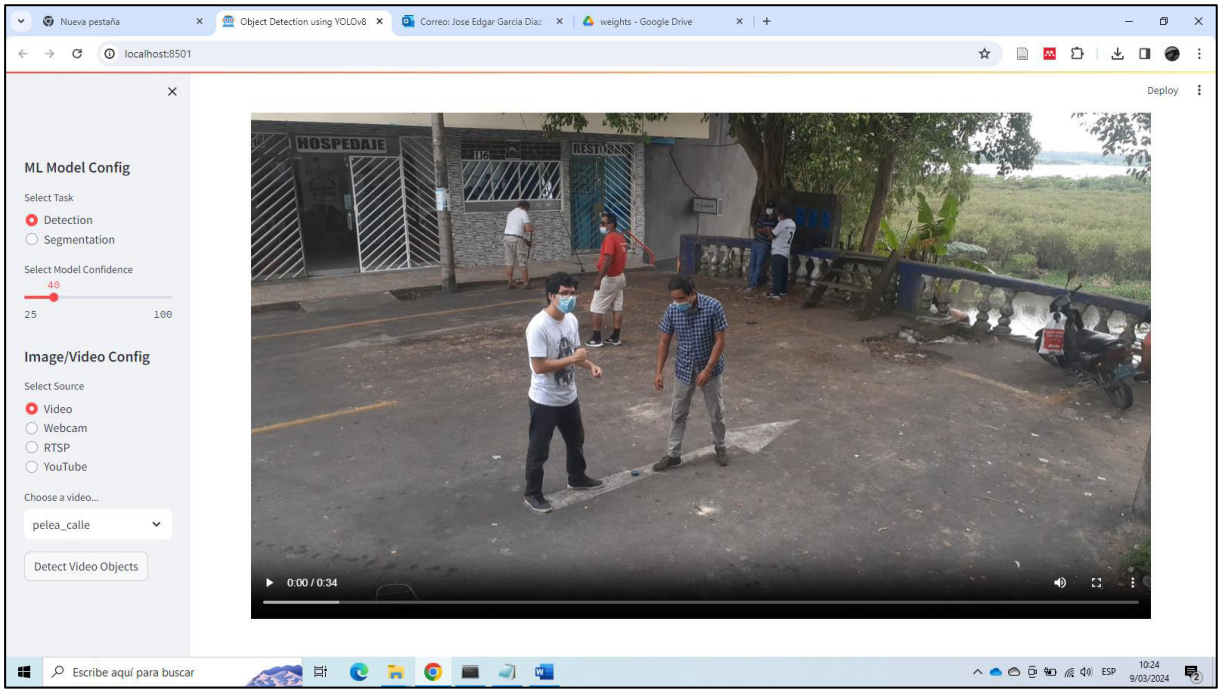
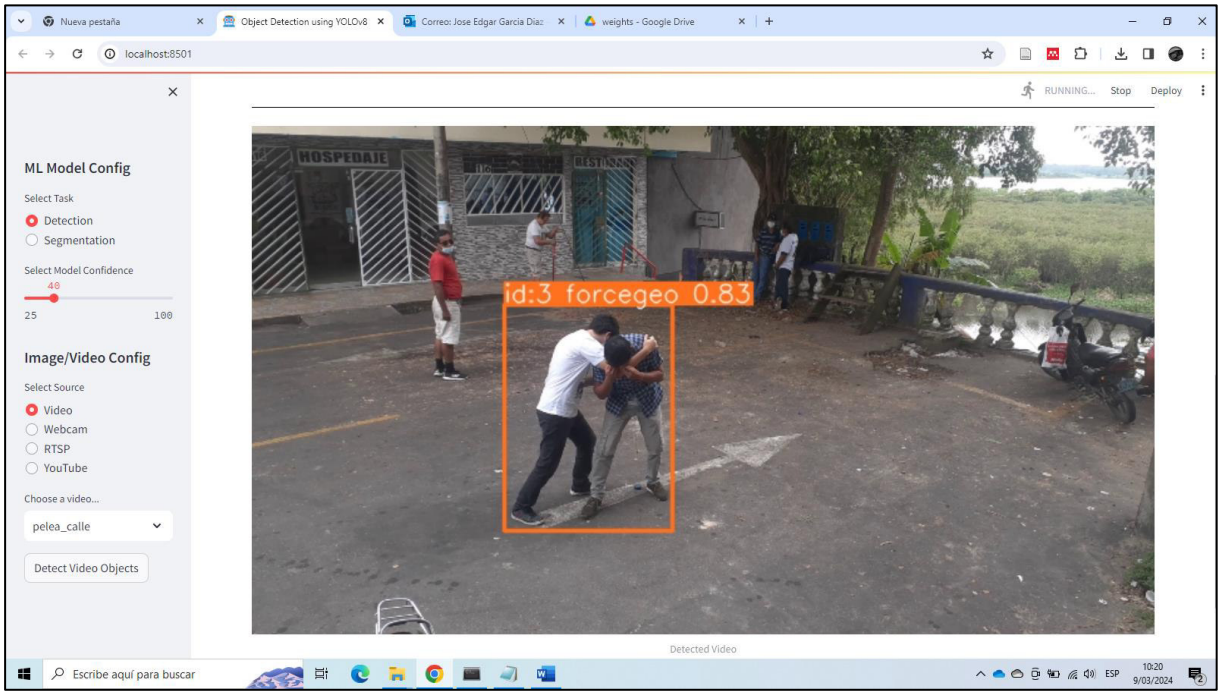


Figura 41

Prueba de funcionamiento del Sistema Web basado en CNN



4.2.2. Pruebas de campo del Sistema Web basado en CNN

Es importante indicar que las pruebas experimentales se realizó con el Sistema Web basado en el algoritmo de YOLOv8 y en el mismo escenario de la zona urbana de la ciudad de Iquitos donde se realizaron las grabaciones para obtener las imágenes del propio dataset, dicha zona se ubica exactamente en la Calle Pevas 2da cuadra en el distrito de Iquitos, provincia de Maynas, departamento de Loreto, cuya posición georreferencial tiene las siguientes coordenadas: 3° 44' 52.871" S, 73° 14' 31.319" W (ver Figuras 42 y 43), asimismo, es preciso mencionar que las pruebas de campo se basó en acciones violentas totalmente simuladas, realizadas por los alumnos del 3er nivel de la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional de la Amazonía Peruana – UNAP, que de manera voluntaria y con pleno consentimiento prestaron apoyo a esta investigación (ver Figuras 44 al 47).

Figura 42

Ubicación satelital de la zona urbana donde se realizó las pruebas de campo.

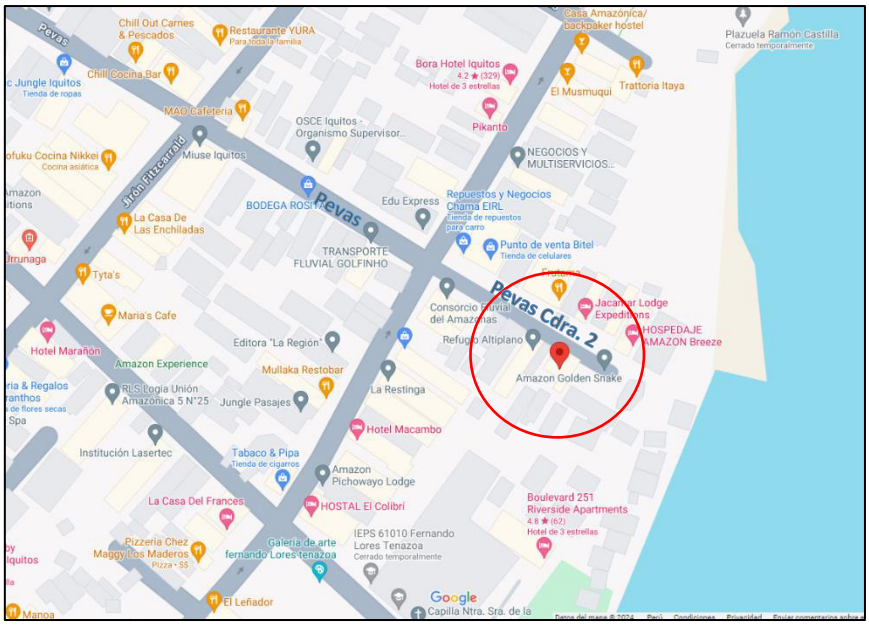


Figura 43

Escenario real de la zona urbana donde se realizó las pruebas de campo.

Figura 44

Prueba de detección de trompada (puñetazo), fue reconocida al 90%

Figura 45

Prueba de detección de forcejeo, fue reconocida al 75%

Figura 46

Prueba de detección de estrangulación, fue reconocida al 86%

Figura 47

Prueba de detección de patada, fue reconocida al 90%

4.2.3. Recolección de los datos de rendimiento del Sistema Web basado en CNN

Una vez realizada las nuevas pruebas de campo in situ, se recolectó nuevos datos de reconocimiento de acciones violentas, para revaluar el rendimiento y eficiencia del Sistema

Web basado en CNN en tiempo real, para lo cual nuevamente se utilizó la matriz de confusión y además se visualizó el comportamiento de precisión de las clases a través de la curva de Precisión – Recall (ver Figuras 48 y 49).

Figura 48

Matriz de confusión del Sistema Web basado en CNN (YOLOv8)

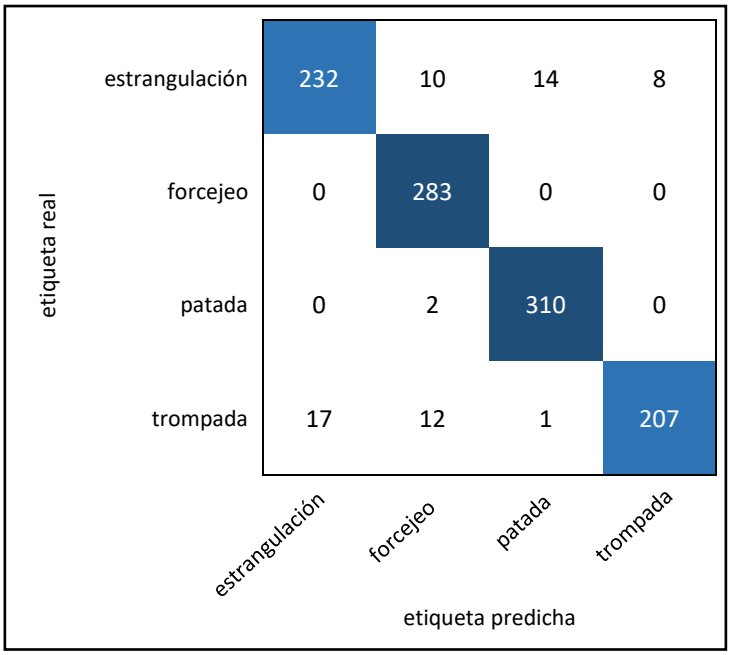
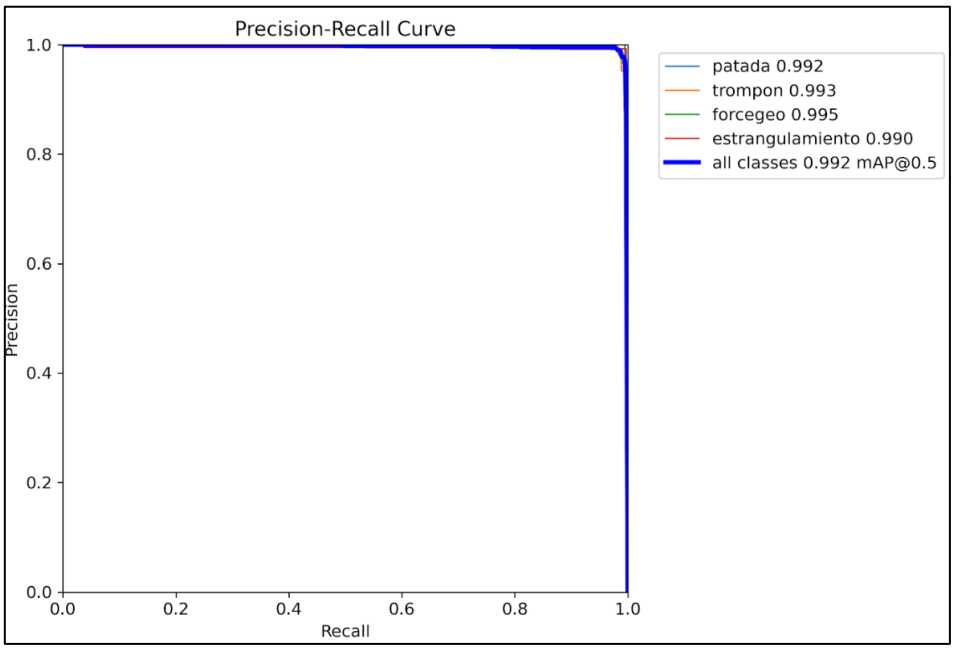


Figura 49

Curva de Precisión – Recall del Sistema Web basado en CNN (YOLOv8)



4.2.4. Análisis de resultados de reconocimiento del Sistema Web basado en CNN

Con los datos de la nueva matriz de confusión se pudo calcular el valor de los Accuracy (exactitud) para cada acción violenta, así como su comportamiento de detección a través de la curva de Precisión – Recall a una escala de mAP@0.5 (mean Average Precision) para comparar las clases estudiadas, tal como se detalla en la Tabla 13:

Tabla 13

Resultados de las pruebas de campo para el reconocimiento de las acciones de violencia

Clases	Accuracy	mAP@0.5
Estrangulación	0,8897	0,990
Forcejeo	0,9348	0,995
Patada	0,9297	0.992
Trompada	0,8851	0,993

En la Tabla 13 se puede apreciar que los datos relacionados a las acciones de "estrangulación" y "trompada" (puñetazo) fueron los menos acertados por el sistema, con niveles de 88.97% y 88.51% de Accuracy (exactitud) y niveles de mAP@0.5 de 86.2% y 84% respectivamente. Asimismo, de las pruebas realizadas, se ha evidenciado que la detección de acciones violentas con el sistema web tiene ciertas limitaciones para zonas con mucho brillo solar, con lluvia, con multitud de gente y para horarios nocturnos, además, se ha encontrado más falsos positivos en la detección de la clase "forcejeo" en vista que esta acción es detectada a pesar de no estar ocurriendo.

Además, se realizaron pruebas experimentales de contrastación, utilizando la Ficha de Observación N° 1 (ver Anexo B) para un total de treinta (30) procesos de reconocimiento simulados, de la cual se simuló 15 procesos con violencia y 15 procesos sin violencia, los resultados de estas pruebas se resume en la Tabla 14, estos resultados se demuestran de manera cruzada, y lo que se busca es contrastar el funcionamiento entre el Método Tradicional de Videovigilancia y el Sistema Web basado en CNN con la finalidad de obtener la mayor cantidad de coincidencias entre ellos (tomando como base al método tradicional) y llegar a un nivel

favorable o aceptable de similitud que permita determinar si “existe” o “no existe” la violencia física.

Tabla 14

Datos cruzados entre el Método Tradicional y el Sistema Web con CNN

		Método Tradicional				**Total**	
		Existe		No existe			
		N	**%**	**N**	**%**	**N**	**%**
Sistema Web con CNN	Existe	12	40	2	6.7	14	46.7
	No existe	3	10	13	43.3	16	53.3
	Total	15	50	15	50	30	100

Entonces, según se observa en la Tabla 14, se tiene que, de quince (15) escenarios donde “existe” violencia se evidenció que doce (12) escenas fueron identificadas o detectada correctamente por el sistema, y de otras quince (15) donde “no existe” violencia, trece (13) fueron identificadas o detectada correctamente por el sistema; eso significa que veinticinco (25) procesos fueron definidos correctamente de un total treinta (30), lo que equivale aproximadamente a un 83% de efectividad.

4.3. Validación del rendimiento del Sistema Web basado en Redes Neuronales Convolucionales con la finalidad de medir el tiempo de respuesta.

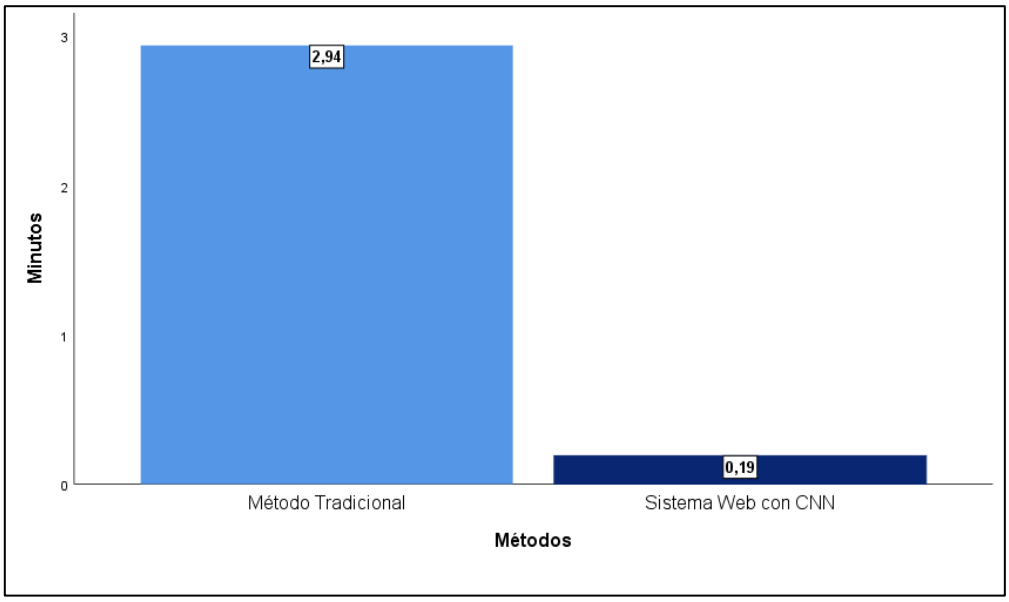
4.3.1. Pruebas de tiempo de respuesta del Sistema Web basado en CNN

En esta etapa se desarrolló la validación del tiempo promedio de respuesta del Sistema Web basado en CNN, teniendo en cuenta el intervalo de tiempo en minutos que se toma desde la detección de un incidente hasta el envío de una alerta sonora por los parlantes, el mismo que tiene como único criterio de activación que exista una secuencia continua de hasta diez (10) detecciones de acciones violentas de manera ininterrumpida.

Para eso, se desarrolló pruebas de campo con escenarios reales de violencia, donde se utilizó la Ficha de Observación N° 2 (ver Anexo B) la misma que consta de treinta (30) procesos de reconocimiento, y que al igual que en el anterior objetivo específico, se busca contrastar el Método Tradicional con el Sistema Web con CNN, pero en este caso con la finalidad de validar las diferencias entre los tiempos de respuestas, optando así por el más eficiente para alertar la violencia detectada, la comparación del tiempo promedio se muestra en la Figura 50.

Figura 50

Comparación del tiempo promedio de respuesta entre el Método Tradicional y el Sistema Web basado en CNN

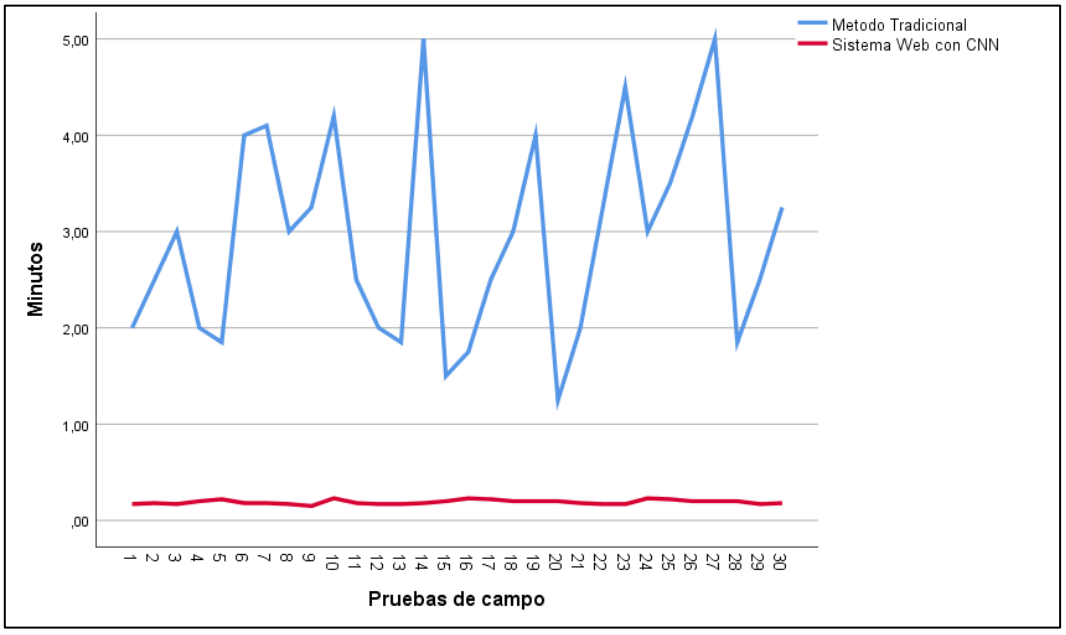


4.3.2. Análisis de la capacidad de respuesta del Sistema Web basado en CNN

Para el análisis de la capacidad de respuesta del sistema web, se realizó un seguimiento de los tiempos de respuesta por cada uno de los procesos de reconocimiento de violencia (treinta 30 procesos) entre el Método Tradicional y el Sistema Web basado en CNN, con la finalidad de evaluar las diferencias entre ambos; se puede observar en la Figura 51, que los tiempos de respuesta con el Método Tradicional son más lentos o demorados, y que pueden llegar a tardar hasta cinco (5) minutos en responder, mientras que los tiempos con el Sistema Web con CNN siempre se mantienen a menos de un minuto de respuesta; en la Figura 51 se observa mejor estas diferencias de tiempos por cada prueba realizada, además en la Tabla 15 se tienen los cálculos de los estadísticos descriptivos más importantes.

Figura 51

Comparación por procesos del tiempo de respuesta de reconocimiento



En la Tabla 15, se observa que la media del tiempo de respuesta con el Método Tradicional es de 2.94 minutos lo que equivale aproximadamente a 2 minutos con 56 segundos, mientras que la media del Sistema Web basado en CNN es de 0.19 minutos lo que equivale aproximadamente a once (11) segundos, además, se puede observar que la moda o tiempo más

frecuente del Método Tradicional es de dos (2) minutos, mientras que la moda del Sistema Web con CNN es de 0.17 minutos lo que equivale aproximadamente a diez (10) segundos. Asimismo, el tiempo mínimo y máximo para el Método Tradicional es de 1.25 y 5 minutos respectivamente, mientras el mínimo y máximo del Sistema Web basado en CNN es de 0.15 y 0.23 minutos respectivamente, a simple vista se deduce una gran diferencia, sin embargo, más adelante esto fue comprobado estadísticamente.

Tabla 15

Estadísticos descriptivos del tiempo de respuesta de reconocimiento

	Método Tradicional	**Sistema Web con CNN**
	Tiempo en min.	Tiempo en min.
Media	2.94	0.19
Mediana	3.00	0.18
Moda	2,00	0,17
Desv. Desviación	1.06	0.02
Mínimo	1.25	0.15
Máximo	5.00	0.23

4.3.3. Pruebas del Sistema Web basado en CNN, bajo diferentes condiciones

Las pruebas de funcionamiento in situ del Sistema Web basado en CNN, también se realizó bajo otras condiciones, como, por ejemplo, climáticas y en horarios nocturnos, específicamente se consideró los días de lluvia, con mucho brillo solar y con poco alumbrado público, con la finalidad de evaluar el nivel de captación de la cámara web ya que en estos escenarios se podría perjudicar de manera directa o indirecta al buen desempeño del sistema, tal como se observan en las Figuras 52 y 53.

Figura 52

Imagen de video de un escenario con mucho brillo solar

Figura 53

Imagen de video de escenario nocturno y con poco alumbrado público

De la evaluación realizada, se estima que el Sistema Web basado en CNN presenta dificultades de reconocimiento y baja su nivel de desempeño, siendo recomendable que su funcionamiento se realice en condiciones climáticas y de tiempo adecuados, es decir durante los días sin lluvia, en horas de la mañana y tarde con suficiente acceso a la luz natural.

4.3.4. Retroalimentación de usuarios potenciales

Para determinar el despliegue eficaz del sistema web, es necesario también considerar el feedback de los usuarios potenciales o expertos, con la finalidad de potenciar las bondades del sistema a través de sus comentarios o recomendaciones, de manera que se podría tomar en cuenta y estudiar su factibilidad de implementación. Entonces, para ello se realizó una entrevista (ver Anexo C) a los responsables del Centro de Operaciones, Emergencia y Monitoreo (COEM) para la seguridad ciudadana de Iquitos, creada el 2016 por la Municipalidad Provincial de Maynas y que se ubica dentro de la Base de Serenazgo, la cual cuenta con 60 cámaras de videovigilancia instaladas en sitios estratégicos de la zona urbana de la ciudad.

Figura 54

Base del COEM - Maynas

El COEM – Maynas trabaja en coordinación con la Fiscalía, Policía Nacional del Perú División de Prevención e Investigación de Robo de Vehículos (DIPROVE), Departamento de Investigación Criminal (DEPINCRI) y las distintas comisarías de Iquitos, contribuyendo con los casos e investigaciones de las imágenes captadas por las cámaras.

4.4. Contrastación de hipótesis

4.4.1. Hipótesis específica 1

H_0: La comparación de las métricas de evaluación de los modelos pre entrenados de Redes Neuronales Convolucionales no permite obtener diferencias significativas en la selección del modelo con mejor rendimiento en el reconocimiento de la violencia física.

H_1: La comparación de las métricas de evaluación de los modelos pre entrenados de Redes Neuronales Convolucionales permite obtener diferencias significativas en la selección del modelo con mejor rendimiento en el reconocimiento de la violencia física.

Para contrastar la primera hipótesis específica, se realizó una comparación de medias de las métricas de evaluación de los modelos pre entrenados seleccionados, a través del estadístico de ANOVA con un factor a un nivel de significancia del 5% ($\alpha = 0.05$), donde al final se obtuvo un p – valor = 0.001 tal como se observa en la Tabla 16.

Entonces, con una probabilidad de error de 0.1% se establece que: La comparación entre las métricas de los modelos pre entrenados permite obtener diferencias significativas para la selección del modelo con mejor rendimiento en el reconocimiento de la violencia física.

Tabla 16

Resumen de la prueba de ANOVA con un factor

	Suma de cuadrados	**gl**	**Media cuadrática**	**F**	**Sig.**
Entre grupos	4245,604	2	2122,802	12,849	**0.001**
Dentro de grupos	2477,998	15	165,199		
Total	6723,602	17			

Asimismo, en la Tabla 17 se muestra los resultados de la prueba post hoc de Tukey a través de comparaciones múltiples, donde se observa un p – valor = 0.191 entre los modelos de YOLOv8 y MobileNet, la misma que es mayor a 0.05, con lo cual estadísticamente se establece que no necesariamente exista una diferencia significativa entre estos dos modelos. Es decir, que el modelo MobileNet también podría ser tomado en cuenta, sin embargo, se trabajó con YOLOv8 porque es un modelo para detección de objetos a través del etiquetado de imágenes y se encuentra en constante evolución tecnológica, siendo éste un factor importante a tener en cuenta, ya que permitirá seguir obteniendo mejores versiones para el sistema en el futuro.

Tabla 17

Comparaciones múltiples de los modelos seleccionados

Modelos pre entrenados de CNN		**Diferencia de medias**	**Desv. Error**	**Sig. (p-valor)**
VGG	MobileNet	-23,545*	7,420	0.016
	YOLO	-37,181*	7,420	0.000
MobileNet	VGG	23,545*	7,420	0.016
	YOLO	-13,636	7,420	**0.191**
YOLO	VGG	37,181*	7,420	0.000
	MobileNet	13,636	7,420	**0.191**

***** La diferencia de medias es significativa en el nivel 0.05.

4.4.2. Hipótesis específica 2

H_0: Con la evaluación del rendimiento del Sistema Web basado en Redes Neuronales Convolucionales no se determina una concordancia significativa con el método tradicional para detectar la existencia de la violencia física dentro de una zona urbana.

H_1: Con la evaluación del rendimiento del Sistema Web basado en Redes Neuronales Convolucionales se determina una concordancia significativa con el método tradicional para detectar la existencia de la violencia física dentro de una zona urbana.

Para contrastar la segunda hipótesis específica se realizó una prueba de asociación a través del Índice de Kappa de Cohen, con un nivel de significancia del 5% ($\alpha = 0.05$), donde al final se observa en la Tabla 18 el valor de kappa = 0.667, con este valor según de Ullibarri Galparsoro (1999) se puede determinar que existe una "buena concordancia" entre el Sistema Web con CNN y el Método Tradicional.

Entonces, estadísticamente se comprueba que: Con la evaluación del rendimiento del Sistema web con CNN se determina una concordancia significativa con el Método Tradicional para detectar la violencia física dentro de una zona urbana.

Tabla 18

Resumen de la prueba del Índice de Kappa de Cohen

		Valor	Error estándar asintótico	T aproximada	Significación aproximada
Medida de acuerdo	Kappa	**0.667**	0.136	3.660	0.000253
N de casos válidos		30			

4.4.3. Hipótesis específica 3

H_0: Con la validación del funcionamiento del Sistema Web basado en Redes Neuronales Convolucionales no se estima una diferencia significativa del tiempo de respuesta para alertar la violencia física.

H_1: Con la validación del funcionamiento del Sistema Web basado en Redes Neuronales Convolucionales se estima una diferencia significativa del tiempo de respuesta para alertar la violencia física.

Para esta tercera prueba de hipótesis, primero se realizó el cálculo de normalidad de los datos, a través de la prueba de Kolmogorov – Smirnov, cuyos resultados obtenidos del software de SPSS se presentan en la Tabla 19, donde se puede observar un p – valor mayor a 0,05 para ambos métodos, con lo cual se determina que si existe normalidad de los datos.

Tabla 19

Resultados de la prueba de normalidad de Kolmogorov – Smirnov

Tipo de sistema	Estadístico	gl	Sig.
Método Tradicional	0.147	30	**0.098**
Sistema Web con CNN	0.159	30	**0.051**

a. Corrección de significación de Lilliefors

Luego, se procedió a utilizar la prueba paramétrica de t – Student para muestras independientes, con un nivel de significancia del 5% ($\alpha = 0.05$), donde al final se obtuvo un p – valor = 3,841E-13 (valor bastante bajo) tal como se observa en la Tabla 20, con la cual se rechaza la hipótesis nula y se acepta la hipótesis alternativa; finalmente, se establece

estadísticamente que: Con la validación del Sistema Web basado en CNN se estima una diferencia significativa de los tiempos de respuestas para alertar la violencia física.

Tabla 20

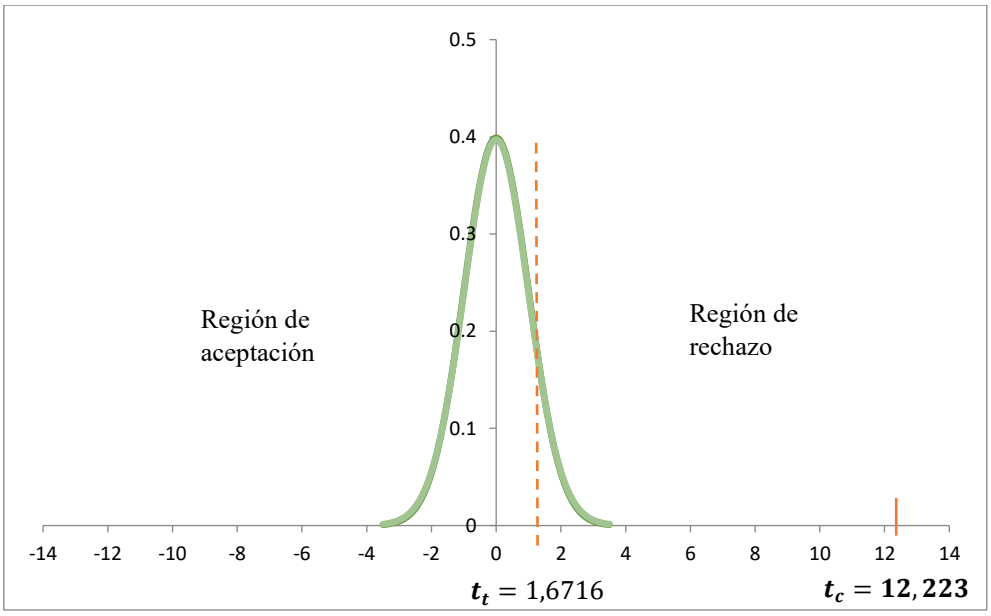
Resumen de la prueba t – Student para muestras independientes

t_c	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	95% de intervalo de confianza de la diferencia	
					Inferior	Superior
12.223	58	**3,841E-13**	2.38000	0.19471	1.98225	2.77775

Asimismo, en la Figura 55 se observa que el valor de t calculado (t_c) supera al t tabulado (t_t), el mismo tiempo que se encuentra en la zona de rechazo para la hipótesis nula.

Figura 55

Campana de Gauss de t - Student



4.4.4. Hipótesis general:

H_0: El desarrollo de un Sistema Web basado en Redes Neuronales Convolucionales no optimiza significativamente el reconocimiento de la violencia física en zonas urbanas.

H_1: El desarrollo de un Sistema Web basado en Redes Neuronales Convolucionales optimiza significativamente el reconocimiento de la violencia física en zonas urbanas.

Para contrasta la hipótesis general, se tomó en consideración todo el tiempo que demora el sistema desde la etapa de reconocimiento para determinar si existe o no existe una acción violenta hasta el momento en que se emite la alarma de alertar a los usuarios sobre dicha acción detectada, para lo cual se utilizó la Ficha de Observación N° 3 (ver Anexo B) la misma que consta de treinta (30) pruebas o procesos de reconocimiento. Con los datos obtenidos se procedió a realizar el cálculo de normalidad, a través de la prueba de Kolmogorov – Smirnov, cuyos resultados según el software de SPSS se presentan en la Tabla 21, donde se puede observar un p – valor mayor a 0,05 para ambos métodos, con lo cual se determina que si existe normalidad de los datos.

Tabla 21

Resultados de la prueba de normalidad de Kolmogorov – Smirnov

Tipo de sistema	Estadístico	gl	Sig.
Método Tradicional	0.147	30	**0.098**
Sistema Web con CNN	0.156	30	**0.061**

a. Corrección de significación de Lilliefors

Al igual que en la hipótesis 3, se procedió a utilizar la prueba paramétrica de t – Student para muestras independientes, con un nivel de significancia del 5% ($\alpha = 0.05$), donde al final se obtuvo un p – valor = 1,846E-14 (valor bastante bajo) tal como se observa en la Tabla 22, con

la cual se rechaza la hipótesis nula y se acepta la hipótesis alternativa; finalmente, estadísticamente se concluye y afirma que: *El Sistema Web basado en CNN optimiza significativamente el reconocimiento de la violencia física en zonas urbanas.*

Tabla 22

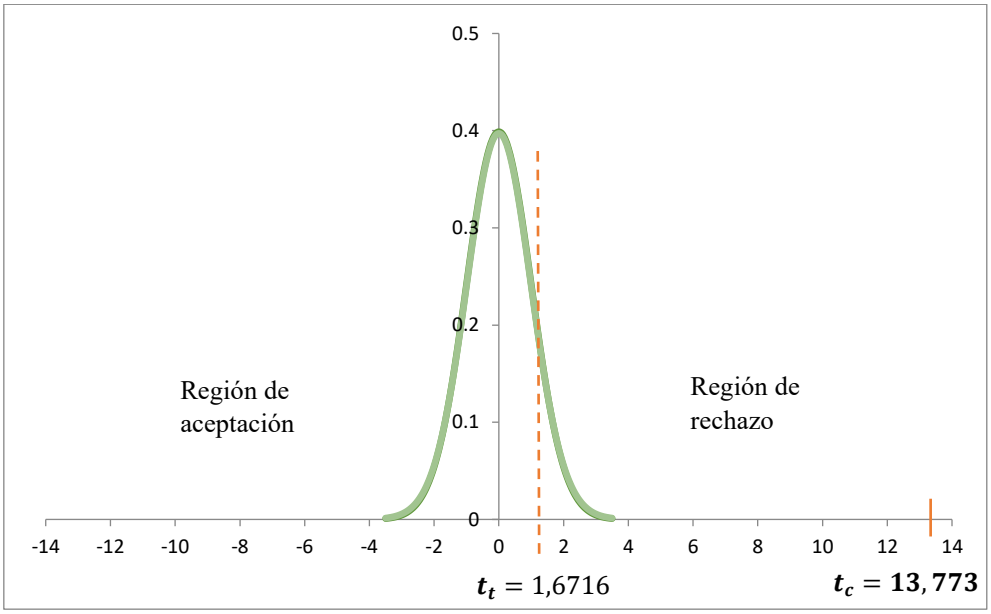
Resumen de la prueba t – Student para muestras independientes

t_c	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	95% de intervalo de confianza de la diferencia	
					Inferior	Superior
13,773	58	**1,846E-14**	2,68166	0,19469	2,28394	2,28394

Asimismo, en la Figura 56 también se observa que el valor de t calculado ($\boldsymbol{t_c}$) supera al t tabulado ($\boldsymbol{t_t}$), el mismo tiempo que se encuentra en la zona de rechazo para la hipótesis nula.

Figura 56

Campana de Gauss de t - Student



V. DISCUSIÓN DE RESULTADOS

De acuerdo a las etapas de esta investigación, se precisa que en la fase de preparación de datos se elaboró un dataset de manera personalizada con imágenes simuladas de violencia, donde se generó al menos un total de dos mil (2000) imágenes, las mismas están clasificadas en cuatro acciones o clases (patada, trompada, forcejeo y estrangulamiento), en contraste, Lumba et al. (2019) indica que para el desarrollo de su investigación se basó en mil (1000) imágenes, las cuales estuvieron conformadas por 500 imágenes obtenidas de los frames de videos de YouTube y otras 500 fueron creadas por los autores, asimismo, con respecto a las clases de violencia, Lumba solo trabajó con dos clases: patada y trompada. Entonces, se estima que la propuesta planteada supera la cantidad de imágenes en un 100%, así como las clases de acciones de violencia fue aumentada, incluyendo forcejeo y estrangulamiento; estas contribuciones de la propuesta incrementan las mejoras para el reconocimiento de la violencia física. Al final, Lumba también indica que, para la fase de entrenamiento de detección de la violencia, utilizaron o compararon tres CNN's con Transfer Learning: MobileNet, InceptionV3 y YOLOv2 obteniendo una mejor precisión con InceptionV3 de 89%; sin embargo, en esta investigación también se realizaron pruebas con tres CNN's que son: VGG16, MobileNet y YOLOv8, obteniéndose la mejor precisión con YOLOv8 de 94.25%, con lo cual se concluye que esta investigación supera a las métricas obtenidas por el referido autor.

Baba et al. (2019), en su estudio presenta una propopuesta de reconocimiento de acciones de violencia a través de la aplicación de su propio modelo de DNN (Deep Neural Networks), donde, para el entrenamiento utilizó dos datasets de videos de violencia de libre disponibilidad: BEHAVE y ARENA, a los cuales Baba los trabajó con la finalidad de obtener dos tipos de inputs (entradas); para los primeros tipos de inputs, Baba consideró las imágenes originales ("sin modificar") de los datasets y para los segundos tipos de inputs, aplicó un

algoritmo de flujo optico, denominado Farneback, que consiste en rescalar los colores de las imágenes y con lo cual se esperaba lograr una detección más rápida de los objetos; sin embargo, Baba menciona que su modelo de DNN consiguió mejores resultados utilizando los primeros tipos de inputs, obteniendo valores de TPR (True Positive Result) y FPR (False Positive Result) de 100% y 26.76% respectivamente y un accuracy de 85.88%; en contraste, en la presente investigación se trabajó con las imágenes originales y sin manipular sus colores, es decir, no se aplicó algoritmos de preprocesamiento antes de entrenar los modelos; donde, con el uso de YOLOv8 se obtuvo un mejor accuracy de 89%; en ese sentido se coincide con Baba al evidenciar que las imágenes "sin modificar" también ayudan a conseguir buenos resultados en las métricas de rendimiento.

Existe una investigación cualitativa exploratoria muy interesante, desarrollada en Esan por Becerra et al. (2019) que trata de abordar los factores determinantes para adoptar la tecnología de la Inteligencia Artificial (IA) en la problemática de la seguridad de los eventos deportivos, donde se busca identificar características contextuales y específicas que debe tener la IA para su aplicación y tratar de prevenir la violencia en el futbol, siguiendo las disposiciones emitidas por la CONMEBOL; asimismo, Becerra busca identificar estratégias para su aplicabilidad en los clubes del futbol peruano y estima que la IA se puede aplicar a través de un sistema inteligente de circuito cerrado de TV para identificar las conductas inadecuadas de los asistentes antes, durante y después del evento, y tener una reacción más rápida de los agentes de seguridad, registrando los hechos e identificar y prohibir el ingreso de personas violentas en los eventos deportivos. Entonces, al comparar con Becerra, en la presente investigación se tiene como un producto adicional, la implementación en un Sistema Web basado en CNN para reconocer las acciones de violencia física en zonas urbanas, el cual puede ser aprovechado como herramienta para detectar la violencia en los alrededores de los estadios, donde casi siempre

existen desmanes antes y después de un partido de futbol y de esta manera, se estaría contribuyento al uso de la IA para manejar una estrategia de seguridad, tal como lo recomientada Becerra en su estudio.

En la UNSA de la ciudad de Arequipa, Machaca (2019) desarrolló una investigación que propone la detección de llamados eventos anómales utilizando CNN a través del algoritmo de YOLOv3 y la técnica de Bounding Box para detectar objetos en la imagen, y así lograr capturar lo que Machaca llama la "persistencia de objetos", donde, básicamente obtiene los datos de velocidad y trayectoria de desplazamiento de las imágenes en el video; asimismo, es preciso indicar que Machaca considera como eventos anómalos a las imágenes con acciones violentas de peleas y asaltos, a los que va midiendo sus cambios repentinos de estados que sufre durante el video, debido a la velocidad con que se manifiestan, pudiendo pasar de un estado normal a otro de manera intempestiva y de esa manera Machaca establece la existencia de un evento anómalo. Al final, Machaca indica que, utilizando YOLOv3 alcanzó una exactitud del 86% y una precisión de 85% para clasificar dos categorías (fight y no – fight). Al respecto se hace énfasis a la manera como Machaca detecta sus acciones violentas, ya que tomó en cuenta la velocidad del cambio de estado de los objetos de la imagen, se puede decir que esto es otra forma de evaluar la existencia de acciones de violencia; sin embargo, en esta investigación se tomó como dato fundamental al porcentaje de precisión de la detección del objeto utilizando YOLOv8; que al final presenta una precisión de 94.25%. Asimismo, es preciso indicar que la investigación de Machaca solo se limita a tener la opción de reconocimiento de la violencia en videos guardados o grabados previamente, pero esta investigación es mejorada, ya que, además de contar con esa opción, también cuenta con la capacidad de reconocer la violencia en tiempo real con el uso de una cámara web.

En una investigación que fue desarrollado en la India por Imran et al. (2020) se propone una estrategia propia para el reconocimiento de acciones de violencia, la cual tiene una arquitectura en “tres fases” para intentar ser más eficiente, robusto y liviano, dichas fases son: la primera, consiste en dividir el video en fragmentos para generar lo que se llama un ADI (Aproximación Dinámica de Imágenes) al cual los autores lo consideran como una mejor técnica que el flujo óptico; la segunda, consiste es utilizar MobileNet para extraer las caracteríscas de cada ADI, y la tercera, es aplicar una Red Neuronal Recurrente (RNN) llamada GRU (Gated Recurret Unit) para reconocer las acciones del video y determinar la clasificación de fight o no fight; para los entrenamientos Imran utilizó los datasets públicos de Hockey Fight, Movies and Violent Flows, obteniendo acuraccy de 100%, 100% y 96.71% respectivamente al realizar pruebas con cada uno de ellos; sin embargo, se discrepa con Imran, puesto que, al realizar un recocimiento basado en “tres fases”, este proceso podría tornarse lento al momento de clasificar, ya que, de acuerdo a la revisión del estado del arte, existen otras técnicas similares de detección de objetos a “dos fases” como: Mask R-CNN y Faster R-CNN que utilizan una combinación de algoritmos pero que a la fecha no satisfacen las espectativas de los expertos para realizar detección en tiempo real; por otro lado, si bien es cierto que, MobileNet es un buen modelo para extraer características por ser un algoritmo más liviano, al mismo tiempo, el peso generado utiliza mucho espacio en disco, como es el caso de Imran, cuyo peso generado por MobileNet alcanza los 49.8 Mb, lo cual podría ir en contra de los objetivos, porque, esto es sinónimo de más tiempo de procesaminto, en contraste, en esta investigación se utilizó YOLOv8, que detecta los objetos en “una fase” o “un solo golpe”, y se generó un peso más liviano en un archivo que ocupa aproximadamente 6 Mb de espacio en disco, lo que garantiza procesar imágenes de manera rápida y en tiempo real.

Dentro del estado del arte, se pudo observar que existen distintas formas de evaluar el procesamiento de imágenes de violencia, algunos autores crean sus propios CNN, otros utilizan transfer learning, otros crean redes neuronales híbridas y otros utilizan CNN para detección de objetos, todos buscan obtener mejores resultados; como es el caso de Sakiba et al. (2023) de la universidad de BRAC en Bangladesh, quienes dentro de su tesis analizaron dos maneras de evaluar la violencia: la primera evaluación fue a las "posturas violentas" mediante una red híbrida, llamada ConvLSTM (donde la parte Conv es un derivado de MobileNet v2 y la parte LSTM corresponde a una red neuronal bi direccional recurrente) con la cual obtuvo un accuracy de 91% y la segunda, fue evaluar a los "individuos con objetos letales" utilizando YOLOv7 para detección de objetos con técnicas de bounding box, con la cual obtuvo un mAP@0.5 de 75.9%. En contraste, la presente investigación realizó algo similar, para evaluar las acciones violentas en imagenes completas, se utilizó MobileNet con la cual se alcanzó un accuracy de 66%, si bien es cierto es un valor bajo, pero, es preciso mencionar que no se realizó ningún híbrido de red; y para la detección de acciones violentas dentro de la imagen se usó YOLOv8 con técnicas de bounding box, con el cual se obtuvo una mAP@0.5 de 91% siendo un resultado mucho mejor que de Sakiba; entonces, se puede concluir que estas formas de evaluar se vienen realizando en otras investigaciones, tanto para la clasificación y detección de acciones violentas.

VI. CONCLUSIONES

- Se comparó tres modelos pre entrenados de CNN con Transfer Learning: VGG16, MobileNet y YOLOv8, a los cuales se realizó el entrenamiento, validación y prueba utilizando un dataset elaborado con acciones de violencia simulada, enfocadas en cuatro clases (patada, trompada, forcejeo y estrangulamiento) compuesto por dos mil (2000) imágenes y distribuidos en 70% para entrenamiento, 30% para validación y 10% para pruebas; cada modelo fue sometido a 150 épocas y al final, se llegó a la conclusión que con el modelo YOLOv8 se logró los mejores resultados para las métricas de rendimiento, con un accuracy global de 89%.
- Se evaluó al Sistema Web basado en CNN para determinar la existencia de la violencia física en una zona urbana en contraste con el método tradicional, teniendo en cuenta el nivel de concordancia o coincidencia entre ambos métodos, los cuales fueron sometidos a treinta (30) pruebas de manera independiente, y de acuerdo a las pruebas estadística de Kappa de Cohen se concluye que existe "buena concordancia" entre ambos métodos, entonces, de esta manera se establece que el Sistema Web basado en CNN si puede determinar coherentemente la existencia de la violencia física en zonas urbanas.
- Se validó al Sistema Web basado en CNN para medir el tiempo de respuesta y alertar la violencia física en una zona urbana, el mismo que también se realizó en contraste con el método tradicional, basado en el tiempo que se toman ambos métodos, a los cuales también se les sometió a treinta (30) pruebas de manera independiente; obteniéndose en promedio 2.94 y 0.19 minutos de tiempo de respuestas respectivamente y que de acuerdo a la prueba estadística de t – Student se establece que existe una diferencia significativa entre ambos, entonces, de esta manera se concluye que el Sistema Web basado en CNN reduce el tiempo de respuesta para alertar y garantizar una acción rápida frente a un escenario violento en una zona urbana.

- Se desarrolló el Sistema Web basado en Redes Neuronales Convolucionales para optimizar el reconocimiento de la violencia física en zonas urbanas, en términos generales el Sistema Web propuesto basado CNN optimiza el reconocimiento de la violencia física en zonas urbanas, esta afirmación se sustenta en las pruebas realizadas durante la evaluación y validación del sistema, tal como se observa en las conclusiones anteriores que sostienen resultados eficientes, objetivos y por la contrastación de las hipótesis de estudio.

VII.RECOMENDACIONES

- Probar con otros modelos pre entrenados de CNN con Transfer Learning a fin de evaluar y mejorar los resultados de las métricas, asimismo, se puede utilizar más de un dataset con más de dos mil (2000) imágenes de violencia, incorporando más clasificaciones de acciones violentas.
- Evaluar la posibilidad de desarrollar un sistema que también detecte la violencia en zonas rurales, considerando el uso de otros métodos de evaluación que optimicen la investigación, con la finalidad de obtener mejores resultados de acuerdo a su adaptabilidad o escenario real.
- Implementar más opciones de envío de alerta del sistema, que podría darse a través de mensajes de texto o correo electrónico con las imágenes del acto violento detectado, asimismo, se podría desarrollar un hardware como una alarma de gran capacidad sonora, con la finalidad mejorar los tiempos de respuestas en la zona urbana y/o tener una reacción más rápida de los involucrados del orden público (serenazgo, junta vecinal, guardianes, etc.).
- Como trabajo futuro se recomienda el desarrollo de una app o aplicativo móvil, que utilice una API de conexión a servidores HPC que permitan realizar el reconocimiento de la violencia en tiempo real a través de la cámara del celular, y teniendo esas evidencias se pueda alertar a los organismos de seguridad desde cualquier lugar.

VIII. REFERENCIAS

Achkoski, J., Trajkovik, V., y Serafimova, N. (2012). A concept for a Smart Web Portal Development in Intelligence Information System Based on SOA [conferencia]. *The 9th Conference for Informatics and Information Technology (CIIT 2012),* Bitola, Macedonia. https://www.researchgate.net/profile/Jugoslav-Achkoski/publication/236886399_A_Concept_for_a_Smart_Web_Portal_Development_in_Intelligence_Information_System_Based_on_SOA/links/0deec519e9ac37ce7b000000/A-Concept-for-a-Smart-Web-Portal-Development-in-Intelligence-Information-System-Based-on-SOA.pdf

Aezion (2018). *The Benefits of Web-Based Systems for Business*. https://www.aezion.com/blogs/the-benefits-of-web-based-systems-for-business/

Akosa, J. S. (2017). Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data. *In Proceedings of the SAS global forum (Vol. 12, pp. 1-4).* Cary, NC, USA: SAS Institute Inc. https://support.sas.com/resources/papers/proceedings17/0942-2017.pdf

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*. Cornell University. https://arxiv.org/abs/1803.01164v2

APD (2019). *¿Cuáles son los tipos de algoritmos del machine learning?* https://www.apd.es/algoritmos-del-machine-learning/

Baba, M., Gui, V., y Pescaru, D. (2019). Deep Learning Approach for Violence Detection in Urban Areas. *ITM Web of Conferences*, *1st International Conference on Computational Methods and Applications in Engineering (ICCMAE 2018)*, Timisoara, Rumania, *29*, 03009. https://doi.org/10.1051/itmconf/20192903009

Becerra Robles, D. E., Godoy Alcarraz, J. D. D., Salazar Llontop, N., & Vallejos Fonseca, F. J. (2019). *Determinantes de la adopción de la inteligencia artificial en la prevención de violencia en eventos deportivos masivos de fútbol en la ciudad de Lima*. [Tesis de Maestría, Universidad ESAN. Escuela de Administración de Negocios para Graduados]. Repositorio Institucional Universidad ESAN. https://hdl.handle.net/20.500.12640/1519

Boden, M. (Ed.). (2016). *Inteligencia Artificial*. Editorial Turner Noema.

Centro de Investigación en Política Pública (2019). *La delincuencia causa más muertos que los conflictos armados: Estudio Global de Homicidios 2019 vía ONU*. https://imco.org.mx/la-delincuencia-causa-mas-muertos-que-los-conflictos-armados-estudio-global-de-homicidios-2019-via-onu/

Comisión Económica para América Latina y el Caribe [CEPAL] (2013). Comisión Económica para América Latina y el Caribe. *Definición de población Urbana y Rural utilizadas en los censos de los países Latinoamericanos.* https://www.cepal.org/sites/default/files/def_urbana_rural.pdf

De Luca, A., y Irigoitia, M. E. (2021). *Análisis de la técnica Transfer Learning en Machine Learning a través de un caso de estudio: La clasificación de productos en el Banco Alimentario de La Plata*. [Tesis de Pregrado, Universidad Nacional de la Plata, Argentina]. Repositorio Institucional de la UNLP. http://sedici.unlp.edu.ar/handle/10915/117398

El Peruano (2015). *Ley para prevenir, sancionar y erradicar la violencia contra las mujeres y los integrantes del grupo familiar - LEY - N° 30364*. Congreso de la República. https://www.gob.pe/74905-ley-n-30364-ley-para-prevenir-sancionar-y-erradicar-la-violencia-contra-las-mujeres-y-los-integrantes-del-grupo-familiar

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, *1*(2), 119–130. https://doi.org/10.1016/0893-6080(88)90014-7

GeeksforGeeks. (11 de junio de 2025). *Web 4.0 - Intelligent Web*. https://www.geeksforgeeks.org/web-4-0-intelligent-web/

González, A. (2023). *¿Qué es Machine Learning?* https://cleverdata.io/que-es-machine-learning-big-data/

Grupo Iberdrola (2023). *¿Qué es la Inteligencia Artificial? ¿Somos conscientes de los retos y principales aplicaciones de la Inteligencia Artificial?* https://www.iberdrola.com/innovacion/que-es-inteligencia-artificial

IAT. (2023). *Machine learning. Tipos, modelos, técnicas y usos*. https://iat.es/tecnologias/inteligencia-artificial/machine-learning/

Imran, J., Raman, B., Singh Rajput, A., y Layer ADI, F. (2020). Robust, Efficient and Privacy-Preserving Violent Activity Recognition in Videos. *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, Brno, República Checa, *8*(20). https://doi.org/10.1145/3341105

INCORE. (2023). *Retrato Regional*. https://incoreperu.pe/portal/index.php/retrato-regional

Instituto de Estudios Peruanos (2018). *VI Ronda del Barómetro de las Américas en el Perú: informe, presentación y vídeos*.

https://iep.org.pe/noticias/iep-presenta-barometro-de-las-americas-2016-2017/

Ketkar, N. y Moolayil, J. (2021). Deep learning with python: Learn Best Practices of Deep Learning Models with PyTorch. In *Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch*. Apress Media LLC. https://doi.org/10.1007/978-1-4842-5364-9

Kienle, H. M. y Distante, D. (2014). *Evolution of web systems. Evolving Software Systems*, 201–228. In: Mens, T., Serebrenik, A., Cleve, A. (eds) Evolving Software Systems. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-45398-4_7

Konasani, V. R., y Kadre, S. (2021). Machine learning and deep learning using python and tensorflow. McGraw-Hill Education. https://www.accessengineeringlibrary.com/content/book/9781260462296

Krizhevsky, A., Sutskever, I., y Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems,* Toronto, Canada. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Latinoamérica21 (2021). *Desigualdad. La violencia interminable en América Latina*. https://latinoamerica21.com/es/la-violencia-interminable-en-america-latina/

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. *2019 XLV Latin American Computing Conference (CLEI)*, Panamá, Panamá, 1–9. https://doi.org/10.1109/CLEI47609.2019.235100

Machaca, L. A. (2019). *Reconocimiento de eventos anómalos en videos obtenidos de cámaras de vigilancia, usando redes convolucionales*. [Tesis de Pregrado, Universidad Nacional de San Agustin de Arequipa]. Repositorio institucional UNSA. http://repositorio.unsa.edu.pe/handle/UNSA/10849

Mayorga, L. C., Riccardi, G. A., Bermeo, O. X. y Guevara V. I. (2022). Sistema Web para los procesos administrativos y de producción en viveros del Cantón Milagro. *Ingeniería y sus Alcances, Revista de Investigación*, *Universidad Agraria del Ecuador, Milagro, Ecuador*, *6*(16), 200-2013. https://repositorio.cidecuador.org/bitstream/123456789/2262/1/Articulo_1_Ingenieria_y_sus_Alcances__N16V6.pdf

Odicio, E. (1992). *Perfil Demográfico De La Region Loreto*. *Documento Técnico N° 01 – Junio 1992.* Instituto de Investigaciones de La Amazonia Peruana – IIAP. http://www.iiap.org.pe/upload/Publicacion/ST001.pdf

Organización Panamericana de la Salud [OPS]. (2021). *Prevención de la violencia - OPS/OMS*. https://www.paho.org/es/temas/prevencion-violencia

Pande, M., Neuman, R., y Cavanagh, R. (Ed.). (2004). *Las claves prácticas de Six Sigma*. Editorial Mc Graw Hill.

Peñafiel, S. (2022). *Entrenamiento del modelo YOLO para detección de una placa vehicular previamente capturada en imagen o video y aplicación de OCR para obtención de sus caracteres*. [Tesis de Pregrado, Pontificia Universidad Católica del Ecuador]. Repositorio Institucional Universidad PUCE. https://repositorio.puce.edu.ec/handle/123456789/27440

Pérez, M. (2023). *Concepto Definición ¿Qué es la Violencia? - Su Definición y Características*. https://conceptodefinicion.de/violencia/

Programa de las Naciones Unidas para el Desarrollo [PNUD]. (2020). *Análisis sobre innovación en seguridad ciudadana y derechos humanos en América Latina y el Caribe. Una perspectiva desde las políticas públicas y la gestión institucional*. https://www.undp.org/sites/g/files/zskgke326/files/migration/latinamerica/undp-rblac-es-Analisis-innovacion-seguridad-ciudadana-derechos-humanos-VF.pdf

Raffino, Equipo editorial, Etecé (28 de noviembre de 2023). *Zona urbana y zona rural*. Enciclopedia Concepto. https://concepto.de/zona-urbana-y-zona-rural/

Redmon, J., Divvala, S., Girshick, R., y Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* Las Vegas, USA, 779–788. https://doi.org/10.1109/CVPR.2016.91

Rettberg, A. (2020). Violencia en América Latina hoy: manifestaciones e impactos. *Revista de Estudios Sociales. Universidad de los Andes, Colombia, 1*(73), 2–17. https://doi.org/10.7440/RES73.2020.01

Ruiz Sarmiento, J. R., Monroy, J., Moreno, F.-A., & González-Jiménez, J. (2020). Tutorial para el reconocimiento de objetos basado en características empleando herramientas Python. *Actas de las XXXIX Jornadas de Automática*, Badajoz, España. https://doi.org/10.17979/SPUDC.9788497497565.0998

Sharma, N. (2023). *What is MobileNetV2? Features, Architecture, Application and More*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2023/12/what-is-mobilenetv2/

Simonyan, K., y Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. Cornell University, Nueva York, EE.UU. https://arxiv.org/abs/1409.1556v6

Sitiobigdata. (2019). *Tres principales enfoques de los modelos de aprendizaje automático.* https://sitiobigdata.com/2019/12/24/3-principales-enfoques-de-los-modelos-de-aprendizaje-automatico/#

Suárez, J. E. (2020). *Arquitectura de detección de actividades criminales basada en análisis de vídeo en tiempo real - Dialnet*. [Tesis de Doctorado, Universidad Politécnica de Valencia]. Repositorio Institucional Universidad UPV. https://riunet.upv.es/entities/publication/f51eaf0e-69a4-4d83-a817-384e824fdf04

Torres, A. (4 de setiembre de 2016). *Los 11 tipos de violencia (y las distintas clases de agresión)*. Psicología y mente. https://psicologiaymente.com/forense/tipos-de-violencia

Vasilev, I., Spacagna, G., Slater, D., Peter, R., y Valentino, Z. (2019). *Python Deep Learning* (2° ed.). Editorial Packt.

IX. ANEXOS

Anexo A

Matriz de Consistencia

Título: SISTEMA WEB BASADO EN REDES NEURONALES CONVOLUCIONALES PARA RECONOCER LA VIOLENCIA FÍSICA EN ZONAS URBANAS

Problema Principal	Objetivo General	Hipótesis General	Variables	Indicadores	Método
¿En qué medida el desarrollo de un Sistema Web basado en Redes Neuronales Convolucionales optimiza el reconocimiento de la violencia física en zonas urbanas? *PE1:* ¿Cómo comparar las métricas de evaluación de los modelos pre entrenados de Redes Neuronales Convolucionales para seleccionar el modelo con el mejor rendimiento en el reconocimiento de la violencia física? *PE2:* ¿De qué manera evaluar el rendimiento del Sistema Web basado en Redes Neuronales Convolucionales para determinar la existencia de la violencia física dentro de una zona urbana? *PE3:* ¿Cómo validar el rendimiento del Sistema Web basado en Redes Neuronales Convolucionales con la finalidad de medir el tiempo de respuesta para alertar la violencia física?	Desarrollar un Sistema Web basado en Redes Neuronales Convolucionales para optimizar el reconocimiento de la violencia física en zonas urbanas. *OE1:* Comparar las métricas de evaluación de los modelos pre entrenados de Redes Neuronales Convolucionales para seleccionar el modelo con mejor rendimiento en el reconocimiento de la violencia física. *OE2:* Evaluar el rendimiento del Sistema Web basado en Redes Neuronales Convolucionales para determinar la existencia de la violencia física dentro de una zona urbana. *OE3:* Validar el rendimiento del Sistema Web basado en Redes Neuronales Convolucionales con la finalidad de medir el tiempo de respuesta para alertar la violencia física.	El desarrollo de un Sistema Web basado en Redes Neuronales Convolucionales optimiza significativamente el reconocimiento de la violencia física en zonas urbanas. *HE1:* La comparación de las métricas de evaluación de los modelos pre entrenados de Redes Neuronales Convolucionales permite obtener diferencias significativas en la selección del modelo con mejor rendimiento en el reconocimiento de la violencia física. *HE2:* Con la evaluación del rendimiento del Sistema Web basado en Redes Neuronales Convolucionales se determina una concordancia significativa con el método tradicional para detectar la existencia de la violencia física dentro de una zona urbana. *HE3:* Con la validación del funcionamiento del Sistema Web basado en Redes Neuronales Convolucionales se estima una diferencia significativa del tiempo de respuesta para alertar la violencia física.	**Variable Independiente:** Sistema Web basado en Redes Neuronales Convolucionales. **Variable Dependiente:** Reconocimiento de la violencia física.	- Rendimiento del Sistema Web basado en CNN - Métricas de evaluación de los modelos pre entrenados de CNN. - Reconocer la violencia física. - Alertar oportunamente la violencia física.	**Tipo de Investigación** El tipo de investigación es aplicada. **Nivel de Investigación** La Investigación es del nivel descriptivo y predictivo. **Diseño de Investigación** El diseño es cuasiexperimental con post – test y grupo de control. **Unidad muestral:** Proceso de reconocimiento de la violencia física en espacios urbanos de las ciudades. A nivel mundial. **Universo:** Todos los procesos de reconocimiento de la violencia en espacios urbanos de las ciudades con cámaras de videovigilancia y equipo servidor de alta gama (workstation) a nivel mundial. N = indeterminado **Muestra:** Proceso de Reconocimiento de la Violencia en espacios urbanos de la ciudad de Iquitos. n = 30 **Tipo de muestro:** Aleatorio.

Anexo B

Fichas de Observación

Ficha de Observación N° 1 **Evaluación del reconocimiento de la existencia de violencia física** **(Método Tradicional vs Sistema Web basado en CNN)**
Indicación: Marque con un aspa (X) en el casillero del ítem que corresponda, según indique la respuesta del experto (método tradicional) y del sistema web basado en CNN.

Procesos a evaluar		**Método Tradicional**		**Sistema Web basado en CNN**	
N°	Escenarios de violencia física	Si existe	No existe	Si existe	No existe
1	Prueba				
2	Prueba				
3	Prueba				
4	Prueba				
5	Prueba				
6	Prueba				
7	Prueba				
8	Prueba				
9	Prueba				
10	Prueba				
11	Prueba				
12	Prueba				
13	Prueba				
14	Prueba				
15	Prueba				
16	Prueba				
17	Prueba				
18	Prueba				
19	Prueba				
20	Prueba				
21	Prueba				
22	Prueba				
23	Prueba				
24	Prueba				
25	Prueba				
26	Prueba				
27	Prueba				
28	Prueba				
29	Prueba				
30	Prueba				

Ficha de Observación N° 2
Evaluación del tiempo de respuesta para alertar la violencia física
(Método Tradicional vs Sistema Web basado en CNN)

Indicación: Apunte el tiempo transcurrido en el casillero del ítem que corresponda, según indique la respuesta del experto (método tradicional) y del sistema web basado en CNN.

Procesos a evaluar		**Método Tradicional**	**Sistema Web basado en CNN**
N°	Escenarios de violencia física	Tiempo en min.	Tiempo en min.
1	Prueba		
2	Prueba		
3	Prueba		
4	Prueba		
5	Prueba		
6	Prueba		
7	Prueba		
8	Prueba		
9	Prueba		
10	Prueba		
11	Prueba		
12	Prueba		
13	Prueba		
14	Prueba		
15	Prueba		
16	Prueba		
17	Prueba		
18	Prueba		
19	Prueba		
20	Prueba		
21	Prueba		
22	Prueba		
23	Prueba		
24	Prueba		
25	Prueba		
26	Prueba		
27	Prueba		
28	Prueba		
29	Prueba		
30	Prueba		
Promedio en minutos			

Ficha de Observación N° 3
Evaluación del tiempo desde el reconocimiento hasta la alerta de la violencia física (Método Tradicional vs Sistema Web basado en CNN)

Indicación: Apunte el tiempo transcurrido en el casillero del ítem que corresponda, según indique la respuesta del experto (método tradicional) y del sistema web basado en CNN.

Procesos a evaluar		**Método Tradicional**	**Sistema Web basado en CNN**
N°	Escenarios de violencia física	Tiempo en min.	Tiempo en min.
1	Prueba		
2	Prueba		
3	Prueba		
4	Prueba		
5	Prueba		
6	Prueba		
7	Prueba		
8	Prueba		
9	Prueba		
10	Prueba		
11	Prueba		
12	Prueba		
13	Prueba		
14	Prueba		
15	Prueba		
16	Prueba		
17	Prueba		
18	Prueba		
19	Prueba		
20	Prueba		
21	Prueba		
22	Prueba		
23	Prueba		
24	Prueba		
25	Prueba		
26	Prueba		
27	Prueba		
28	Prueba		
29	Prueba		
30	Prueba		
	Promedio en minutos		

Anexo C

Preguntas de la Entrevista a los expertos del COEM

1. ¿El COEM cuenta con Sistema de Videovigilancia basado en Inteligencia Artificial (IA)?

   ……

2. ¿Qué opinas sobre el hecho de poder contar con IA para combatir la delincuencia?

   ……

3. ¿Crees que el aplicativo que te presentamos podría ayudar en algo o hasta qué nivel llegaría?

   ……

4. ¿Si Ud. estaría dispuesto a aplicar nuestro sistema web que condición nos pondría?

   ……

5. ¿Cree Ud. que el sistema web influiría en la mejora de su trabajo como monitor municipal?

   ……

6. ¿Cree Ud. que el sistema web se adapta mejor para la seguridad pública o privada?

   ……

7. ¿De acuerdo a su experiencia puede indicar algunas ventajas del sistema web con respecto al método tradicional?

   ……

8. ¿El tiempo de respuesta para emitir una alerta sobre alguna acción violenta detectada, necesariamente debe ser rápido o muy rápido para combatir la delincuencia?

   ……

9. ¿La forma de alertar del sistema web es la más adecuada o que podría recomendar?

   ……

10. ¿Realmente cree que un futuro estos tipos de sistemas deberían ser tomados más en cuenta y/o debería recibir más capacitación sobre estos temas?

    ……

Anexo 6: PREGUNTAS DE LA ENTREVISTA A LOS EXPERTOS DEL COEM

1. ¿El COEM cuenta con Sistema de Videovigilancia basado en Inteligencia Artificial (IA)?

   NO.

2. ¿Qué opinas sobre el hecho de poder contar con IA para combatir la delincuencia?

   APLICARLO SERÍA EXELENTE

3. ¿Crees que el aplicativo que te presentamos podría ayudar en algo o hasta qué nivel llegaría?

   SERIA MUY IMPORTANTE PARA EL DESARROLLO

4. ¿Si Ud. estaría dispuesto a aplicar nuestro sistema web que condición nos pondría?

   QUE SEA MUY EFICIENTE Y FASIL DE MANEJAR

5. ¿Cree Ud. que el sistema web influiría en la mejora de su trabajo como monitor municipal?

   CLARO QUE SI LA TECNOLOGIA ES BUENA.

6. ¿Cree Ud. que el sistema web se adapta mejor para la seguridad pública o privada?

   PRIVADA

7. ¿De acuerdo a su experiencia puede indicar algunas ventajas del sistema web con respecto al método tradicional?

   ES MODERNO ESTAS ACTUALIZADO

8. ¿El tiempo de respuesta para emitir una alerta sobre alguna acción violenta detectada, necesariamente debe ser rápido o muy rápido para combatir la delincuencia?

   CREO QUE SERIA MAS RAPIDO

9. ¿La forma de alertar del sistema web es la más adecuada o que podría recomendar?

   SI MAS ADECUADA

10. ¿Realmente cree que un futuro estos tipos de sistemas deberían ser tomados más en cuenta y/o debería recibir más capacitación sobre estos temas?

    POR SUPUESTO QUE SI

Richard Ramirez Nicodemi

DNI: N°

Anexo 6: PREGUNTAS DE LA ENTREVISTA A LOS EXPERTOS DEL COEM

1. ¿El COEM cuenta con Sistema de Videovigilancia basado en Inteligencia Artificial (IA)?

Aún no contamos con este tipo de sistema

2. ¿Qué opinas sobre el hecho de poder contar con IA para combatir la delincuencia?

Nos facilitaría mucho contar con un sistema así ya nos daría ventaja en diferentes aspectos del trabajo.

3. ¿Crees que el aplicativo que te presentamos podría ayudar en algo o hasta qué nivel llegaría?

Claro, porque nos facilita mucho los lugares exactos e imágenes para poder tener una intervención.

4. ¿Si Ud. estaría dispuesto a aplicar nuestro sistema web que condición nos pondría?

Se podría decir que es optima pos que cumple con la condición básica de la app, aunque, faltan cosas como detalles para mejor (por lo que se vio)

5. ¿Cree Ud. que el sistema web influiría en la mejora de su trabajo como monitor municipal?

Desde luego porque, como ya mencione nos ayuda a ser mejor y a llegar al lugar de los hechos.

6. ¿Cree Ud. que el sistema web se adapta mejor para la seguridad pública o privada?

Creo que la función de esta app se adapte mejor en la seguridad pública ya que nuestra área y otras facilitan su trabajo con esta.

7. ¿De acuerdo a su experiencia puede indicar algunas ventajas del sistema web con respecto al método tradicional?

Con esta app no tenemos que esperar llamadas ya que podemos visualizar y actualizar nos donde está la incidencia.

8. ¿El tiempo de respuesta para emitir una alerta sobre alguna acción violenta detectada, necesariamente debe ser rápido o muy rápido para combatir la delincuencia?

Debe ser muy rápido en casos de violencia ya que son casos muy riesgosos en los que se debe actuar con la mayor rápidez posible.

9. ¿La forma de alertar del sistema web es la más adecuada o que podría recomendar?

En casos como violencia, robos o emergencia medica deberia tener prioridad, alertas mas vistosas?

10. ¿Realmente cree que un futuro estos tipos de sistemas deberían ser tomados más en cuenta y/o debería recibir más capacitación sobre estos temas?

Claro, estos tipos de apps nos van a ayudar mucho a la efectividad del trabajo y también a optimizarlo, en muchos casos estas apps ya tendríamos que estar usando ya que nos estamos quedando atras con lo que respecta al avance tecnológico.

Norma Riva Flores.

Anexo 6: PREGUNTAS DE LA ENTREVISTA A LOS EXPERTOS DEL COEM

1. ¿El COEM cuenta con Sistema de Videovigilancia basado en Inteligencia Artificial (IA)?
   El Coem aun no cuenta.

2. ¿Qué opinas sobre el hecho de poder contar con IA para combatir la delincuencia?
   Nos Facilitaria en el Monitoreo

3. ¿Crees que el aplicativo que te presentamos podría ayudar en algo o hasta qué nivel llegaría?
   Si tubiera el I.A. Fuera de Utilidad en el COEM.

4. ¿Si Ud. estaría dispuesto a aplicar nuestro sistema web que condición nos pondría?
   Que Siempre Estubieran Actualizando el Sistema.

5. ¿Cree Ud. que el sistema web influiría en la mejora de su trabajo como monitor municipal?
   Si mejoraria en el trabajo con el Monitor del COEM.

6. ¿Cree Ud. que el sistema web se adapta mejor para la seguridad pública o privada?
   Si creo mejor para la Seguridad Publica

7. ¿De acuerdo a su experiencia puede indicar algunas ventajas del sistema web con respecto al método tradicional?
   La Velocidad.

8. ¿El tiempo de respuesta para emitir una alerta sobre alguna acción violenta detectada, necesariamente debe ser rápido o muy rápido para combatir la delincuencia?
   debeia de Ser muy napida con evaluación inmediata.

9. ¿La forma de alertar del sistema web es la más adecuada o que podría recomendar?
   Buscar una Forma de prevención cuando el Fluido se Va

10. ¿Realmente cree que un futuro estos tipos de sistemas deberían ser tomados más en cuenta y/o debería recibir más capacitación sobre estos temas?
    Capacitacion al Personal, Actualización al Sistema Para una mejor atención.

ALEYDA ZUMAETA MANGUINURI

DNI 06353338

Anexo D

Fotografías de la visita al COEM

Anexo E

Código del archivo principal del sistema “app.py”

```
# Python In-built packages
from pathlib import Path
import PIL

# External packages
import streamlit as st

# Local Modules
import settings
import helper

# Setting page layout
st.set_page_config(
    page_title="Object Detection using YOLOv8",
    page_icon="🤖",
    layout="wide",
    initial_sidebar_state="expanded"
)

# Main page heading
st.title("Object Detection using YOLOv8")

# Sidebar
st.sidebar.header("ML Model Config")

# Model Options
model_type = st.sidebar.radio(
    "Select Task", ['Detection', 'Segmentation'])

confidence = float(st.sidebar.slider(
    "Select Model Confidence", 25, 100, 40)) / 100

# Selecting Detection Or Segmentation
if model_type == 'Detection':
    model_path = Path(settings.DETECTION_MODEL)
elif model_type == 'Segmentation':
    model_path = Path(settings.SEGMENTATION_MODEL)

# Load Pre-trained ML Model
try:
    model = helper.load_model(model_path)
except Exception as ex:
    st.error(f"Unable to load model. Check the specified path: {model_path}")
    st.error(ex)

st.sidebar.header("Image/Video Config")
source_radio = st.sidebar.radio(
    "Select Source", settings.SOURCES_LIST)

source_img = None
if source_radio == settings.VIDEO:
    helper.play_stored_video(confidence, model)

elif source_radio == settings.WEBCAM:
    helper.play_webcam(confidence, model)

elif source_radio == settings.RTSP:
```

```
    helper.play_rtsp_stream(confidence, model)

elif source_radio == settings.YOUTUBE:
    helper.play_youtube_video(confidence, model)

else:
    st.error("Please select a valid source type!")
```

Código del archivo de configuración “settings.py”

```
from pathlib import Path
import sys

# Get the absolute path of the current file
file_path = Path(__file__).resolve()

# Get the parent directory of the current file
root_path = file_path.parent

# Add the root path to the sys.path list if it is not already there
if root_path not in sys.path:
    sys.path.append(str(root_path))

# Get the relative path of the root directory with respect to the current working
directory
ROOT = root_path.relative_to(Path.cwd())

# Sources
# IMAGE = 'Image'
VIDEO = 'Video'
WEBCAM = 'Webcam'
RTSP = 'RTSP'
YOUTUBE = 'YouTube'

SOURCES_LIST = [VIDEO, WEBCAM, RTSP, YOUTUBE]

# Images config
IMAGES_DIR = ROOT / 'images'
DEFAULT_IMAGE = IMAGES_DIR / 'office_4.jpg'
DEFAULT_DETECT_IMAGE = IMAGES_DIR / 'office_4_detected.jpg'

# Videos config
VIDEO_DIR = ROOT / 'videos'
VIDEO_1_PATH = VIDEO_DIR / 'video_1.mp4'
VIDEO_5_PATH = VIDEO_DIR / 'video_5.mp4'
VIDEO_6_PATH = VIDEO_DIR / 'video_rodo_jhon.mov'
VIDEO_7_PATH = VIDEO_DIR / 'pelea.mp4'
VIDEOS_DICT = {
    'pelea_calle': VIDEO_5_PATH,
    'video_rodo_jhon': VIDEO_6_PATH,
    'pelea': VIDEO_7_PATH,
    'detect_object': VIDEO_1_PATH,
}

# ML Model config
MODEL_DIR = ROOT / 'weights'
DETECTION_MODEL = MODEL_DIR / 'best.pt'
#DETECTION_MODEL = MODEL_DIR / 'best_weight_yolov8_full_integer_quant.tflite'
# In case of your custome model comment out the line above and
```

```
# Place your custom model pt file name at the line below
# DETECTION_MODEL = MODEL_DIR / 'my_detection_model.pt'

SEGMENTATION_MODEL = MODEL_DIR / 'yolov8n-seg.pt'

# Webcam
WEBCAM_PATH = 0
```

Código del archivo “helper.py”

```
from ultralytics import YOLO
import time
import streamlit as st
import cv2
from pytube import YouTube

import settings


def load_model(model_path):
    """
    Loads a YOLO object detection model from the specified model_path.

    Parameters:
        model_path (str): The path to the YOLO model file.

    Returns:
        A YOLO object detection model.
    """
    model = YOLO(model_path)
    return model


def display_tracker_options():
    display_tracker = st.radio("Display Tracker", ('Yes', 'No'))
    is_display_tracker = True if display_tracker == 'Yes' else False
    if is_display_tracker:
        tracker_type = st.radio("Tracker", ("bytetrack.yaml", "botsort.yaml"))
        return is_display_tracker, tracker_type
    return is_display_tracker, None


def _display_detected_frames(conf, model, st_frame, image, is_display_tracking=None,
tracker=None):
    """
    Display the detected objects on a video frame using the YOLOv8 model.

    Args:
    - conf (float): Confidence threshold for object detection.
    - model (YoloV8): A YOLOv8 object detection model.
    - st_frame (Streamlit object): A Streamlit object to display the detected video.
    - image (numpy array): A numpy array representing the video frame.
    - is_display_tracking (bool): A flag indicating whether to display object tracking
(default=None).

    Returns:
    None
    """

    # Resize the image to a standard size
    image = cv2.resize(image, (720, int(720*(9/16))))
```

```
    # Display object tracking, if specified
    if is_display_tracking:
        res = model.track(image, conf=conf, persist=True, tracker=tracker)
    else:
        # Predict the objects in the image using the YOLOv8 model
        res = model.predict(image, conf=conf)

    # # Plot the detected objects on the video frame
    res_plotted = res[0].plot()
    st_frame.image(res_plotted,
                   caption='Detected Video',
                   channels="BGR",
                   use_column_width=True
                   )


def play_youtube_video(conf, model):
    """
    Plays a webcam stream. Detects Objects in real-time using the YOLOv8 object
detection model.

    Parameters:
        conf: Confidence of YOLOv8 model.
        model: An instance of the `YOLOv8` class containing the YOLOv8 model.

    Returns:
        None

    Raises:
        None
    """
    source_youtube = st.sidebar.text_input("YouTube Video url")

    is_display_tracker, tracker = display_tracker_options()

    if st.sidebar.button('Detect Objects'):
        try:
            yt = YouTube(source_youtube)
            stream = yt.streams.filter(file_extension="mp4", res=720).first()
            vid_cap = cv2.VideoCapture(stream.url)

            st_frame = st.empty()
            while (vid_cap.isOpened()):
                success, image = vid_cap.read()
                if success:
                    _display_detected_frames(conf,
                                             model,
                                             st_frame,
                                             image,
                                             is_display_tracker,
                                             tracker,
                                             )
                else:
                    vid_cap.release()
                    break
        except Exception as e:
            st.sidebar.error("Error loading video: " + str(e))


def play_rtsp_stream(conf, model):
```

```
    """
    Plays an rtsp stream. Detects Objects in real-time using the YOLOv8 object
detection model.

    Parameters:
        conf: Confidence of YOLOv8 model.
        model: An instance of the `YOLOv8` class containing the YOLOv8 model.

    Returns:
        None

    Raises:
        None
    """
    source_rtsp = st.sidebar.text_input("rtsp stream url:")
    st.sidebar.caption('Example URL:
rtsp://admin:12345@192.168.1.210:554/Streaming/Channels/101')
    is_display_tracker, tracker = display_tracker_options()
    if st.sidebar.button('Detect Objects'):
        try:
            vid_cap = cv2.VideoCapture(source_rtsp)
            st_frame = st.empty()
            while (vid_cap.isOpened()):
                success, image = vid_cap.read()
                if success:
                    _display_detected_frames(conf,
                                             model,
                                             st_frame,
                                             image,
                                             is_display_tracker,
                                             tracker
                                             )
                else:
                    vid_cap.release()
                    # vid_cap = cv2.VideoCapture(source_rtsp)
                    # time.sleep(0.1)
                    # continue
                    break
        except Exception as e:
            vid_cap.release()
            st.sidebar.error("Error loading RTSP stream: " + str(e))


def play_webcam(conf, model):
    """
    Plays a webcam stream. Detects Objects in real-time using the YOLOv8 object
detection model.

    Parameters:
        conf: Confidence of YOLOv8 model.
        model: An instance of the `YOLOv8` class containing the YOLOv8 model.

    Returns:
        None

    Raises:
        None
    """
    source_webcam = settings.WEBCAM_PATH
    is_display_tracker, tracker = display_tracker_options()
    if st.sidebar.button('Detect Objects'):
        try:
```

```
            vid_cap = cv2.VideoCapture(source_webcam)
            st_frame = st.empty()
            while (vid_cap.isOpened()):
                success, image = vid_cap.read()
                if success:
                    _display_detected_frames(conf,
                                             model,
                                             st_frame,
                                             image,
                                             is_display_tracker,
                                             tracker,
                                             )
                else:
                    vid_cap.release()
                    break
        except Exception as e:
            st.sidebar.error("Error loading video: " + str(e))


def play_stored_video(conf, model):
    """
    Plays a stored video file. Tracks and detects objects in real-time using the
YOLOv8 object detection model.

    Parameters:
        conf: Confidence of YOLOv8 model.
        model: An instance of the `YOLOv8` class containing the YOLOv8 model.

    Returns:
        None

    Raises:
        None
    """
    source_vid = st.sidebar.selectbox(
        "Choose a video...", settings.VIDEOS_DICT.keys())

    is_display_tracker, tracker = display_tracker_options()

    with open(settings.VIDEOS_DICT.get(source_vid), 'rb') as video_file:
        video_bytes = video_file.read()
    if video_bytes:
        st.video(video_bytes)

    if st.sidebar.button('Detect Video Objects'):
        try:
            vid_cap = cv2.VideoCapture(
                str(settings.VIDEOS_DICT.get(source_vid)))
            st_frame = st.empty()
            while (vid_cap.isOpened()):
                success, image = vid_cap.read()
                if success:
                    _display_detected_frames(conf,
                                             model,
                                             st_frame,
                                             image,
                                             is_display_tracker,
                                             tracker
                                             )
                else:
                    vid_cap.release()
                    break
```

```
except Exception as e:
    st.sidebar.error("Error loading video: " + str(e))
```