Departamento de Informática y Automática

Facultad de Ciencias

Tesis Doctoral

Inteligencia artificial fiable para la detección de violencia en vídeo

Autor:

D. Pablo Negre Rodríguez

Directores:

Dr. D. Javier Prieto Tejedor

Dr. D. Ricardo Serafín Alonso Rincón

Julio de 2025

University of Salamanca

Department of Computer Science and Automation

Faculty of Sciences

Doctoral Thesis

Trustworthy artificial intelligence for video violence detection

Author:

D. Pablo Negre Rodríguez

Advisors:

Dr. D. Javier Prieto Tejedor

Dr. D. Ricardo Serafín Alonso Rincón

July, 2025

Universidad de Salamanca

Departamento de Informática y Automática

Facultad de Ciencias

Inteligencia artificial fiable para la detección de violencia en vídeo

Autor:

D. Pablo Negre Rodríguez

Directores:

Dr. D. Javier Prieto Tejedor

Dr. D. Ricardo Serafín Alonso Rincón

Tribunal

Presidente:

Vocal:

Secretario:

Suplentes:

Fecha de lectura:

Calificación:

A mi madre

A mi hermana

A Max

Solicitud de Presentación de Tesis Doctoral

Estimado Coordinador del Programa de Doctorado en Ingenería Informática:

D. Pablo Negre Rodríguez, con DNI 209******, y alumno del programa de DOCTORADO EN INGENIERÍA INFORMÁTICA, matriculado en el plan de estudios D015 INGENIERÍA INFORMÁTICA (R.D. 99/2011) y con número de expediente 160:

Solicita que se tenga en consideración la información aportada en este documento con el objeto de poder presentar la Tesis Doctoral con título *Inteligencia artificial fiable para la detección de violencia en vídeo.*

A continuación se detallan los documentos adjuntos en esta solicitud:

- Autorización de los directores para la presentación de la Tesis Doctoral.
- Introducción y resumen de la Tesis Doctoral presentada.
  - Introducción.
  - Técnicas de detección de violencia en vídeo mediante IA.
  - CNN y su combinación con un número mínimo de fotogramas violentos.
  - VGG-19 combinada con RNN.
  - Arquitectura explicable basada en VGG-19 preentrenada con capas LSTM/Bi-LSTM.
  - Conclusiones.

En Salamanca, a 23 de junio de 2025

El doctorando

D. Pablo Negre Rodríguez

Firmado por PABLO NEGRE RODRIGUEZ - NIF:***0422** el día 26/06/2025 con un certificado emitido por ACCV RSA1 CLIENTE

Autorización de los Directores

En Salamanca, a 18 de junio de 2025

HACEMOS CONSTAR:

Que, como directores de la Tesis Doctoral de **Pablo Negre Rodríguez**, con DNI 209******, autorizamos a presentar la Tesis Doctoral **"Inteligencia artificial fiable para la detección de violencia en vídeo"** al disponer de todos los requisitos mínimos requeridos por la Universidad de Salamanca (USAL) para el depósito de la misma.

Los directores:

Dr. Javier Prieto Tejedor.

Firmado por PRIETO TEJEDOR JAVIER - ***6981** el día 27/06/2025

Dr. Ricardo Serafín Alonso Rincón

Firmado por ALONSO RINCON RICARDO SERAFIN - ***8531** el día

"Estás hecho de la madera de los grandes, pero tienes que tomar el timón y trazar tu propio curso.

No te rindas, a pesar de las borrascas.

Y cuando llegue el momento, tendrás la oportunidad de probar el corte de tus velas y demostrar lo que vales.

Y yo, espero poder ver la luz que tus velas henchidas despedirán ese día".

...

"Harás vibrar a las estrellas"

John Silver. El planeta del tesoro.

Resumen

Las agresiones físicas son un problema grave y generalizado, como lo demuestra el hecho de que más de una cuarta parte (27 %) de las mujeres de entre 15 y 49 años a nivel global declaran haber sido sometidas a algún tipo de violencia física y/o sexual por parte de su pareja íntima. La Inteligencia Artificial y específicamente las técnicas de Visión Artificial, ofrecen una solución eficaz para detectar la violencia en tiempo real, reduciendo la necesidad de supervisión humana constante. La Inteligencia Artificial, y en particular las técnicas de Visión Artificial, pueden contribuir a identificar episodios de violencia en tiempo real en lugares previamente delimitados, respetando los marcos éticos y legales establecidos. Sin embargo, el aumento del uso de la inteligencia artificial ha generado preocupación sobre la fiabilidad de los algoritmos, lo que ha llevado a la creación de informes destinados a establecer estándares y guías, con organizaciones como la Comisión Europea liderando estos esfuerzos. En este respecto, existen múltiples propuestas de algoritmos para la detección de violencia, donde la combinación de arquitecturas más comúnmente empleada es la de Redes Neuronales Convolucionales (CNN) y Redes de Memoria a Corto y Largo Plazo (LSTM), la cual obtiene excelentes resultados, si bien todavía persisten desafíos; sin embargo, hasta donde se conoce, ningún trabajo en el estado del arte ha abordado la detección de violencia mediante el uso de inteligencia artificial explicable, lo que limita la comprensión y confianza en los resultados obtenidos. Por ello, el objetivo principal de esta Tesis Doctoral es investigar, diseñar, desarrollar y validar algoritmos basados en técnicas de inteligencia artificial fiable orientadas en la detección de violencia en vídeo, con foco en arquitecturas basadas en la combinación de CNN junto con capas LSTM. En base a ello, en este trabajo se ha llevado a cabo un análisis y categorización de todos los procesos que involucran la detección de violencia en vídeo. Posteriormente se han investigado, diseñado, desarrollado y validado tres arquitecturas que utilizan la arquitectura VGG-19 preentrenada, una red neuronal convolucional conocida por su capacidad para extraer características visuales, combinadas con: características manuales, capas LSTM y capas Bi-LSTM. Por último, a partir de estas arquitecturas se han implementado técnicas de inteligencia artificial explicable como GradCAM y se ha creado un algoritmo que cuantifica el nivel de importancia para la detección de violencia por parte de las capas LSTM y Bi-LSTM. Los resultados obtenidos demuestran que el uso de capas Bi-LSTM supera al rendimiento obtenido por capas LSTM, si bien esta mejora no supera el 4 % de exactitud. No se han encontrado valores o combinaciones de hiperparámetros para las arquitecturas que utilizan capas LSTM y Bi-LSTM que mejoren de una forma estadísticamente significativa la *accuracy* obtenida. Las arquitecturas desarrolladas han obtenido buenos reusltados como, por ejemplo, la combinación de VGG-19 preentrenada con capas Bi-LSTM, que obtiene un 97 % de exactitud utilizando el dataset *Hockey Fights*. Por último, se ha conseguido hacer más explicable el proceso de detección con las técnicas implementadas.

Abstract

Physical aggressions constitute a serious and widespread issue in society. Studies indicate that in 2015, at least half of the children in Asia, Africa, and North America experienced violence. Although solutions have been explored for medium and long-term interventions, real-time violence detection through artificial intelligence offers a direct and efficient solution that can save lives and reduce the need for constant human supervision. On the other hand, the increasing use of artificial intelligence has raised concerns about the development of reliable algorithms, leading to the creation of reports to define and standardize these terms. Major organizations such as the European Comission are leading this effort. There are multiple algorithm proposals for violence detection, with the most commonly employed combination being Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, which yield excellent results. However, there are still issues to address, such as the actual impact of using LSTM layers instead of just CNN, how much violence detection improves with CNN combined with Bi-LSTM layers instead of LSTM layers, or if certain values and combinations of hyperparameters yield better results. Lastly, the use of reliable artificial intelligence remains very limited. Based on this, this work has developed a systematic literature review with the analysis and categorization of: 21 challenges associated with violence detection, 28 public datasets on violence vídeos, and 13 evaluation metric methods; among others. Three architectures have been developed using pre-trained VGG-19 combined with: manual features, LSTM layers, and Bi-LSTM layers. It is evident that the use of Bi-LSTM layers outperforms the performance obtained by LSTM layers, although this improvement does not exceed 3% accuracy. No values or combinations of hyperparameters that significantly improve the obtained accuracy have been found statistically. The developed architectures have achieved good results, such as the combination of pre-trained VGG-19 with Bi-LSTM layers, which achieves 97% accuracy using the Hockey Fights dataset and 90% using the Violent Flow dataset. Lastly, the use of explainable artificial intelligence techniques on the proposed architectures, where YoloV8 and Frame Difference are used for the extraction of characteristic frames, GradCAM to highlight the areas VGG-19 focuses on for each convolutional layer, and a proprietary algorithm quantifies the level of importance for violence detection by LSTM and Bi-LSTM layers in violence detection.

Agradecimientos

A mi director, Ricardo, por ser el primero que me abrió la puerta al sector de la Inteligencia Artificial, que se ha convertido en mi profesión y especialidad en esta Tesis Doctoral. Por exigirme un Trabajo de Fin de Máster de excelencia. Por animarme a embarcarme en este maravilloso proyecto que ha sido esta Tesis Doctoral. Por tener reuniones y responder a mensajes, fuese de noche o fin de semana (a veces, ambas). Por facilitar y estar presente en todo lo que ha conllevado esta Tesis Doctoral. En definitiva, por ser el mejor director de tesis que nadie pueda tener, con total seguridad. Gracias y mil veces gracias.

A mi director, Javier Prieto, y a Juan Manuel Corchado, por sus correcciones, y por su ayuda con la documentación y trámites necesarios para llevar esta Tesis Doctoral a buen puerto.

Al profesor Dr. Dang Nhan Cach, al profesor Ho Dand The y a todo el equipo de la Ho Chi Minh University of Transport, por permitirme realizar una estancia doctoral en su universidad y por la maravillosa y tan diferente experiencia que supuso para mi Tesis Doctoral y a nivel personal.

Al profesor Dr. Paulo Novais y a todo el equipo de la Universidade do Minho, por acogerme dentro del equipo como uno más y brindarme la oportunidad de colaborar juntos durante mi estancia doctoral en Braga, Portugal.

A mi madre, por ser la principal razón y motivo de la persona en la que me he convertido. Por apostar y confiar en mí en todo momento, sin excepción. Todo lo que soy y he conseguido te lo debo a ti.

A mi hermana, por ser capaz de animarme y ver la luz al final del túnel en cualquier situación. Por crear una complicidad entre hermanos que a muchos les gustaría tener y que muy pocos logran encontrar. Sin tu apoyo, no habría podido conseguir ni afrontar de la misma forma muchos de los objetivos que he logrado.

A Max, cuya constante necesidad de dormir dieciséis horas diarias, cambiando estratégicamente de posición para descansar más plácidamente, así como la gran cantidad de amor que nos has dado, te convierte en el mejor perro que podría haber imaginado. Gracias por todas esas largas horas en casa conmigo mientras trabajaba.

Índice general

Índice de figuras

Índice de tablas

Índice de listados

Siglas y acrónimos

AI	***A**rtificial **I**ntelligence*
Bi-LSTM	***Bi-**directional **L**ong **S**hort **T**erm **M**emory*
CCTV	***C**losed **C**ircuit **T**elevision*
CNN	***C**onvolutional **N**eural **N**etwork*
Conv-LSTM	***Conv**olutional **L**ong **S**hort **T**erm **M**emory*
CV	***C**omputer **V**ision*
DL	***D**eep **L**earning*
GRU	***G**ated **R**ecurrent **U**nit*
HD	***H**igh **D**efinition*
IA	**I**nteligencia **A**rtificial
LSTM	***L**ong **S**hort **T**erm **M**emory*
ML	***M**achine **L**earning*
PT	***P**re**T**rained*
RLVS	***R**eal **L**ife **V**iolent **S**ituations*
RNN	***R**ecurrent **N**eural **N**etwork*
RWF-2000	***R**eal **W**orld **F**ight dataset*
SMS	***S**ystematic **M**apping **S**tudy*
SSMI	*Segmentation based on **S**tructural and **S**tatistical **M**etrics for **I**mages*
SOTA	***S**tate **O**f **T**he **A**rt*
XAI	*e**X**plainable **A**rtificial **I**nteligence*

Capítulo 1

Introducción

Capítulo 1

Introducción

Índice

Las agresiones físicas constituyen un problema social de gran relevancia y con múltiples consecuencias. Diversas investigaciones se han llevado a cabo para enfrentar esta problemática desde diferentes perspectivas, con variados niveles de intervención. Entre las posibles soluciones, la detección en tiempo real de actos violentos en vídeo emerge como una respuesta crucial, actuando como un recurso final para identificar a las víctimas de agresiones físicas. Este avance ha sido posible gracias a factores determinantes como el uso extendido de cámaras de seguridad, la disponibilidad de plataformas con datos a gran escala y el desarrollo de algoritmos avanzados que aplican inteligencia artificial al análisis de imágenes y vídeos.

En esta Tesis Doctoral se propone el desarrollo de técnicas de violencia en vídeo mediante inteligencia artificial fiable (TAI), con el objetivo de abordar ciertos aspectos del estado del arte todavía no estudiados como, por ejemplo, si existen combinaciones óptimas de hiperparámetros que obtienen una mayor exactitud o si es posible hacer más explicable el proceso de detección de violencia en vídeo. Para ello, se desarrollan tres arquitecturas basadas en la combinación de *Convolutional Neural Networks* (CNN) junto con: *(1)* características manuales, *(2)* capas *Long Short Term Memory* (LSTM), *(3)* capas *Bidirectional Long Short Term Memory* (Bi-LSTM) (Mumtaz et al., 2022b); respectivamente. Además, se aplican en estas arquitecturas técnicas de inteligencia artificial explicable (XAI) herramientas como GradCAM y algoritmos de desarrollo

propio (Abhishek & Kamath, 2022), con el fin de hacer más comprensible el proceso de toma de decisiones del algoritmo.

El presente Capítulo introductorio hace un repaso a los antecedentes en relación a la detección de la violencia en vídeo mediante inteligencia artificial fiable (TAI), tema principal de esta Tesis Doctoral. Así, la Sección 1.1 realiza una introducción acerca de los antecedentes del tema objeto de estudio y describe el problema a abordar. La Sección 2.5 expone la hipótesis de trabajo, así como los objetivos de los distintos trabajos de investigación que han conducido a la elaboración de esta Tesis Doctoral. La Sección 1.3 detalla la metodología seguida durante dichos trabajos de investigación. Finalmente, la Sección 1.3 describe la estructura de esta memoria.

1.1. Descripción del problema

Esta Sección expone la descripción del problema que trata esta Tesis Doctoral: la detección de violencia en vídeo mediante inteligencia artificial precisa y fiable. La Sección se divide en dos partes. La Sección 1.1.1 expone datos actuales sobre las agresiones físicas en el mundo, así como las diferentes propuestas que se han realizado para afrontarla en distintos plazos de tiempo. La Sección 1.1.2 expone el aumento de la relevancia del uso de inteligencia artificial fiable y los pilares fundamentales que la componen.

1.1.1. La violencia en la sociedad y sus soluciones existentes

Las agresiones físicas afectan a prácticamente todos los aspectos de la vida, no solo a las víctimas directas, sino también a las familias, la sociedad y el desarrollo diario de los países del mundo actual (Muarifah et al., 2022). Asimismo, estas agresiones afectan a la economía (compras, viajes o turismo) y la salud mental de los ciudadanos debido al estado permanente de inseguridad (Long et al., 2021). Algunos estudios muestran datos alarmantes, como que las agresiones físicas son la principal causa de mortalidad entre los jóvenes de 15 a 44 años en todo el mundo en la primera década del siglo XXI (Shubber & Al-Ta'i, 2022). Otros estudios afirman cómo un 27 % de las mujeres a nivel global declaran haber sufrido violencia por parte de su pareja íntima (White et al., 2024). Por todo esto, se enfatiza el problema de la agresión física en las sociedades de todo el mundo.

Se han propuesto **múltiples soluciones a largo, medio y corto plazo**. Las soluciones a **largo plazo** consisten en comprender las razones o situaciones que exacerban la violencia para que las generaciones futuras puedan ser educadas para evitarlas (Long et al., 2021; Martínez-González et al., 2021). Algunos estudios han investigado un aumento de la violencia por razones como la exposición a un entorno de violencia y agresividad, lo que conduce a una tendencia a usar la violencia como una herramienta (Martínez-González et al., 2021). Estos estudios también investigan cómo la relación materna y el nivel de autoestima han mostrado tener una relación directa con los niveles de agresiones (Muarifah et al., 2022) o cómo la influencia de las tecnologías genera un aumento de la violencia debido al uso de teléfonos inteligentes o al consumo de vídeojuegos agresivos (Fekih-Romdhane et al., 2022). Asimismo, los autores han observado que la rápida urbanización en China durante el siglo XXI, junto con el crecimiento de la población, ha llevado a más delincuencia (Long et al., 2021). Otros estudios analizan sobre cómo el COVID-19, más específicamente las consecuencias que ha tenido en la vida de las personas, ha llevado a un aumento de la violencia, produciendo una «tormenta perfecta» de factores (Killgore et al., 2021).

Las **soluciones a medio plazo** consisten en enfrentar la violencia tratando de relacionar la densidad de población callejera y el paisaje urbano con las tasas de criminalidad. Por ejemplo, algunos estudios han utilizado datos de teléfonos móviles y de redes sociales para medir la densidad de población en áreas urbanas y relacionarla con la tasa de criminalidad utilizando técnicas estadísticas centradas en el análisis espacial (Vomfell et al., 2018), aunque existe el inconveniente de que no es posible diferenciar entre poblaciones en interiores y exteriores. Asimismo, investigaciones recientes han analizado cómo se reduce la criminalidad en áreas donde hay zonas verdes, empleando para ello el uso de CNN (Hipp et al., 2021). Los métodos a medio plazo que analizan y comparan la cantidad de personas o tráfico en las calles y el tipo de paisaje urbano con la tasa de criminalidad proporcionan información real y específica sobre qué áreas son más propensas a agresiones físicas y cómo construir nuevos paisajes urbanos que induzcan menos a la criminalidad (Jing et al., 2021). Sin embargo, es un enfoque fuera de línea que no protege a las víctimas ante una agresión. Por ello, **las soluciones a corto plazo existentes consisten en la detección de la violencia en tiempo real** (Jing et al., 2021).

En cuanto a **las soluciones a corto plazo**, estas se centran en la detección de la

violencia en tiempo real.Hasta ahora, las cámaras de seguridad se han utilizado en gran medida como prueba para las personas afectadas ante compañías de seguros, policía o jueces (Shukla & Pandey, 2020). Sin embargo, existen varios trabajos que han desarrollado algoritmos que permiten la **detección de violencia en tiempo real a través del uso de cámaras de seguridad utilizando algoritmos de inteligencia artificial (IA)** (Cheng et al., 2021b; Jaafar & Lachiri, 2023; Zhou et al., 2023). Este es un campo en crecimiento (Omarov et al., 2022) cuyo desarrollo **ha sido posible debido a la evolución de tres pilares principales: el aumento del uso de cámaras de seguridad** (Afra & Alhajj, 2020; Vosta & Yow, 2022), **el desarrollo tecnológico de plataformas de *Big Data*** (Ageed & Zeebaree, 2021; Alonso et al., 2020) **y la implementación de algoritmos de inteligencia artificial** (Collins et al., 2021) que permiten el análisis de imágenes y vídeo (Ding et al., 2021).

En relación con el primer pilar, la instalación de cámaras de vigilancia ha aumentado en todo el mundo por razones de monitoreo y seguridad, entre otras (Afra & Alhajj, 2020; Vosta & Yow, 2022). Es importante destacar que, si bien el uso de imágenes y vídeo presenta una gran cantidad de aplicaciones que pueden mejorar la calidad de vida de los ciudadanos, ya existen estudios que hablan sobre el riesgo para la privacidad personal que representan las grabaciones a gran escala (Kostka et al., 2020). En relación con el segundo pilar, las plataformas de *Big Data* actualmente nos ofrecen la capacidad de obtener y procesar una gran cantidad de datos del entorno que nos rodea (Ali et al., 2018; Alonso et al., 2020). Estas herramientas nos permiten capturar, almacenar y analizar en tiempo real diferentes tipos de información como imágenes, vídeos y otros tipos de datos (Ageed & Zeebaree, 2021). En relación con el tercer pilar, el uso de algoritmos basados en IA para el análisis de imágenes y vídeo ha crecido exponencialmente en los últimos años. No solo ha aumentado su uso, sino también su variedad y su mejora en precisión (Ding et al., 2021).

Vale la pena preguntarse por qué es necesaria la aplicación de inteligencia artificial para la detección de agresiones físicas, que es el tema que concierne a esta Tesis Doctoral. La realidad es que el análisis constante de imágenes y vídeo por parte de humanos implica o bien un alto costo de personal (salarios o instalaciones) o un bajo nivel de supervisión en el caso de que unas pocas personas monitoreen un gran número de escenarios, lo que puede llevar a la pérdida de vidas (Mugunga et al., 2021a). Es por ello que el uso de

técnicas de análisis de imágenes y vídeo juega un papel crucial en el desarrollo de la detección de violencia (Omarov et al., 2022).

En conclusión, las agresiones físicas representan una preocupación social con amplias implicaciones. Numerosas investigaciones han centrado sus esfuerzos en abordar este problema a través de diversos enfoques, cada uno con diferentes grados de directividad. La identificación en tiempo real de la violencia destaca como la solución más inmediata, sirviendo como la última línea de defensa para identificar a las personas sometidas a daño físico (Cheng et al., 2021b; Jaafar & Lachiri, 2023; Zhou et al., 2023). Este logro solo es posible gracias a los tres pilares anteriores: la creciente utilización de vídeos de seguridad, el avance de plataformas de datos a gran escala y la evolución de algoritmos capaces de analizar imágenes y vídeo potenciados por inteligencia artificial (Mugunga et al., 2021a). Sin embargo, para obtener resultados útiles, es esencial el uso de IA fiable, tal y como refleja la siguiente Sección.

1.1.2. La inteligencia artificial fiable

La búsqueda de una inteligencia artificial fiable no es reciente. Desde sus inicios, la IA ha evolucionado significativamente, pasando de simples algoritmos determinísticos a complejas redes neuronales capaces de aprender patrones intrincados a partir de grandes cantidades de datos. En este contexto, la idea de "fiabilidad.[en] la IA surgió como una preocupación clave, especialmente en las décadas recientes, cuando estas tecnologías comenzaron a influir en áreas críticas como la salud, la justicia y la seguridad pública. Este creciente interés ha llevado a la creación de enfoques interdisciplinarios para garantizar que las decisiones basadas en IA sean transparentes, justas y robustas (Kaur et al., 2022).

En particular, la intersección entre IA y análisis de vídeo ha sido un área de notable crecimiento y estudio. El uso de IA en este campo ha demostrado ser valioso en múltiples aplicaciones, desde la vigilancia del tráfico hasta el análisis de contenido audiovisual en tiempo real. Sin embargo, estas capacidades también han generado preguntas éticas y de fiabilidad, ya que los sistemas deben operar de manera consistente y sin prejuicios en contextos altamente dinámicos. Tecnologías avanzadas como las redes neuronales convolucionales y los algoritmos de visión por computadora han permitido mejorar la precisión en el reconocimiento de objetos y actividades, pero también han evidenciado

la necesidad de explicabilidad en sus decisiones (Chamola et al., 2023), especialmente en escenarios sensibles como la identificación de comportamientos sospechosos o actos violentos.

La detección de violencia mediante IA es una de las aplicaciones más críticas y socialmente relevantes. Estas herramientas se utilizan en sistemas de videovigilancia para identificar incidentes en tiempo real, lo que puede ser vital para prevenir o mitigar situaciones peligrosas. Sin embargo, su implementación plantea retos únicos. Por un lado, la complejidad de las situaciones reales puede dificultar que los algoritmos operen con precisión, especialmente en ambientes caóticos o mal iluminados. Por otro lado, existe el riesgo de sesgos que puedan discriminar a ciertos grupos sociales, subrayando nuevamente la importancia de que estos sistemas sean no solo robustos, sino también explicables y éticamente diseñados (Negre et al., 2024a).

Por todo ello, el creciente uso de sistemas basados en la IA ha llevado a grandes organizaciones a preguntarse cómo de fiables son estos algoritmos (Kaur et al., 2022). Varias grandes organizaciones como la Comisión Europea (European Commission, 2019) o la Organización Internacional de Normalización (ISO) (ISO/IEC, 2020) se han esforzado por crear definiciones, normas y legislación al respecto. En particular, la Comisión Europea ha publicado un informe sobre inteligencia artificial (European Commission, 2019) que define la IA fiable. En su definición, **la IA fiable se basa en tres pilares**:

- *Licitud*: garantizar el cumplimiento de todas las leyes y reglamentos pertinentes en todas las circunstancias posibles.
- *Eticidad*: dar prioridad a los principios y valores que defienden las consideraciones morales y normas éticas fundamentales.
- *Robustez*: asegurar que los sistemas de IA, aunque bienintencionados, no causen daños accidentales tanto técnica como socialmente.

El informe de la Comisión Europea define, además, cómo cada uno de estos componentes es necesario, pero no suficiente para que la IA sea digna de confianza. La opacidad inherente de la mayoría de los modelos de aprendizaje automático profundo (también conocida como el concepto de caja negra o *black box* en inglés),

dificulta la comprensibilidad de sus procesos y resultados y, por tanto, de su fiabilidad (Abhishek & Kamath, 2022). Para aumentar esta comprensibilidad, la comunidad investigadora ha estudiado y discutido ampliamente la ***inteligencia artificial explicable* (XAI)**sustentada en el interés por comprender y transparentar las decisiones de los sistemas de inteligencia artificial (Mahmoudi et al., 2023; Speith, 2022; Swamy et al., 2022).

La falta de transparencia en los modelos de IA puede ser una barrera para su aceptación y adopción, especialmente en aplicaciones críticas donde es importante entender el razonamiento detrás de las decisiones automatizadas (Mahmoudi et al., 2023). Precisamente, la XAI se refiere a la capacidad de entender y explicar cómo un modelo de IA toma decisiones. Speith (2022) ejemplificó por qué los sistemas de detección de agresiones deben utilizar inteligencia artificial explicable, afirmando que, de lo contrario, existiría el riesgo de que un algoritmo tomara decisiones sesgadas, por ejemplo, basándose en el color de la piel de la persona. Por todo ello, es esencial que un sistema para la detección violencia sea explicable y, en consecuencia, fiable, especialmente debido a su papel en la protección de los ciudadanos (Chamola et al., 2023).

1.2. Hipótesis y objetivos

En esta Tesis Doctoral se propone, en primer lugar, el **desarrollo de un estudio de mapeo** que analice todas las partes del proceso de detección de violencia en vídeo, dada la necesidad de una revisión completa y actualizada de este sector. Por otra parte, habiendo obtenido la arquitectura VGG-16 muy buenos resultados en la detección de violencia en video (Negre et al., 2024b) y dado que no se ha explorado en profundidad el potencial de la red VGG-19 en el estado del arte para la detección de violencia, se propone el **desarrollo de tres arquitecturas partiendo de VGG-19 preentrenada** en el dataset de *ImageNet* (Deng et al., 2009). Una de las arquitecturas consiste en la combinación de VGG-19 preentrenada combinada con una lógica manual mediante la cual solo se considera violento un vídeo si la predicción realizada por la arquitectura de la red VGG-19 preentrenada ha predicho como violentos un cierto número de *frames* individuales. La segunda de las arquitecturas combina el uso de la red VGG-19 preentrenada junto con capas LSTM. La tercera y última arquitectura utiliza la red VGG-19 preentrenada junto con capas Bi-LSTM.

Por otro lado, dada la inherente opacidad de los modelos de inteligencia artificial con la consecuente preocupación de grandes organismos como la Comisión Europea (European Commission, 2019), se propone también el **desarrollo de un modelo que utilize técnicas de inteligencia artificial explicable** a partir de las arquitecturas propuestas de forma que todo el proceso de detección de violencia sea más comprensible entorno a su toma de decisión. Se utiliza YoloV8 y la diferencia de fotogramas como algoritmos intrínsicamente explicables para la extracción de frames característicos, GradCAM como algoritmo de explicabilidad enfocado en CNN y un algoritmo de desarrollo propio que cuantifica la relevancia de un conjunto de frames en la detección de violencia en vídeo.

En base a lo anterior, a continuación se presentan la hipótesis principal y las hipótesis específicas de esta Tesis Doctoral, junto con el objetivo principal y los objetivos específicos correspondientes.

La **hipótesis general** de esta Tesis Doctoral se basa en que la combinación de técnicas de Convolutional Neural Networks (CNN) con capas Long Short Term Memory (LSTM) en arquitecturas de inteligencia artificial explicable permitirá mejorar significativamente la precisión y fiabilidad en la detección de violencia en vídeo.

(**HIP1**) Analizar las técnicas y algoritmos utilizados en la detección de violencia en vídeo revelará que, aunque existen enfoques innovadores, las soluciones actuales aún enfrentan limitaciones en cuanto a la precisión y fiabilidad de la detección, lo que justifica la necesidad de desarrollar arquitecturas más avanzadas combinando CNN y LSTM.

(**HIP2**) Diseñar arquitecturas de detección de violencia en vídeo, combinando CNN con características manuales, capas LSTM y Bi-LSTM, conducirá a una mejora sustancial en la capacidad de capturar patrones espaciales y temporales, logrando una detección más precisa, explicable y robusta.

(**HIP3**) Crear arquitecturas explicables para la detección de violencia en vídeo permitirá un entendimiento sustancial del proceso de inferencia, lo que contribuirá no solo al desarrollo del algoritmo, sino también a facilitar su comprensión por parte de los usuarios finales.

(**HIP4**) Implementar y comparar las arquitecturas propuestas, frente a otras soluciones del estado del arte, demostrará que las nuevas arquitecturas basadas en CNN y LSTM ofrecen un rendimiento superior en términos de precisión, fiabilidad y capacidad de generalización en la detección de violencia en vídeo.

Las hipótesis planteadas sirven como base para la formulación y desarrollo de los objetivos, los cuales buscan validar las premisas iniciales y explorar nuevas soluciones en esta Tesis Doctoral.

El **objetivo principal** de esta Tesis Doctoral es investigar, diseñar e implementar algoritmos basados en técnicas de inteligencia artificial explicable orientadas a la detección de violencia en vídeo, con foco en arquitecturas basadas en el uso conjunto de *Convolutional Neural Networks* (CNN) junto con capas *Long Short Term Memory* (LSTM). La aplicación de estas técnicas logrará obtener mejores resultados a los obtenidos por otros trabajos del estado del arte, y conseguirá un proceso de detección de violencia más fiable.

Para alcanzar el objetivo principal, es necesario definir un listado de **objetivos específicos**, que se describen a continuación:

(**OB1**) Analizar las técnicas y algoritmos empleados para la detección de violencia en vídeo mediante el uso de inteligencia artificial fiable, así como los retos actuales a los que se enfrenta.

(**OB2**) Diseñar arquitecturas de detección de violencia en vídeo mediante inteligencia artificial fiable, combinando CNN con características manuales y capas RNN, que mejore los métodos actuales y permita la detección de violencia en escenarios con condiciones cambiantes (escenario de alta visibilidad, escenario de alta densidad poblacional y escenario de baja visibilidad).

(**OB3**) Investigar el marco teórico y diseñar una arquitectura explicable para la detección de violencia en vídeo, integrando técnicas de XAI como GradCAM y desarrollando métodos específicos que evalúen y cuantifiquen la relevancia de los frames individuales dentro del vídeo.

(**OB4**) Implementar y demostrar la idoneidad de las arquitecturas de detección de violencia comparándolos entre sí y con otras arquitecturas propuestas en el estado del arte.

1.3. Metodología de investigación

De acuerdo con Baskerville (1999), en el contexto de una revisión sistemática, los estudios desarrollados bajo esta metodología comparten cuatro características comunes:

1. Orientación a la acción y al cambio.
2. Enfoque del problema.
3. Proceso orgánico que implica etapas sistemáticas y, a veces, iterativas.
4. Colaboración entre los participantes.

La metodología de investigación-acción, propuesta en la década de 1940, se enfoca en abordar problemas prácticos mediante la combinación de acción e investigación para generar conocimiento aplicable a contextos específicos. En general, esta metodología consta de cinco fases (Eden & Ackermann, 2018; Tüzün et al., 2019), que, para esta Tesis Doctoral, se describen a continuación:

1. **Diagnóstico:** esta fase se corresponde con la identificación y planteamiento del problema expuesto en esta Tesis Doctoral. Es necesario delimitar el alcance, proponer una hipótesis y definir los objetivos establecidos. A este respecto, se identifican y analizan los requisitos existentes en la detección de violencia en vídeo y el beneficio del uso de técnicas de XAI. Ello íncluye el análisis de retos actuales, motivaciones y problemas existentes a la hora de aplicar dichas técnicas, expuesto en el Capítulo 2.
2. **Revisión de la literatura:** esta fase realiza un estudio del estado del arte con el objeto de identificar las soluciones existentes al problema planteado y sentar las bases para el diseño de una solución innovadora. En este trabajo de investigación se incluye una revisión del estado del arte en la detección de violencia en vídeo

mediante inteligencia artificial, centrado especialmente en las agresiones físicas y en el uso de técnicas de inteligencia artificial fiable. Se recopilan en el Capítulo 2, un total de 63 artículos publicados entre junio de 2021 y junio de 2023.

3. **Solución:** esta fase propone una solución que responda a los objetivos establecidos a partir de los resultados obtenidos en las dos fases anteriores. En esta Tesis Doctoral se plantea el desarrollo de tres arquitecturas de detección de violencia en vídeo mediante la combinación de VGG-19 preentrenada así como características manuales, capas LSTM y capas Bi-LSTM. Además, a partir de estas arquitecturas se aplican técnicas de inteligencia artificial fiable para hacer el proceso de detección de violencia comprensible en todas sus fases. Por ejemplo, se propone el uso de GradCAM como algoritmo de explicabilidad para arquitecturas basadas en CNN, así como YoloV8 para la extracción de frames característicos.

4. **Evaluación:** esta fase lleva a cabo la evaluación de las arquitecturas propuestas. Para ello se entrenan los modelos a partir de una serie de datasets públicos de vídeos de violencia, estableciendo una serie de hiperparámetros para obtener modelos que maximizen la exactitud obtenida. Se lleva a cabo en las Secciones tituladas «Análisis y comparación de resultados», en los Capítulos 2, 3 y 4.

5. **Resultados:** Comprende la última fase del ciclo de investigación-acción, donde se analizan y comparan los resultados de los modelos implementados entre sí y con otros estudios publicados en el estado del arte actual que hayan desarrollado modelos con arquitecturas similares. Se lleva a cabo en las Secciones tituladas «Análisis y comparación de resultados», en los Capítulos 2, 3 y 4.

1.4. Estructura de la Tesis Doctoral

Esta Tesis Doctoral parte de un análisis de las revisiones con el objetivo de comprobar la hipótesis establecida para esta investigación y abordar los objetivos definidos, que parten de un **análisis de las revisiones actualizadas** en detección de violencia en vídeo mediante inteligencia artificial (OB1).

Tras ello, se desarrolla un **estudio sistemático de mapeo** de los artículos publicados recientemente que traten la detección de violencia en vídeo mediante inteligencia artificial

fiable. En este estudio se analizan los retos actuales así como los últimos avances y algoritmos utilizados en cada parte del proceso de detección de violencia (OB2).

Posteriormente se **diseñan, implementan y comparan** los resultados obtenidos con otros trabajos del estado del arte de **tres arquitecturas propuestas** mediante el uso combinado de la red VGG-19 preentrenada junto con características manuales, capas LSTM y capas Bi-LSTM (OB3 y OB5).

A partir de dichas arquitecturas se **diseña, implementa y exponen** los resultados obtenidos de una **arquitectura propuesta que utiliza técnicas de inteligencia artificial explicable** para aportar fiabilidad a todo el proceso de detección de violencia (OB4 y OB6), utilizando técnicas de explicabilidad como GradCAM enfocadas en CNN y un algoritmo de desarrollo propio para cuantificar el nivel de relevancia de un grupo de frames en la detección de violencia en vídeo.

De este modo, y tal como se ha adelantado, con el fin de facilitar el seguimiento de la investigación, se estructura la memoria de la presente Tesis Doctoral a través de seis Capítulos. El primero de ellos se corresponde con la actual **Introducción (Capítulo 1)**, en la que se describe el problema a resolver planteado, se formula la hipótesis de trabajo, se detallan los diferentes objetivos y se describe la metodología de investigación seguida para la culminación de esta Tesis Doctoral.

Los siguientes Capítulos describen cómo se estructura la memoria de la presente Tesis Doctoral para alcanzar los objetivos anteriores. Así, la Tesis Doctoral se estructura en 5 Capítulos adicionales que realizan una revisión sistemática de las técnicas de detección de violencia en vídeo (Capítulo 2), describen las arquitecturas de VGG-19 combinada con una característica manual (Capítulo 3), proponen la combinación de VGG-19 junto con capas LSTM y capas Bi-LSTM (Capítulo 4), diseñan una arquitectura explicable basada en el modelo descrito en el Capítulo 4 (Capítulo 5), y exponen las principales conclusiones y trabajo futuro (Capítulo 6). En concreto:

El **Capítulo 2** ofrece una revisión sistemática exhaustiva de las técnicas de detección de violencia en vídeo mediante TAI. Este Capítulo examina las revisiones recientes que destacan los avances y las limitaciones en el campo. Además, explora los retos que enfrenta esta tecnología, junto con un análisis detallado de las técnicas utilizadas en cada fase del proceso de detección.

El **Capítulo 3** describe la arquitectura, el entrenamiento y el testeo de la red VGG-19 preentrenada, así como de la combinación de VGG-19 preentrenada con una lógica manual que considera un vídeo como violento solo si un mínimo número de fotogramas han sido detectados como violentos. El objetivo es analizar la mejora que supone el uso de dicha lógica manual comparado con la combinación de capas LSTM de forma secuencial tras el uso de CNN.

El **Capítulo 4** expone la arquitectura, el entrenamiento y el testeo del modelo propuesto en esta Tesis Doctoral que consiste en la combinación de la red VGG-19 preentrenada junto con capas LSTM o capas Bi-LSTM. El objetivo de esta arquitectura es analizar si existe una cierta combinación de hiperparámetros que obtenga una mejor exactitud de forma sistemática, así como estudiar la mejora que suponen las capas Bi-LSTM respecto a las capas LSTM. Por último, el artículo compara los resultados de los modelos desarollados en el Capítulo 3 con los modelos desarrollados en este Capítulo.

El **Capítulo 5** expone una arquitectura explicable basada en el modelo descrito en el Capítulo 4. Se utiliza YoloV8 y la diferencia de fotogramas para realizar una selección explicable de fotogramas característicos, complementada con la implementación de GradCAM para generar mapas de saliencia que identifican la relevancia espacial en los fotogramas durante la detección de violencia. Finalmente, se desarrolla un algoritmo que cuantifica la importancia temporal de cada grupo de cinco fotogramas en la detección de violencia en el vídeo.

El **Capítulo 6** detalla las principales conclusiones que se han obtenido a partir del trabajo de investigación desarrollado en esta Tesis Doctoral y las aportaciones más relevantes realizadas. Además, define las líneas para el desarrollo de trabajo futuro que se abren a partir de los resultados de la presente Tesis Doctoral.

Finalmente, el **Apéndice A** presenta las publicaciones científicas resultantes del trabajo realizado en esta Tesis Doctoral, tanto en revistas de alto impacto, como en congresos especializados. Asimismo, se detallan las estancias llevadas a cabo en organismos de investigación internacionales realizadas durante las diversas etapas de las investigaciones de esta Tesis Doctoral.

Chapter 1

Introduction

Chapter 1

Introduction

Physical assaults represent a highly relevant social issue with multiple consequences. Various research efforts have been undertaken to address this problem from different perspectives and intervention levels. Among the potential solutions, real-time detection of violent acts in video emerges as a crucial response, acting as a final resource to identify victims of physical aggression. This advancement has been made possible by key factors such as the widespread use of surveillance cameras, the availability of large-scale data platforms, and the development of advanced algorithms applying artificial intelligence to image and video analysis.

This Doctoral Thesis proposes the development of video violence detection techniques through trustworthy artificial intelligence (TAI), with the aim of addressing certain aspects of the state of the art that have not yet been studied. For instance, whether there are optimal combinations of hyperparameters that yield greater accuracy, or whether it is possible to make the process of video violence detection more explainable. To this end, three architectures are developed based on the combination of *Convolutional Neural Networks* (CNN) together with: *(1)* handcrafted features, *(2) Long Short Term Memory* (LSTM) layers, and *(3) Bidirectional Long Short Term Memory* (Bi-LSTM) layers (Mumtaz et al., 2022b); respectively. Moreover, these architectures incorporate explainable artificial intelligence (XAI) techniques such as GradCAM and custom-developed algorithms (Abhishek & Kamath, 2022), in order to enhance the comprehensibility of the algorithm's decision-making process.

This introductory Chapter provides a review of the background related to the detection of violence in video using trustworthy artificial intelligence (TAI), the main topic of this Doctoral Thesis. Section 1.1 introduces the background of the study topic and outlines the problem to be addressed. Section 2.5 presents the working hypothesis, as well as the

objectives of the various research efforts that led to the development of this Doctoral Thesis. Section 1.3 details the methodology followed during these research projects. Finally, Section 1.3 describes the structure of this document.

1.1. Problem Description

This Section describes the problem addressed in this Doctoral Thesis: the detection of violence in video through accurate and trustworthy artificial intelligence. The Section is divided into two parts. Section 1.1.1 presents current data on physical assaults worldwide, as well as the various proposals that have been made to tackle them across different time horizons. Section 1.1.2 discusses the rising importance of trustworthy artificial intelligence and the fundamental pillars it comprises.

1.1.1. Violence in Society and Existing Solutions

Physical assaults affect virtually every aspect of life, not only for direct victims but also for families, society, and the daily development of countries in today's world (Muarifah et al., 2022). These assaults also impact the economy (shopping, travel, or tourism) and citizens' mental health due to a persistent sense of insecurity (Long et al., 2021). Some studies show alarming data, such as physical assaults being the leading cause of death among youth aged 15 to 44 globally during the first decade of the 21st century (Shubber & Al-Ta'i, 2022). Other studies report that 27% of women worldwide have experienced intimate partner violence (White et al., 2024). All of this underscores the global societal problem of physical aggression.

Multiple long-, medium-, and short-term solutions have been proposed. **Long-term solutions** focus on understanding the root causes or circumstances that exacerbate violence, so that future generations can be educated to prevent it (Long et al., 2021; Martínez-González et al., 2021). Some studies have examined the increase in violence due to exposure to violent and aggressive environments, leading to a tendency to use violence as a tool (Martínez-González et al., 2021). These studies also explore how maternal relationships and self-esteem levels are directly linked to aggression (Muarifah et al., 2022), or how technology influences violence due to smartphone use or violent video games (Fekih-Romdhane et al., 2022). Moreover, authors have observed that rapid

urbanization in 21st-century China, along with population growth, has led to increased crime (Long et al., 2021). Other studies discuss how COVID-19, and specifically its effects on daily life, have contributed to increased violence, creating a "perfect storm" of factors (Killgore et al., 2021).

Medium-term solutions involve addressing violence by correlating street population density and urban landscapes with crime rates. For example, some studies have used mobile phone and social media data to measure urban population density and correlate it with crime rates using spatial analysis techniques (Vomfell et al., 2018), though these face the limitation of being unable to distinguish between indoor and outdoor populations. Other recent research has shown crime reduction in areas with green spaces, using CNNs (Hipp et al., 2021). These approaches provide concrete insights into which areas are more prone to physical assaults and how to design new urban landscapes that discourage criminal behavior (Jing et al., 2021). However, this is an offline approach and does not protect victims during an assault. Therefore, **existing short-term solutions focus on real-time violence detection** (Jing et al., 2021).

Short-term solutions are centered on real-time violence detection. So far, surveillance cameras have largely been used as evidence for affected individuals when dealing with insurance companies, law enforcement, or the judiciary (Shukla & Pandey, 2020). However, several works have developed algorithms that enable **real-time violence detection through the use of surveillance cameras employing artificial intelligence (AI) algorithms** (Cheng et al., 2021b; Jaafar & Lachiri, 2023; Zhou et al., 2023). This is a growing field (Omarov et al., 2022) whose development has been driven by three main pillars: **the increased use of surveillance cameras** (Afra & Alhajj, 2020; Vosta & Yow, 2022), **the technological development of Big Data platforms** (Ageed & Zeebaree, 2021; Alonso et al., 2020), and **the implementation of artificial intelligence algorithms** (Collins et al., 2021) for image and video analysis (Ding et al., 2021).

Regarding the first pillar, surveillance camera installation has grown globally for monitoring and security purposes, among others (Afra & Alhajj, 2020; Vosta & Yow, 2022). Although video and image use offers many applications that improve citizens' quality of life, some studies warn of privacy risks due to large-scale recordings (Kostka et al., 2020). For the second pillar, Big Data platforms now allow the collection and

processing of vast environmental data (Ali et al., 2018; Alonso et al., 2020), enabling real-time analysis of images, videos, and other data types (Ageed & Zeebaree, 2021). As for the third pillar, the use of AI algorithms for image and video analysis has grown exponentially, not only in usage but also in variety and accuracy (Ding et al., 2021).

A key question is why AI is necessary for detecting physical assaults—central to this Doctoral Thesis. Human-based video analysis requires either high staffing costs (salaries, facilities) or results in low supervision levels when few people monitor many scenarios, potentially leading to loss of life (Mugunga et al., 2021a). Therefore, image and video analysis techniques play a critical role in the development of violence detection (Omarov et al., 2022).

In conclusion, physical aggression is a major societal concern with broad implications. Many research initiatives have aimed to tackle this issue through various approaches with different levels of directness. Real-time violence identification stands out as the most immediate solution, serving as the last line of defense for identifying individuals at risk (Cheng et al., 2021b; Jaafar & Lachiri, 2023; Zhou et al., 2023). This progress has only been possible due to the aforementioned three pillars: increased use of surveillance video, advancements in large-scale data platforms, and the evolution of AI-powered image and video analysis algorithms (Mugunga et al., 2021a). However, achieving useful results requires the use of trustworthy AI, as discussed in the following Section.

1.1.2. Trustworthy Artificial Intelligence

The quest for trustworthy artificial intelligence is not new. Since its inception, AI has evolved significantly—from simple deterministic algorithms to complex neural networks capable of learning intricate patterns from large datasets. In this context, the idea of "trustworthiness" in AI has become a key concern, especially in recent decades, as these technologies began to impact critical sectors such as health, justice, and public safety. This growing interest has led to interdisciplinary approaches aimed at ensuring that AI-based decisions are transparent, fair, and robust (Kaur et al., 2022).

The intersection between AI and video analysis has been a particularly prominent and expanding area. AI has proven useful in various applications—from traffic surveillance to real-time audiovisual content analysis. However, these capabilities have also raised

ethical and reliability concerns, since systems must operate consistently and without bias in highly dynamic contexts. Advanced technologies like convolutional neural networks and computer vision algorithms have improved object and activity recognition accuracy but have also highlighted the need for decision explainability (Chamola et al., 2023), particularly in sensitive settings like identifying suspicious or violent behaviors.

Violence detection using AI is one of the most critical and socially relevant applications. These tools are used in video surveillance systems to detect incidents in real time, which can be vital for preventing or mitigating dangerous situations. However, their implementation poses unique challenges. On one hand, the complexity of real-world situations can hinder algorithmic accuracy, especially in chaotic or poorly lit environments. On the other, there is the risk of bias that may discriminate against specific social groups, emphasizing the importance of systems that are not only robust but also explainable and ethically designed (Negre et al., 2024a).

Consequently, the growing use of AI-based systems has led major organizations to question how trustworthy these algorithms are (Kaur et al., 2022). Several institutions, such as the European Commission (European Commission, 2019) and the International Organization for Standardization (ISO) (ISO/IEC, 2020), have made efforts to define standards and legislation in this regard. Notably, the European Commission has published a report on artificial intelligence (European Commission, 2019) defining trustworthy AI. According to their definition, **trustworthy AI is built on three pillars**:

- *Lawfulness*: ensuring compliance with all applicable laws and regulations under any circumstances.
- *Ethics*: prioritizing principles and values that uphold fundamental moral and ethical norms.
- *Robustness*: ensuring that AI systems, even when well-intentioned, do not cause unintended harm, either technically or socially.

The Commission's report further states that each of these components is necessary, but not sufficient on its own, for AI to be trustworthy. The inherent opacity of most deep learning models—also known as the "black box" concept—makes their processes

and results difficult to interpret, and thus, their reliability questionable (Abhishek & Kamath, 2022). To increase this interpretability, the research community has extensively studied and discussed ***explainable artificial intelligence* (XAI)**, motivated by the need to understand and demystify AI system decisions (Mahmoudi et al., 2023; Speith, 2022; Swamy et al., 2022).

Lack of transparency in AI models can hinder their acceptance and adoption, especially in critical applications where understanding the reasoning behind automated decisions is vital (Mahmoudi et al., 2023). XAI specifically refers to the ability to understand and explain how an AI model makes decisions. Speith (2022) illustrated why aggression detection systems must use explainable AI, arguing that otherwise, there is a risk that an algorithm could make biased decisions, such as relying on a person's skin color. For these reasons, any system intended to detect violence must be explainable and, consequently, trustworthy—especially given its role in citizen protection (Chamola et al., 2023).

1.2. Hypothesis and Objectives

This Doctoral Thesis proposes, first, the **development of a mapping study** to analyze all stages in the process of violence detection in video, due to the need for a comprehensive and updated review of this field. Additionally, considering that the VGG-16 architecture has yielded excellent results in video violence detection (Negre et al., 2024b), and that the potential of VGG-19 has not yet been thoroughly explored in the state of the art, this Thesis proposes the **development of three architectures based on VGG-19 pretrained** on the *ImageNet* dataset (Deng et al., 2009). One architecture combines the pretrained VGG-19 with manual logic, whereby a video is considered violent only if a certain number of individual frames are classified as violent. The second architecture combines pretrained VGG-19 with LSTM layers. The third and final architecture integrates pretrained VGG-19 with Bi-LSTM layers.

Furthermore, due to the inherent opacity of AI models—and in line with the concerns raised by major bodies such as the European Commission (European Commission, 2019)—this Thesis also proposes the **development of a model incorporating explainable artificial intelligence techniques** based on the proposed architectures, to make the entire violence detection process more interpretable. YoloV8 and frame

differencing are used as intrinsically explainable algorithms for key frame extraction, GradCAM is employed as a CNN-focused explainability tool, and a custom algorithm quantifies the relevance of frame sets in the context of violence detection in video.

Based on the above, the following presents the main hypothesis and specific hypotheses of this Doctoral Thesis, along with the main objective and corresponding specific objectives.

The **general hypothesis** of this Doctoral Thesis is that combining Convolutional Neural Networks (CNN) with Long Short Term Memory (LSTM) layers within explainable artificial intelligence architectures will significantly improve the accuracy and reliability of violence detection in video.

(**HIP1**) To analyze the techniques and algorithms used in video violence detection will reveal that, although there are innovative approaches, current solutions still face limitations in terms of detection accuracy and reliability, thus justifying the need for more advanced architectures combining CNN and LSTM.

(**HIP2**) To design video violence detection architectures by combining CNN with handcrafted features, LSTM, and Bi-LSTM layers will lead to substantial improvements in capturing spatial and temporal patterns, achieving more accurate, explainable, and robust detection.

(**HIP3**) To create explainable architectures for violence detection in video will enable substantial understanding of the inference process, contributing not only to algorithm development but also to facilitating user comprehension.

(**HIP4**) To implement and compare the proposed architectures against other state-of-the-art solutions will demonstrate that the new CNN and LSTM-based architectures offer superior performance in terms of accuracy, reliability, and generalization capability in video violence detection.

The proposed hypotheses serve as the foundation for formulating and developing the objectives, which aim to validate the initial premises and explore new solutions:

1. **Diagnosis:** this phase corresponds to the identification and formulation of the problem addressed in this Doctoral Thesis. It is necessary to define the

scope, propose a hypothesis, and set the established objectives. In this regard, existing requirements for violence detection in video and the benefits of using XAI techniques are identified and analyzed. This includes the analysis of current challenges, motivations, and existing problems when applying these techniques, as presented in Chapter 2.

2. **Literature Review:** this phase involves a state-of-the-art study with the goal of identifying existing solutions to the stated problem and laying the groundwork for designing an innovative solution. This research includes a review of the state of the art in video violence detection using artificial intelligence, with a particular focus on physical aggression and the use of trustworthy AI techniques. A total of 63 articles published between June 2021 and June 2023 are collected in Chapter 2.

3. **Solution:** this phase proposes a solution that meets the established objectives based on the results obtained in the two previous phases. This Doctoral Thesis proposes the development of three video violence detection architectures through the combination of a pretrained VGG-19 network with handcrafted features, LSTM layers, and Bi-LSTM layers. Furthermore, trustworthy artificial intelligence techniques are applied to these architectures to make the violence detection process understandable in all its phases. For example, the use of GradCAM is proposed as an explainability algorithm for CNN-based architectures, as well as YoloV8 for extracting key frames.

4. **Evaluation:** this phase performs the evaluation of the proposed architectures. To do this, models are trained using a series of public video violence datasets, setting several hyperparameters to obtain models that maximize achieved accuracy. This is carried out in the Sections titled «Analysis and comparison of results», in Chapters 2, 3, and 4.

5. **Results:** this is the final phase of the action-research cycle, where the results of the implemented models are analyzed and compared with each other and with other studies from the current state of the art that have developed similar architectures. This is presented in the Sections titled «Analysis and comparison of results», in Chapters 2, 3, and 4.

1.3. Structure of the Doctoral Thesis

This Doctoral Thesis begins with an analysis of reviews with the objective of verifying the established hypothesis for this research and addressing the defined objectives, starting with an **analysis of updated reviews** in video violence detection using artificial intelligence (OB1).

Subsequently, a **systematic mapping study** is conducted on recently published articles dealing with video violence detection through trustworthy artificial intelligence. This study analyzes current challenges, as well as recent advancements and algorithms used in each part of the violence detection process (OB2).

Next, the **design, implementation, and comparison** of the results obtained with other works from the state of the art of **three proposed architectures** is carried out, using the pretrained VGG-19 network combined with handcrafted features, LSTM layers, and Bi-LSTM layers (OB3 and OB5).

From these architectures, an **explainable architecture using trustworthy artificial intelligence techniques** is **designed, implemented, and the results presented** to provide reliability throughout the violence detection process (OB4 and OB6), using explainability techniques such as GradCAM focused on CNNs and a proprietary algorithm to quantify the relevance level of a group of frames in video violence detection.

Thus, as previously mentioned, and to facilitate the follow-up of the research, the body of this Doctoral Thesis is structured into six Chapters. The first corresponds to the current **Introduction (Chapter 1)**, which describes the problem to be solved, formulates the working hypothesis, details the various objectives, and outlines the research methodology followed to complete this Doctoral Thesis.

The following Chapters describe how the body of this Doctoral Thesis is structured to achieve the aforementioned objectives. Thus, the Doctoral Thesis is structured into 5 additional Chapters that conduct a systematic review of video violence detection techniques (Chapter 2), describe the VGG-19 architecture combined with a handcrafted feature (Chapter 3), propose the combination of VGG-19 with LSTM and Bi-LSTM layers (Chapter 4), design an explainable architecture based on the model described in

Chapter 4 (Chapter 5), and present the main conclusions and future work (Chapter 6). Specifically:

Chapter 2 offers an exhaustive systematic review of video violence detection techniques using TAI. This Chapter examines recent reviews highlighting advances and limitations in the field. Additionally, it explores the challenges faced by this technology, along with a detailed analysis of the techniques used in each phase of the detection process.

Chapter 3 describes the architecture, training, and testing of the pretrained VGG-19 network, as well as the combination of VGG-19 with a handcrafted logic that considers a video violent only if a minimum number of frames are detected as violent. The goal is to analyze the improvement brought by this handcrafted logic compared to the sequential combination of LSTM layers after using CNN.

Chapter 4 presents the architecture, training, and testing of the model proposed in this Doctoral Thesis, which consists of the combination of the pretrained VGG-19 network with LSTM or Bi-LSTM layers. The aim of this architecture is to analyze whether a certain combination of hyperparameters can systematically achieve better accuracy, as well as to study the improvement provided by Bi-LSTM layers over LSTM layers. Lastly, the Chapter compares the results of the models developed in Chapter 3 with those developed in this Chapter.

Chapter 5 presents an explainable architecture based on the model described in Chapter 4. YoloV8 and frame difference techniques are used to perform an explainable selection of key frames, complemented with the implementation of GradCAM to generate saliency maps that identify spatial relevance in frames during violence detection. Finally, an algorithm is developed to quantify the temporal importance of each group of five frames in video violence detection.

Chapter 6 details the main conclusions derived from the research work developed in this Doctoral Thesis and the most relevant contributions made. It also defines future research directions stemming from the results of this Doctoral Thesis.

Finally, **Appendix A** presents the scientific publications resulting from the work conducted in this Doctoral Thesis, both in high-impact journals and specialized conferences. It also details the research stays carried out in international research institutions during the various stages of this Doctoral Thesis.

Capítulo 2

Técnicas de detección de violencia en vídeo mediante IA

Capítulo 2

Técnicas de detección de violencia en vídeo mediante IA

Índice

Las agresiones físicas tienen un amplio impacto en diversos aspectos de la vida, afectando no solo a las personas directamente involucradas, sino también a sus familias, comunidades y la economía (Long et al., 2021; Muarifah et al., 2022). Ciertos estudios muestran que la violencia física es una causa principal de mortalidad entre los jóvenes en todo el mundo (Shubber & Al-Ta'i, 2022). Además, investigaciones como las de Hillis et al. (2016) señalan que una gran cantidad de niños en diferentes regiones experimentan violencia. Por otra parte, el estudio desarrollado por FRA (2023) expuso cómo, en Europa, una parte significativa de la población fue víctima de acoso o agresiones físicas, en base a 35.000 personas encuestadas entre enero y junio de 2019 FRA (2023). Ello subraya la presencia y gravedad del problema de la violencia física a nivel global.

Para abordar el problema que suponen las agresiones físicas, existen **soluciones a largo, medio y corto plazo**. Las soluciones a largo plazo incluyen entender las causas subyacentes de la violencia para educar a las futuras generaciones y prevenir su ocurrencia (Fekih-Romdhane et al., 2022; Martínez-González et al., 2021; Muarifah et al., 2022). Factores como la rápida urbanización y el impacto de eventos como la pandemia de COVID-19 también contribuyen a este aumento de la violencia (Killgore et al., 2021; Long et al., 2021).Las soluciones a corto plazo buscan estrategias para prevenir actos de violencia física, como la detección de violencia en tiempo real (Hipp et al., 2021; Jing et al., 2021; Vomfell et al., 2018; Yue et al., 2022). Sin embargo, la detección de violencia en tiempo real ha surgido como una estrategia crucial a corto plazo para prevenir actos de violencia física (Cheng et al., 2021b; Jaafar & Lachiri, 2023; Shukla & Pandey, 2020; Zhou et al., 2023). Entre otras soluciones, destaca el uso de algoritmos de IA para analizar imágenes y vídeos de seguridad, que ha sido posible gracias a la evolución de la tecnología de cámaras, procesadores y algoritmos (Afra & Alhajj, 2020; Ageed & Zeebaree, 2021; Collins et al., 2021).

Este Capítulo lleva a cabo un **estudio de las técnicas existentes** para la detección de violencia en vídeo mediante el uso de IA, por lo que desarrolla un mapeo sistemático de la literatura, realizando un análisis exhaustivo de 63 trabajos seleccionados publicados entre junio de 2020 y junio de 2023. El Capítulo analiza estudios abordando todos los procesos que conlleva la detección de violencia tales como 21 retos actuales asociados con la detección de violencia, 28 *datasets* públicos sobre vídeos de violencia, 13 métodos de métricas de evaluación, entre otros. Finalmente, evidencia que el uso de algoritmos basados en inteligencia artificial fiable en la detección de violencia aún es limitado, si bien algunos trabajos utilizan algoritmos que pueden aportar cierta explicabilidad al proceso, aun no siendo este el objetivo de estos.

Este Capítulo, por lo tanto, realiza un mapeo sistemático a lo largo de 9 secciones. La Sección 2.1.1 presenta un análisis de las revisiones recientes de detección de violencia mediante inteligencia artificial y expone la metodología implementada para la realización del análisis sistemático de la literatura de detección de violencia en vídeo mediante IA. La Sección 2.2 aborda las principales etapas que constituyen los algoritmos de detección de violencia analiza las clasificaciones propuestas en los trabajos seleccionados sobre este tema. Además, desarrolla los desafíos y limitaciones presentes en la detección de violencia en vídeo mediante IA, así como describe las métricas de evaluación utilizadas para cuantificar la bondad de las clasificaciones realizadas por los algoritmos. La Sección 2.3 analiza los datasets, las técnicas de selección de frames característicos y los métodos de preprocesados empleados en la detección de violencia en vídeo mediante IA. La Sección analiza los algoritmos de detección de violencia en vídeo y su exactitud en los principales conjuntos de datos. La Sección 2.5 analiza si los algoritmos implementados en los trabajos seleccionados emplean técnicas de inteligencia artificial fiable o, en su defecto, técnicas basadas en alguno de sus pilares. La Sección 2.6 expone las conclusiones obtenidas durante el presente Capítulo.

2.1. Punto de partida

Esta Sección se divide en dos Secciones. La Sección 2.1.1 presenta un resumen de los estudios de revisión previamente publicados sobre la detección de violencia en vídeo mediante inteligencia artificial. La Sección 2.1.2 describe la metodología utilizada en esta Tesis Doctoral para llevar a cabo el estudio sistemático de la literatura.

2.1.1. Análisis de las revisiones existentes

Esta Sección realiza un resumen de los trabajos de revisión ya publicados que analizan la detección de violencia en vídeo utilizando IA. Este resumen se realiza para estudiar cuánto se ha estudiado sobre la IA fiable, evidenciando que es necesario investigar en este área, comenzando por realizar un nuevo mapeo. Ninguno de estos trabajos hace referencia al uso de inteligencia artificial fiable. La Sección incluye las revisiones publicadas entre junio de 2021 y julio de 2023, y muestra los contenidos de las revisiones existentes. Los contenidos de las revisiones existentes se muestran categorizados en la Tabla 2.1 y en la Tabla 2.2.

Yao y Hu (2023) reconocen en la introducción de su investigación el desafío de definir la violencia, destacando su ambigüedad inherente al distinguirla de otras acciones, y señalan su ocurrencia poco frecuente, lo que dificulta obtener grabaciones de casos reales. También discuten la disparidad entre los métodos convencionales (con menor *accuracy*) y las técnicas de aprendizaje profundo, para detectar la agresión. Sin embargo, no proporcionan información sobre las bases de datos utilizadas, los criterios de búsqueda, el período de tiempo y los filtros de selección. Además, no hacen referencia a trabajos relacionados. Su revisión categoriza los trabajos de vanguardia en dos grupos: el marco tradicional y el aprendizaje profundo. Además, analiza el tipo de extracción de características empleado en los trabajos seleccionados, como por ejemplo: el movimiento, la apariencia, basadas en trayectoria y descriptores de características. También estudia los clasificadores empleados en la detección de violencia, donde son ampliamente utilizados SVM (Support Vector Machine) y KNN (K-Nearest Neighbors).

Siddique et al. (2022) proporcionan una visión general de la detección de violencia, delineando los pasos y conceptos básicos junto con un esquema visual y una tabla resumen. Si bien no realizan una revisión sistemática, ofrecen una metodología de investigación integral, que incluye consultas de búsqueda, bases de datos, rangos de fechas y procesos de filtrado de trabajos. Esta metodología categoriza los algoritmos de detección de agresiones en 19 subgrupos, si bien varios de estos subgrupos tienen solo un trabajo dentro de ellos, lo que sugiere la necesidad de una agrupación más amplia para mejorar la claridad. El trabajo no aporta los resultados obtenidos por los algoritmos citados, el tipo de clasificador empleado en cada arquitectura, ni el dataset con el que se realiza el proceso de entrenamiento y testeo. Por otro lado, no analiza qué

tipo de extracción de características es la más frecuentemente utilizada, además de no investigar qué tipo de algoritmos tienden a obtener mejores resultados. Por último, de los datasets recopilados, no se especifica más información que el número de clips y el año de publicación.

Kaur y Singh (2022) presentan un estudio que resume revisiones recientes del estado del arte en el campo de la detección de agresiones desde 2016. Los autores categorizan los trabajos revisados en dos grupos principales: aquellos que emplean métodos tradicionales, con subcategorías adicionales basadas en las técnicas de extracción de características y clasificación empleadas. Aquellos que emplean métodos de aprendizaje profundo. El documento hace especial hincapié en varios estudios que aprovechan las redes neuronales convolucionales y las redes neuronales *Long Short Term Memory*, destacando estas como las técnicas más recientes y prometedoras en el campo. También aborda varios desafíos encontrados en la detección de agresiones, incluidos problemas como cambios en la iluminación, objetos superpuestos y confusión con otros tipos de acciones. La revisión concluye que no parece haber una solución universal para estos problemas como colectivo, sino que deben abordarse individualmente.

Shubber y Al-Ta'i (2022) categorizan los algoritmos de detección de violencia en dos grupos principales: métodos tradicionales y métodos de aprendizaje profundo, siendo estos últimos los que demuestran una mayor exactitud. Se exponen las divisiones realizadas en extracción de características y clasificadores. Tambiénproporcionan una tabla que detalla los diversos métodos de extracción de características utilizados en diferentes trabajos y presentan cuatro métodos de clasificación, citando los estudios que los han aplicado. Por otro lado, enfatizan que el desarrollo de técnicas de aprendizaje profundo ha sido posible gracias a la disponibilidad sustancial de datos y alta capacidad de procesamiento. Además, citan los datasets utilizados en los trabajos revisados, especificando el número de clips de vídeo, el número de fotogramas y el contenido de estos datasets (por ejemplo, vídeos de hockey, películas, escenas reales).

Omarov et al. (2022) presentan la única revisión sistemática que analiza 80 trabajos sobre detección de agresiones en vídeo mediante el uso de inteligencia artificial publicados entre 2015 y 2021. Los autores detallan elementos de revisión sistemática como preguntas de investigación, definición de trabajos y pasos de filtrado. Además, ilustran la creciente relevancia de la detección de agresiones y el creciente uso de algoritmos de aprendizaje

profundo. Por otro lado, estos autores categorizan los trabajos seleccionados entre aquellos que emplean técnicas de aprendizaje automático y aprendizaje profundo, donde en ambos casos se analiza el tipo de técnica de extracción de características y clasificador utilizados.

La Tabla 2.1 y la Tabla 2.2 contienen los contenidos de las revisiones analizadas en esta Sección. Estas Tablas muestran cómo, en general, las revisiones no abordan todas las partes de la detección de violencia. De entre todas las revisiones expuestas, el artículo de revisión más detallado, que además es la única revisión sistemática entre la literatura identificada, es el de Omarov et al. (2022). Sin embargo, los trabajos recopilados son de entre 2015 y 2021 y ninguno de los análisis identificados menciona el uso de inteligencia artificial explicable (XAI) o la importancia de la fiabilidad en la detección de agresiones físicas. Por lo tanto, **se considera relevante y necesario producir un SMS actualizado con trabajos recientes y contenido integral sobre todo el campo de la detección de violencia en vídeo mediante IA**.

En resumen, las revisiones analizadas que han sido publicadas durante los últimos tres años no cubren todos los aspectos de la detección de violencia. Por lo tanto, existe una necesidad de investigar en el ámbito de la inteligencia artificial para la detección de violencia en vídeo, especialmente desde el punto de vista de la AI explicable. Para ello, se parte por hacer una revisión sistemática, que además, como se ha evidenciado en esta Sección, no existe ninguna que cubra estas tecnologías.

Tabla 2.1: Resumen de trabajos de revisión relacionados que abordan la agresión física. Parte 1.

Referencia	**Título**	**Año**	**SMS**	**Trabajo relacionado**	**Metodología búsqueda**	**Retos**	**Parámetros Evaluación**	Dataset						
								Nombre	**Ref.**	**Año**	**Número vídeos**	**Tipo**	**Calidad**	**Citas**
(Yao & Hu, 2023)	A survey of vídeo violence detection.	2021				✓		✓		✓	✓	✓		
(Siddique et al., 2022)	State-of-the-Art Violence Detection Techniques: A review.	2022			✓			✓	✓	✓			✓	
(Kaur & Singh, 2022)	Violence Detection in vídeos Using Deep Learning: A Survey.	2022		✓		✓								
(Shubber & Al-Ta'i, 2022)	A review on vídeo violence detection approaches.	2022				✓		✓		✓	✓	✓	✓	
(Omarov et al., 2022)	State-of-the-art violence detection techniques in vídeo surveillance security systems: a systematic review.	2022	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓

Tabla 2.2: Resumen de trabajos de revisión relacionados que abordan la agresión física. Parte 2.

Referencia	Métodos Tradicionales					Métodos basados en *Deep Learning*				
	Preprocesado	**Detección de Objetos**	**Extracción de Características**	**Clasificación**	**Dataset**	**Preprocesado**	**Detección de Objetos**	**Extracción de Características**	**Clasificación**	**Dataset**
(Yao & Hu, 2023)			✓	✓				✓	✓	
(Siddique et al., 2022)		✓	✓				✓	✓		
(Kaur & Singh, 2022)			✓					✓	✓	✓
(Shubber & Al-Ta'i, 2022)	✓		✓		✓	✓		✓		✓
(Omarov et al., 2022)		✓	✓	✓	✓		✓	✓	✓	✓

2.1.2. Metodología para un análisis sistemático

Esta Sección expone la metodología empleada para la realización del estudio sistemático de la literatura de esta Tesis Doctoral. La Sección 2.1.2.1 describe el plan de la investigación, el cual detalla la motivación y relevancia del estudio, sus objetivos y las preguntas de la investigación. La Sección 2.1.2.2 trata el desarrollo del estudio, abordando la estrategia de búsqueda y los criterios de inclusión de los trabajos seleccionados. La Sección 2.1.2.3 presenta un informe del mapeo realizado durante el estudio, el cual incluye el estudio de filtrado de los trabajos seleccionados, el proceso de clasificación de estos y la validación del proceso de mapeo.

2.1.2.1. Planificación de la investigación

Esta Sección expone en qué consiste un SMS y la relevancia de estos, ya que han ganado importancia en los últimos años debido a su contribución vital al campo de la investigación científica (Petersen et al., 2008). Además, establece la motivación, la relevancia, los objetivos y las preguntas de investigación.

Un **SMS** es una metodología de investigación utilizada para clasificar, analizar e identificar un campo de interés. El objetivo general de un SMS es establecer el alcance de estudio de un campo de investigación dado mediante el análisis de la frecuencia de publicaciones según diversas categorías de manera esquemática.

Kitchenham et al. (2011) proporcionan una visión general de las prácticas de revisión sistemática en ingeniería de software, identificando desafíos comunes y ofreciendo recomendaciones para mejorar la calidad de las revisiones sistemáticas. La comprensión mejorada de las prácticas de revisión sistemática ha contribuido a una mayor adopción de SMS como enfoque inicial para mapear y clasificar el conocimiento existente en áreas de investigación específicas. Por otro lado, Petersen et al. (2015) establecen pautas metodológicas específicas para llevar a cabo SMS, lo que ha permitido una mayor estandarización y rigor en la realización de estos estudios. Los objetivos particulares de un SMS son los siguientes (Petersen et al., 2008):

- Identificar las principales áreas de investigación y temas relevantes en un campo específico y clasificar las publicaciones científicas existentes sobre el tema de estudio.
- Examinar la distribución temporal de la investigación y su evolución a lo largo del tiempo.
- Identificar las metodologías y enfoques más comúnmente utilizados en la investigación sobre el tema.
- Descubrir posibles lagunas en el conocimiento y áreas poco investigadas.
- Proporcionar una visión general del estado actual del conocimiento en el campo de estudio.
- Detectar posibles tendencias y temas emergentes que podrían ser objeto de futuras investigaciones.

La **motivación** del SMS presentado en esta Tesis Doctoral (Negre et al., 2024b) consiste en compilar los últimos estudios sobre el uso de técnicas de inteligencia artificial en la detección de agresiones físicas en vídeo. Se pretende, además, analizar si estos trabajos abordan el uso de inteligencia artificial fiable o alguno de los pilares en los que esta se fundamenta como la inteligencia artificial lícita, ética o robusta. Ello es relevante de cara a garantizar que las tecnologías desarrolladas no solo sean efectivas, sino también seguras y justas, promoviendo así su confianza por parte del público y de las instituciones.

La *relevancia* de esta revisión sistemática radica en el hecho de que solo ha habido 5 revisiones sobre la detección de agresión física en vídeo mediante el uso de inteligencia artificial en los últimos dos años hasta donde llega el conocimiento del autor de esta Tesis Doctoral (Kaur & Singh, 2022; Omarov et al., 2022; Shubber & Al-Ta'i, 2022; Siddique et al., 2022; Yao & Hu, 2023), de las cuales solo una es un SMS (Omarov et al., 2022). Además, ninguna de las revisiones sistemáticas aborda la fiabilidad de los algoritmos estudiados o su explicabilidad. Por todas estas razones, este SMS se considera relevante para evaluar los últimos estudios publicados en el creciente campo de la detección de agresión física en vídeo (Omarov et al., 2022), con un enfoque especial en la aplicación de otro campo en crecimiento, a saber, el de la inteligencia artificial fiable y explicable (European Commission, 2019; ISO/IEC, 2020).

En base a ello, el **objetivo** de este SMS es compilar los últimos estudios que abordan la detección de agresión física en vídeo, con un énfasis especial en el uso o mención de inteligencia artificial fiable y/o explicable. Esta revisión sistemática proporciona una perspectiva actualizada sobre los avances en este campo de investigación. Por otra parte, esta investigación compila y describe en detalle los datasets que se utilizan, así como los algoritmos utilizados en cada parte de la detección de agresión. En base a ello se pretende conocer qué combinación de técnicas es la más óptima en cada caso. Finalmente, este SMS intenta identificar qué técnicas de inteligencia artificial fiable y explicable se mencionan o utilizan en los trabajos seleccionados.

A partir del objetivo establecido, las **preguntas de investigación** guían y enfocan la investigación, permitiendo identificar y clasificar estudios relevantes para mapear y comprender el estado actual del conocimiento en un campo de investigación específico. Se seleccionan dos preguntas clave de investigación, que se presentan a continuación:

- ¿Cuáles son las últimas técnicas de vanguardia aplicadas en la detección de agresión física en vídeo?
- ¿Qué técnicas de inteligencia artificial fiable y/o explicable se utilizan o discuten en la literatura actual sobre detección de agresiones?

Con todo, se analiza en esta Sección en qué consiste un SMS y su relevancia. Además, se establece la motivación, objetivos y preguntas de investigación del estudio.

2.1.2.2. Desarrollo del estudio

Esta Sección establece la *Estrategia de Búsqueda* y los *Criterios de Inclusión/Exclusión* de este SMS. Por un lado, la estrategia de búsqueda describe cómo se recopila la información relevante, utilizando bases de datos específicas y términos cuidadosamente seleccionados. Por otro, los criterios de inclusión/exclusión establecidos sirven como pautas para seleccionar estudios relevantes y descartar aquellos inadecuados, asegurando así la calidad y fiabilidad de los resultados obtenidos en este SMS.

Estrategia de Búsqueda

Una parte clave de un SMS es el desarrollo de **la estrategia de búsqueda**. Tanto Kitchenham et al. (2011) como Petersen et al. (2015) proponen la estrategia PICO

(Población, Intervención, Comparación y Resultados) para llevar a cabo el SMS. Esta estrategia consiste en construir cadenas de búsqueda a partir de palabras clave relevantes en las preguntas de investigación formuladas. A continuación, se define qué significan los términos que componen PICO en este SMS.

- Población: en este SMS, la población comprende la colección de los documentos de investigación que se están analizando.
- Intervención: se refiere a las técnicas, modelos o soluciones que tratan sobre la detección de agresión física en vídeo.
- Comparación: la investigación seleccionada se contrasta según diferentes categorías.
- Resultados: no aplicable a este SMS.

Dado que existen numerosos términos alternativos para la detección de agresión física, la cadena de búsqueda creada es lo más completa posible, incluyendo todas las terminologías posibles de estudios que tratan este tema. Se utiliza la cadena de búsqueda empleada por Siddique et al. (2022) como punto de partida, la cual se muestra en el Listado 2.1, ya que es muy completa y pertenece a uno de los dos únicos SMS publicados en los últimos dos años. En esta, los asteriscos `(*)` denotan los derivados y variantes de términos clave, lo que permite una búsqueda más completa y exhaustiva.

Listado 2.1: Cadena de búsqueda del trabajo de Siddique et al. (2022).

```
(violen* OR fight* OR anomal*) AND
(activity OR event OR scene OR sequence) AND
(detect* OR recogni* AND from) AND
(surveillance OR cctv*) AND
(vi* OR motion) AND
(using OR through OR by OR via) AND
(machine learning OR computer vision OR deep learning)
OR artificial intelligence)
```

Se realizan cambios en la cadena de búsqueda mostrada en el Listado 2.1, los cuales se exponen en el Listado 2.2, y mejoran los criterios de búsqueda y la hacen más específica para las particularidades de este estudio. El término `anomal*` se reemplaza por `physical aggression` y `physical attack`, incluyendo la palabra `physical`

Tabla 2.3: Distribución de los términos de las cadenas de búsqueda.

Descripción Acción	*1*	physical aggression	physical attack	violence	fight
	2	activity	event	scene	sequence
	3	detect*	recogni*		
Medio	*4*	surveillance	cctv		
	5	vi*	motion	camera	
Tecnología	*6*	artificial intelligence, AI	machine learning, ML	deep learning, DL	computer vision, CV
	7	trustworth*	explainab*, XAI	ethic*	robust*

para que no pueda confundirse con otro tipo de agresiones. También se eliminan los conectores `from`, `using`, `through`, `by` y `via`, ya que no agregan valor a la búsqueda y pueden limitar resultados relevantes. Por otro lado, se agrega el término `camera` a los términos `vídeo` o `motion`, además de los acrónimos `artificial intelligence (IA)`, `machine learning (ML)`, `deep learning (DL)` y `computer vision (CV)` para ampliar más la cobertura de búsqueda. Finalmente, se añade una operador `AND` al final de la cadena con términos relacionados con técnicas de IA fiable y/o explicable: `trustworth*/explainab*/XAI/ethic*/robust*`.

Listado 2.2: Cadena de búsqueda ajustada para la selección de trabajos en el SMS.

```
(physical aggression OR physical attack OR
violence OR fight) AND
(activity OR event OR scene OR sequence) AND
(detect* OR recogni*) AND
(surveillance OR CCTV) AND
(vi* OR motion OR camera) AND
(artificial intelligence OR AI OR machine learning OR ML OR
deep learning OR DL OR computer vision OR CV) AND
(trustworth* OR explainab* OR XAI OR ethic* OR robust*)
```

La Tabla 2.3 permite visualizar y analizar fácilmente cómo se distribuyen los términos de la cadena de búsqueda en las diferentes categorías, cada término AND en la cadena de búsqueda se encuentra en una fila diferente. Además, los términos AND se agrupan en tres grupos principales: «Descripción de la acción» agrupa los primeros tres elementos, «Medio» agrupa los elementos 4-5 y «Tecnología» agrupa los elementos 6-7.

Tabla 2.4: Palabras clave posibles palabras derivadas de los términos incluidos en la cadena de búsqueda completa del estudio de mapeo sistemático.

Palabras clave	**Posibles palabras derivadas**
detect*	detect, detection, detectable, etc.
recogni*	recognition, recognizing, recognizer, etc.
vi*	vídeo, visual, visualization, etc.
trustworth*	trustworthy, trustworthiness, trustworthily
explainab*	explainable, explainability, explainably
ethic*	ethical, ethics, ethically, etc.
robust*	robust, robustness, robustly, etc.

Como ya se ha comentado, el uso de asteriscos para englobar las palabras derivadas de los términos clave permite realizar una búsqueda más completa y exhaustiva. En la Tabla 2.4 se muestran las palabras utilizadas en las cadenas de búsqueda que contienen asteriscos junto con algunas de sus posibles palabras derivadas.

Se seleccionan tres bases de datos para la búsqueda de artículos: *Web of Science*, *Scopus* y *Springer*. Estas son opciones adecuadas por su amplia cobertura interdisciplinar, contenido revisado por pares, enfoque en áreas relevantes y acceso a investigaciones actualizadas, lo que permite una revisión sólida y exhaustiva de la literatura. La elección de la cadena de texto adecuada es un reto, ya que la búsqueda debe contener un número de trabajos que no sea ni demasiado bajo ni demasiado alto. La cadena de búsqueda que se muestra en la Tabla 2.4 es muy completa y abarca todos los términos que definen este tema de investigación, pero ciertos términos pueden restringir demasiado la búsqueda o introducir trabajos que corresponden a otro tema de estudio.

En definitiva, se trata de un proceso iterativo de ensayo y error. Tras probar múltiples combinaciones, añadiendo y eliminando elementos, se decide eliminar dos cláusulas. El primero es la inclusión de los elementos `surveillance` o `CCTV`, ya que se observa que limitan la búsqueda de trabajos relacionados con la detección de la violencia que simplemente no utilizaban estos términos. Por otro lado, la inclusión de los términos relacionados con la inteligencia artificial fiable (`trustworth*` or `explainab*` or `XAI` or `ethic*` or `robust*`) obtuvo resultados muy bajos. Como conclusión, se expone la cadena definitiva empleada para este SMS tras la comprobación experimental de cómo los términos afectaban a los resultados obtenidos en el Listado 2.3.

Listado 2.3: Cadena de búsqueda definitiva ajustada al propósito de este SMS para la selección de trabajos tras comprobar la mejor combinación de los términos empleados.

```
(physical aggression OR physical attack OR
violence OR fight) AND
(activity OR event OR scene OR sequence) AND
(detect* OR recogni*) AND
(vi* OR motion OR camera) AND
(artificial intelligence OR AI OR machine learning OR ML OR
deep learning OR DL OR computer vision OR CV)
```

En el caso de las bases de datos de *Scopus* y *Web of Science*, se realiza la búsqueda sobre el *abstract*. Sin embargo, en el caso de la base de datos de *Springer*, esto no es posible debido a que el sistema de búsqueda no lo permite, por lo que la búsqueda se realiza para el texto completo. Si bien ello aumenta significativamente el número de trabajos a descartar, *Springer* se trata de una base de datos con un alto número de trabajos publicados, por lo que se considera relevante su inclusión en este estudio.

- **Base de datos *Scopus***: el listado 2.4 contiene la cadena de búsqueda específica para la base de datos de *Scopus*.

Listado 2.4: Cadena de búsqueda en *Scopus* para la selección de trabajos.

```
ABS(
((physical AND aggression) OR (physical AND attack) OR violence OR fight)
AND
(activity OR event OR scene OR sequence) AND
(detect* OR recogni*) AND
(vi* OR motion OR camera) AND
((artificial AND intelligence) OR (ai) OR (machine AND learning) OR (ml)
OR
(deep AND learning) OR (dl) OR (computer AND vision) OR (cv))
)
```

- **Base de datos *Web of Science***: el listado 2.5 contiene la cadena de búsqueda específica para la base de datos de *Web of Science*.

Listado 2.5: Cadena de búsqueda en *Web of Science* para la selección de trabajos.

```
AB = (
(((physical) AND (aggression)) OR ((physical) AND (attack)) OR
(violence) OR (fight)) AND
```

```
((activity) OR (event) OR (scene) OR (sequence)) AND ((detect*) OR
(recogni*)) AND
((vi*) OR (motion) OR (camera)) AND
(((artificial) AND (intelligence)) OR (ai) OR ((machine) AND (learning))
OR (ml) OR ((deep) AND (learning)) OR (dl) OR ((computer) AND (vision))
OR (cv))
)
```

- **Base de datos *Springer***: el listado 2.6 contiene la cadena de búsqueda específica para la base de datos de *Springer*.

Listado 2.6: Cadena de búsqueda en *Springer* para la selección de trabajos.

```
("physical aggression" OR "physical attack" OR "violence" OR
"fight") AND
("activity" OR "event" OR "scene" OR "sequence") AND
("detect*" OR "recogni*") AND
("vi*" OR "motion" OR "camera") AND
("artificial intelligence" OR "ai" OR "machine learning" OR
"ml" OR "deep learning" OR "dl" OR "computer vision" OR "cv ")
```

Criterios de inclusión/exclusión

Los criterios de inclusión/exclusión para los trabajos recopilados resultan de gran importancia, ya que garantiza la calidad de los resultados obtenidos en el SMS (Kitchenham et al., 2011). En este caso, se definen criterios de exclusión para filtrar los documentos que habían sido seleccionados a través de las cadenas de búsqueda creadas.

- Artículos que no aparecen como resultado de aplicar la cadena de búsqueda seleccionada, buscando en el resumen.
- Documentos que no sean artículos de revistas o ponencias de conferencias.
- Documentos que no hayan sido publicados dentro del rango de fechas definido, desde junio de 2020 hasta junio de 2023.
- Documentos que no estén escritos en inglés.
- Documentos duplicados.
- Documentos que no aborden la detección de agresión física en vídeo.

2.1.2.3. Informe de mapeo

El informe de mapeo implica reflejar el proceso seguido en la selección de trabajos y su clasificación. Por lo tanto, se divide en tres partes diferenciadas: estudios de filtrado, proceso de clasificación y validación del estudio de mapeo sistemático.

Estudios de filtrado

A partir de los articulos encontrados al emplear las cadenas de búsqueda expuestas en la Sección 2.1.2.2 para cada base de datos, se aplican los criterios de exclusión enumerados también en dicha sesión. Como se ha mencionado anteriormente, la base de datos de *Springer* no permite filtrar por resumen. Por lo tanto, los resultados se filtran según el cuerpo completo del artículo. Aunque esta decisión implica analizar un número mucho mayor de trabajos, se procede con la búsqueda en esta base de datos.

Es importante señalar que el término «Detección de violencia» es ambiguo y puede incluir: tiroteos, vandalismo, robos, entre otros. Por ello, en esta Tesis Doctoral, la detección de violencia se refiere específicamente a las agresiones físicas cuerpo a cuerpo y sin el uso de armas. Se seleccionan aquellos trabajos que han utilizado al menos un dataset centrado en la detección de agresión física, en lugar de en la violencia en general (armas de fuego, explosiones, etc.). En total, después de la búsqueda inicial, se recuperan un total de 13.137 artículos de revistas y conferencias (393 de *Scopus*, 240 de *Web of Science* y 12,504 de *Springer*); se seleccionan un total de 63 artículos siguiendo la aplicación de los criterios de exclusión.

La Tabla 2.5 muestra el número de trabajos seleccionados después de aplicar cada criterio de exclusión; la primera columna desde la izquierda es el primer filtro aplicado y la última columna desde la izquierda es el último filtro aplicado. Cada base de datos se divide en dos: *A* para artículos de revistas y *C* para artículos de conferencias. La columna *Artículos con dudas* contiene el número de trabajos seleccionados después de decidir excluir ciertos artículos por diversos motivos, entre ellos: no estar centrados en la detección de la violencia por agresión física, no explicar de forma clara la arquitectura implementada, no aparecer información relevante que deba incluirse en este estudio o no tener la rigurosidad suficiente para la inclusión en este trabajo.

Una vez completada la selección de los trabajos, se lleva a cabo el proceso de clasificación. De las obras seleccionadas, 27 son artículos de revistas y 36 son ponencias de conferencias.

Tabla 2.5: Número de trabajos seleccionados tras cada criterio de exclusión.

		Resultados búsqueda	Tipo Artículo	Año & Mes	Agresión Física	Artículos con Dudas
Scopus	**A**	393	162	114	23	16
	C	393	144	102	37	30
Web of Science	**A**	240	164	106	14	14
	C	240	65	63	7	6
Springer	**A**	43734	9637	2679	13	12
	C	43734	2867	1167	13	12

Todos los documentos presentan un modelo para detectar violencia utilizando IA, 12 de ellos incluyen un nuevo dataset que se presenta en el documento y 2 de ellos presentan una arquitectura para un sistema de detección de agresión física en vídeo desarrollado (Huszár et al., 2023; Ullah et al., 2022).

Validación del estudio de mapeo sistemático

Petersen et al. (2015) introducen un marco para la validación y evaluación de estudios de mapeo sistemático (SMS). Esta guía de evaluación comprende 26 tareas que deben llevarse a cabo durante la ejecución de un SMS. Según los autores, un SMS competente debería abarcar un mínimo del 33 % de estas tareas y, como mínimo, desarrollar 9 tareas para lograr un SMS de calidad intermedia. Según las recomendaciones de Petersen et al. (2008), para evaluar la legitimidad de un SMS, es imperativo calcular la relación entre el número de tareas ejecutadas y las 26 tareas delineadas en las pautas.

La Tabla 2.6 expone la guía de evaluación propuesta por Petersen et al. (2015), la cual consta de tres fases en las que se segmenta el desarrollo de este SMS: planificación, desarrollo e informe; que coincide con las Secciones 2.1.2.1, 2.1.2.2 y 2.1.2.3, respectivamente. El SMS desarrollado en esta Tesis Doctoral incorpora 16 de las 26 tareas prescritas por el marco de Petersen et al. (2008), siendo el porcentaje de tareas desarrolladas del 54 %. Por lo tanto, el SMS supera la calidad mediana según lo establecido.

Tabla 2.6: Actividades de la rúbrica de evaluación de Petersen et al. (2015) llevadas a cabo en este estudio de mapeo sistemático (SMS).

			Aplicado
Planificación	**Necesidad del mapa**	Motivar la necesidad y relevancia	Sí
		Definir objetivos y preguntas	Sí
		Consultar al público objetivo para definir las preguntas	No
Desarrollo	**Identificación del estudio/ Elección de la estrategia de búsqueda**	Búsqueda en cadena	No
		Manual	Sí
		Realizar búsqueda en bases de datos	Sí
	Búsqueda en desarrollo	PICO	Sí
		Consultar con bibliotecarios o expertos	No
		Intentar iterativamente encontrar más artículos relevantes	Sí
		Palabras clave de artículos conocidos	Sí
		Usar estándares, enciclopedias y tesauros	No
	Evaluar la búsqueda	Conjunto de prueba de artículos conocidos	Sí
		Evaluación de expertos de los resultados	No
		Buscar páginas web de autores clave	No
		Prueba-reprueba	No
	Inclusión-Exclusión	Identificar criterios objetivos para la decisión	Sí
		Añadir revisor adicional, resolver desacuerdos entre ellos cuando sea necesario	No
		Reglas de decisión	Sí
Informe	**Proceso de extracción**	Identificar criterios objetivos para la decisión	Sí
		Ocultar información que podría sesgar	No
		Añadir revisor adicional, resolver desacuerdos entre ellos cuando sea necesario	No
		Prueba-reprueba	No
	Esquema de clasificación	Tipo de investigación	Sí
		Método de investigación	No
		Tipo de lugar	Sí
	Discusión de validez	Discusión de validez / Limitaciones proporcionadas	Sí

2.2. Clasificación de Algoritmos y Desafíos en la Detección de Violencia mediante IA

Esta Sección clasifica los distintos algoritmos de detección de violencia en vídeo y presenta los desafíos que estos implican. Para ello se divide en dos Secciones. La Sección 2.2.1 describe las etapas fundamentales para la implementación de algoritmos de detección de violencia en vídeo y ofrece una categorización propia de estos algoritmos. La Sección 2.2.2 aborda los desafíos y limitaciones asociados a la detección de violencia, recopilando todas las menciones relevantes presentes en las distintas secciones de los estudios analizados.

2.2.1. Etapas principales y clasificación de los algoritmos de detección de violencia en vídeo mediante IA

Esta Sección expone las etapas principales en la implementación de los algoritmos de detección de violencia en vídeo, además de presentar una categorización propia de los algoritmos de detección de violencia en vídeo.

La Figura 2.1 representa las **etapas principales en la detección de violencia en vídeo mediante IA** (Omarov et al., 2022; Siddique et al., 2022). Dichas etapas se dividen en dos fases: *entrenamiento* (la cual incluye el entrenamiento y testeo del modelo con datasets etiquetados) y *situación en la vida real*. La etapa de *entrenamiento* contiene los pasos para entrenar un algoritmo de detección de violencia mediante el uso de inteligencia artificial: comenzar con un conjunto de datos etiquetado (vídeos de violencia y no violencia) donde los vídeos generalmente se recortan para contener el mismo número de fotogramas y el mismo tamaño de imagen (número de píxeles de alto y ancho); tras ello, preprocesar los datos, convirtiendo los vídeos al formato de datos de entrada para ser pasados al algoritmo de detección de violencia y realizar la clasificación de si la escena se predice o no como violenta.

Existen muchos tipos de algoritmos de detección de violencia (Omarov et al., 2022; Siddique et al., 2022; Yao & Hu, 2023), pero esencialmente se basan en extraer características de los vídeos (espaciales, temporales, etc.) y entrenar un algoritmo que aprenda a partir de estas características. Una vez se realiza este proceso, las

características extraídas se introducen en un clasificador que decide si la escena es violenta o no. Este proceso requiere experimentar múltiples combinaciones de hiperparámetros y evaluar métricas hasta encontrar una estructura óptima.

Por otro lado, la *situación en la vida real* consiste en recibir una señal continua de imagen o vídeo (Shubber & Al-Ta'i, 2022). La imagen debe ser preprocesada de la misma manera que se realiza el preprocesamiento en el entrenamiento, para que el algoritmo entrenado reciba el mismo tipo de entrada. Dado que el análisis en tiempo real es computacional y económicamente costoso, existen técnicas computacionalmente más ligeras que se pueden utilizar para analizar únicamente grupos de fotogramas que pueden contener potencialmente violencia; la extracción de fotogramas característicos es opcional y no todos los trabajos de vanguardia la proponen. Finalmente, aquellas imágenes o fotogramas que puedan contener violencia se introducen en el algoritmo entrenado que clasificará la información como violenta y no violenta.

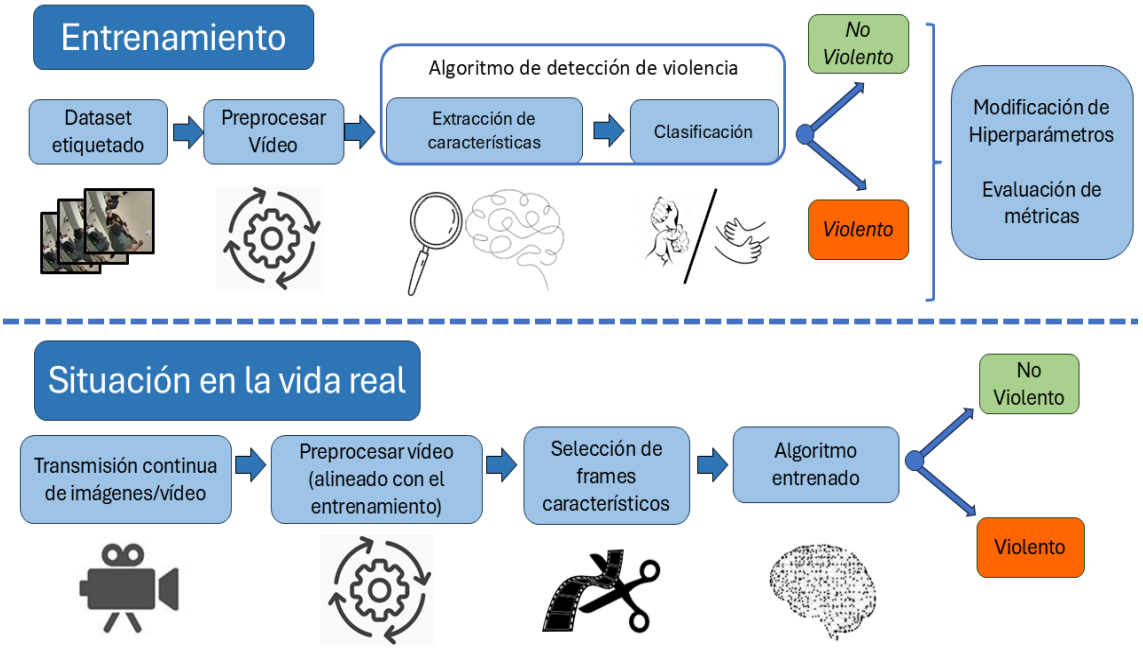

FIGURA 2.1: Pasos básicos de los algoritmos de detección de violencia en vídeo (Omarov et al., 2022; Siddique et al., 2022).

A partir de los trabajos analizados, esta Tesis Doctoral propone una **clasificación de los algoritmos de detección de violencia en vídeo mediante IA**.

Atendiendo a las clasificaciones realizadas en la literatura, un 87 % de los trabajos seleccionados clasifican los algoritmos de detección de violencia en *métodos tradicionales* y *métodos de aprendizaje profundo*. Entre los *métodos tradicionales* se incluyen aquellos

TABLA 2.7: Categorizaciones de trabajos relacionados a partir de los artículos seleccionados.

Referencia del trabajo	**Criterio de categorización**	**Categorías**
(Huszár et al., 2023)	Enfoque del algoritmo	Modelado de patrones normales, Aprendizaje de múltiples instancias, Aprendizaje supervisado
(Bi et al., 2022)		Algoritmos de segmentación semántica, Métodos de extracción de fotogramas clave, Modelos de detección de violencia
(Ahmed et al., 2021)		Métodos basados en trayectoria, Métodos no centrados en objetos, Métodos basados en aprendizaje profundo
(Shang et al., 2022)	Naturaleza del algoritmo	Enfoque de reconocimiento de vídeos violentos, Aprendizaje auto-supervisado, Destilación de conocimiento
(Singh et al., 2022)		Redes neuronales convolucionales, Basado en detección de movimiento, Otros algoritmos de aprendizaje automático
(Zhou, 2022)	Etapa del proceso de detección	Estimación de la pose humana, Reconocimiento de acciones
(Mohtavipour et al., 2022; Monteiro & Durães, 2022)	Enfoque espacial	Enfoque local, Enfoque global, (Métodos de aprendizaje profundo)

que emplean técnicas de aprendizaje automático o extracción manual de características. Esta división tradicional puede resultar vaga e imprecisa dada la amplia variedad de algoritmos utilizados. Solo ocho de los trabajos seleccionados en la Tabla 2.7 categorizan los métodos de manera diferente, los cuales se han agrupado en esta Tesis Doctoral en cuatro categorías: *enfoque del algoritmo*, *naturaleza del algoritmo*, *etapa del proceso de detección* y *enfoque espacial*.

Basándose en los trabajos analizados en esta Tesis Doctoral, se considera fundamental establecer una clasificación clara y precisa de los distintos tipos de algoritmos empleados en la detección de violencia en vídeo mediante IA. A continuación, se presenta la propuesta de categorización desarrollada en esta investigación, la cual ofrece un desglose detallado de la naturaleza de los algoritmos utilizados en la literatura:

- **Algoritmos manuales**. Estos algoritmos suelen centrarse en la extracción de características a partir de factores como el movimiento y texturas (Hu et al., 2022). Requieren un proceso de extracción de características y un proceso de entrenamiento de esas características extraídas (Ehsan & Mohtavipour, 2020).

- **CNN**. Las CNN son comúnmente utilizadas en la detección de violencia debido a su capacidad para capturar características espaciales en los fotogramas de video, lo que las hace efectivas para reconocer patrones asociados con acciones violentas (Ehsan & Mohtavipour, 2020). Las redes neuronales convolucionales (CNN) preentrenadas también son ampliamente utilizadas en los trabajos seleccionados (Freire-Obregón et al., 2022). Los subgrupos en los que se dividen las CNN son: *CNN* y CNN preentrenadas (*CNN PT*).

- **LSTM**. Las LSTM son las RNN más utilizadas en problemas de detección de violencia. Permiten aprender y controlar el flujo de información a lo largo de secuencias temporales de manera más efectiva que las RNN tradicionales, ya que resuelven el problema de desvanecimiento del gradiente que ralentiza el proceso de aprendizaje (Halder & Chatterjee, 2020; Jahlan & Elrefaei, 2021).

- **Basados en esqueleto**. Los algoritmos basados en esqueleto se fundamentan en el seguimiento y análisis de posturas y movimientos humanos representados como esqueletos en los videos. Estos algoritmos utilizan información de las articulaciones y conexiones entre partes del cuerpo para identificar patrones de movimiento anormales o violentos. Los algoritmos basados en esqueleto se dividen en dos subcategorías: técnicas de aprendizaje profundo basadas en esqueleto *skeleton-based (DL)* (Hung et al., 2021; Zhou, 2022) y técnicas manuales basadas en esqueleto *skeleton-based (manual)*.

- **Transformer**. Los algoritmos Transformer son también utilizados para la detección de violencia en video (Aktı et al., 2022) (Ehsan et al., 2023) (Kumar et al., 2023), los cuales dividen la imagen en parches, añaden información de posición, calculan la atención entre parches, aplican transformaciones en capas y generan salidas para tareas de clasificación. Estos modelos son una alternativa efectiva a las redes convolucionales en aplicaciones de procesamiento de imágenes.

- **Basados en audio**. Los algoritmos basados en audio para la detección de violencia analizan características sonoras, como tono, ritmo y eventos sonoros violentos (como gritos o golpes). Estos se dividen en dos subcategorías: técnicas de extracción manual de características basadas en audio *audio (manual)* o técnicas de aprendizaje profundo basadas en audio *audio (DL)* (Mahalle & Rojatkar, 2021b; Zheng et al., 2021).

Como se detalla más adelante en la Sección 2.4.1, muchos de los algoritmos de detección de violencia actuales combinan múltiples enfoques. Por ello, la clasificación presentada y creada en esta Tesis Doctoral categoriza las distintas técnicas empleadas, incluso cuando estas constituyen una mezcla de métodos diferentes.

2.2.2. Desafíos y parámetros de la detección de violencia

Esta Sección presenta los desafíos y limitaciones de la detección de violencia, compilando todas las menciones que aparecen en cualquier sección de los trabajos seleccionados. El objetivo es proporcionar una vista completa y actualizada de los desafíos encontrados en este campo, si bien tan solo el 20 % (12 de los 63 artículos) mencionan explícitamente estos desafíos a lo largo de sus documentos. Además, se exponen los parámetros de detección de violencia utilizados para cuantificar la bondad de las predicciones realizadas.

En primer lugar se presentan los **retos** mencionados en los trabajos seleccionados, los cuales se agrupan y categorizan en la Tabla 2.8. La columna *Tipo de Desafío* contiene cuatro categorías en las que se agrupan los desafíos, ya que varios de ellos están relacionados entre sí; estos grupos son: «*problemas de detección y monitoreo*», «*calidad de imagen y condiciones de iluminación*», «*consideraciones de hardware y tiempo real*» y «*cambios en la escena y entorno*». *Posición de la cámara* es el desafío más mencionado, con un total de 6 trabajos que lo mencionan. En segundo lugar, con cuatro citas, *ocultación de elementos* y *fondo no estacionario*.

En segundo lugar, se definen los **parámetros de evaluación** utilizados en los trabajos seleccionados para cuantificar y analizar los resultados obtenidos. La detección de violencia en vídeo se enfoca usualmente como un problema con un resultado binario, es decir, si la agresión física está presente o no en un vídeo. Por ello, los datasets con los que se entrenan los algoritmos suelen tener solamente dos etiquetas que clasifican las escenas entre violentas o no violentas.

La selección de parámetros de evaluación es vital para cuantificar y evaluar correctamente los resultados (Negre et al., 2024b). Se recopilan todos los parámetros de evaluación utilizados en los trabajos seleccionados en la Figura 2.2, donde debido a que la detección de violencia se enfoca como un problema binario, en el que la escena es violenta o no lo es, gran parte de las métricas parten del uso de la matriz de confusión

Tabla 2.8: Categorización de los retos mencionados en los artículos seleccionados.

Nombre del reto	**Tipo de reto**	**Conteo Número Citas**
Oclusión de elementos	Problemas de detección y monitoreo	4
Variación de escala		3
Diferente acción dependiendo de la persona		1
Apariencias similares		1
Variación en la iluminación	Calidad de imagen y condiciones de iluminación	4
vídeo con poca luz		4
Baja resolución de vídeo		4
Desenfoque de movimiento		3
Condiciones climáticas		2
Efectos de iluminación		1
Posición de la cámara	Hardware y consideraciones de tiempo real	4
Costo de procesamiento en tiempo real		4
Disponibilidad limitada de vídeos		2
Movimiento de la cámara		2
Punto de vista de la cámara		1
Imágenes aéreas		1
Tamaño reducido de la persona		1
Fondo no estacionario	Cambios en la escena y entorno	4
Escena abarrotada		3
Cambios bruscos en movimiento		2
Cambios en la apariencia de objetos		2
Fondo complejo		1
Desorden en la escena		1

binaria (Negre et al., 2024a). Así, *accuracy* es claramente la métrica más utilizada, ya que 54 trabajos la emplean, seguida de *recall*, la cual ha sido utilizada en 18 trabajos.

Se exponen a continuación la explicación y las ecuaciones a partir de las que calcular cada uno de los parámetros de evaluación, en orden descendente desde la columna *Recuento de Citas del Artículo* de la Figura 2.2. Como se verá, la mayoría de ellos son métricas que se calculan a partir de los valores obtenidos en la evaluación del modelo y que se reducen a los elementos de la matriz de confusión.

En una **matriz de confusión**, los principales elementos son los siguientes: los **Verdaderos Positivos (VP)**, que representan la cantidad de casos positivos

correctamente identificados por el modelo; los **Verdaderos Negativos (VN)**, que corresponden a los casos negativos clasificados correctamente; los **Falsos Positivos (FP)**, que son los casos negativos que el modelo ha clasificado erróneamente como positivos; y los **Falsos Negativos (FN)**, que indican los casos positivos que el modelo ha identificado incorrectamente como negativos. Donde $VP \in \mathbb{Z}$, $VN \in \mathbb{Z}$, $FP \in \mathbb{Z}$, $FN \in \mathbb{Z}$, siendo $\mathbb{Z}$ el conjunto de los números enteros.

- **Exactitud (*Accuracy*)** (Ecuación 2.1): mide la proporción de instancias clasificadas correctamente respecto al número total de instancias. Es una métrica simple e intuitiva, pero puede no funcionar bien con datasets desequilibrados, aunque es fácil de entender.

$$Accuracy = \frac{(VP + VN)}{(VP + VN + FP + FN)} \tag{2.1}$$

- **Sensibilidad (*Recall*)** (Ecuación 2.2): calcula la proporción de instancias positivas predichas correctamente respecto a todas las instancias positivas reales. Es crucial en situaciones donde evitar falsos negativos es una prioridad.

$$Recall = \frac{VP}{(VP + FN)} \tag{2.2}$$

- **Puntuación F1 (*F1 score*)** (Ecuación 2.3): combina *accuracy* y recall en una sola métrica, proporcionando una medida equilibrada del rendimiento de un modelo. Es útil cuando equilibrar falsos positivos y falsos negativos es importante. Como punto positivo, logra un equilibrio entre *accuracy* y recall.

$$F1\_score = \frac{2 \cdot (accuracy \cdot recall)}{(accuracy + recall)} \tag{2.3}$$

- **Precisión (*Precision*)** (Ecuación 2.4): cuantifica la proporción de instancias positivas predichas correctamente respecto a todas las instancias positivas predichas. Es valioso cuando minimizar falsos positivos es crítico. Es útil para aplicaciones donde minimizar falsos positivos es importante pero ignora falsos negativos.

$$Precision = \frac{VP}{(VP + FP)} \tag{2.4}$$

- **Especificidad (*Specificity*)** (Ecuación 2.5): también conocida como selectividad o tasa de Verdaderos Negativos (TNR), mide la capacidad de un modelo para identificar correctamente instancias negativas. Calcula la proporción de verdaderos negativos (VN) correctamente predichos entre todas las instancias negativas reales. Una alta especificidad indica una baja tasa de falsos positivos.

$$Specificity = \frac{\text{VN}}{\text{VN} + \text{FP}} \tag{2.5}$$

- **ROC (Característica de Operación del Receptor o *Receiver Operating Characteristic*)** (Ecuación 2.6): es una representación gráfica del rendimiento de un modelo en diferentes umbrales. Grafica la Tasa de Verdaderos Positivos (Sensibilidad) contra la Tasa de Falsos Positivos ($1 - Especificidad$). Aunque muestra el equilibrio entre sensibilidad y especificidad en varios umbrales, puede no ser ideal para datasets desequilibrados.

- **AUC (Área Bajo la Curva ROC)** (Ecuación 2.6): el área bajo la Curva ROC (AUC) cuantifica el rendimiento general de un modelo de clasificación binaria. Representa la probabilidad de que el modelo clasifique correctamente una instancia positiva elegida al azar más alta que una instancia negativa elegida al azar. Un AUC más alto indica una mejor capacidad de discriminación, con un AUC de 0.5 que representa un rendimiento aleatorio y un AUC de 1.0 que representa una discriminación perfecta. Es una métrica ampliamente utilizada para evaluar modelos de clasificación, especialmente cuando se trata con datasets desequilibrados o cuando elegir un umbral óptimo es importante. La Tasa de Falsos Positivos se expresa como: $FPR = \text{FP}$ (VN + FP).

$$AUC = \int_0^1 \text{Recall} \cdot \text{d}(FPR) \tag{2.6}$$

- **Tiempo de Computación**: mide cuánto tiempo tarda un modelo en ser entrenado y en inferir datos en tiempo real. Se mide en segundos y es esencial para evaluar la eficiencia de un modelo, especialmente en aplicaciones en tiempo real o con recursos limitados, reflejando la eficiencia y escalabilidad del modelo.

- **AP (*Average Precision*)** (Ecuación 2.7): mide el área bajo la curva de precisión-recall, promediando la precisión en distintos niveles de *recall*. Donde N

representa el número total de niveles de *recall*.

$$AP = \sum_{n=1}^{N} P(n) \cdot \Delta R(n) \tag{2.7}$$

- **MAP** (Ecuación 2.8): extiende AP a múltiples consultas o tareas, calculando el AP promedio a través de un conjunto de consultas o tareas. Es común en sistemas de recuperación de información. Es útil para evaluar modelos en escenarios que involucran múltiples tareas de recuperación. Donde K es el número total de consultas o tareas.

$$\text{MAP} = \frac{1}{K} \sum_{k=1}^{K} AP_k \tag{2.8}$$

- **G-mean (Media Geométrica)** (Ecuación 2.9): es una métrica que equilibra sensibilidad y especificidad, centrándose en la capacidad de un modelo para clasificar correctamente tanto instancias positivas como negativas. Es adecuada para datasets desequilibrados, enfatiza el rendimiento equilibrado.

$$\text{G-mean} = \sqrt{\textit{Recall} \cdot \textit{Specificity}} \tag{2.9}$$

En resumen, se exponen en esta Sección las métricas de evaluación utilizadas en los trabajos seleccionados. La Figura 2.2 recoge las métricas utilizadas. Con todo, como se ha expuesto anteriormente, la *accuracy* es la métrica más utilizada con 54 artículos, seguida de *recall* con 18. En resumen, se compilan y exponen en esta Sección las métricas actualmente utilizadas en la detección de agresión física en vídeos.

2.3. Datasets, selección de fotogramas característicos y preprocesamiento en la detección de violencia en vídeo mediante IA

Esta Sección expone los pasos previos a la inserción de los datos en el modelo de detección de violencia en los trabajos seleccionados. Dichos pasos se muestran en la Figura 2.1 y consisten en la recopilación de datasets relevantes, la selección de fotogramas característicos, así como el tipo de datos de entrada que se introduce en el algoritmo de

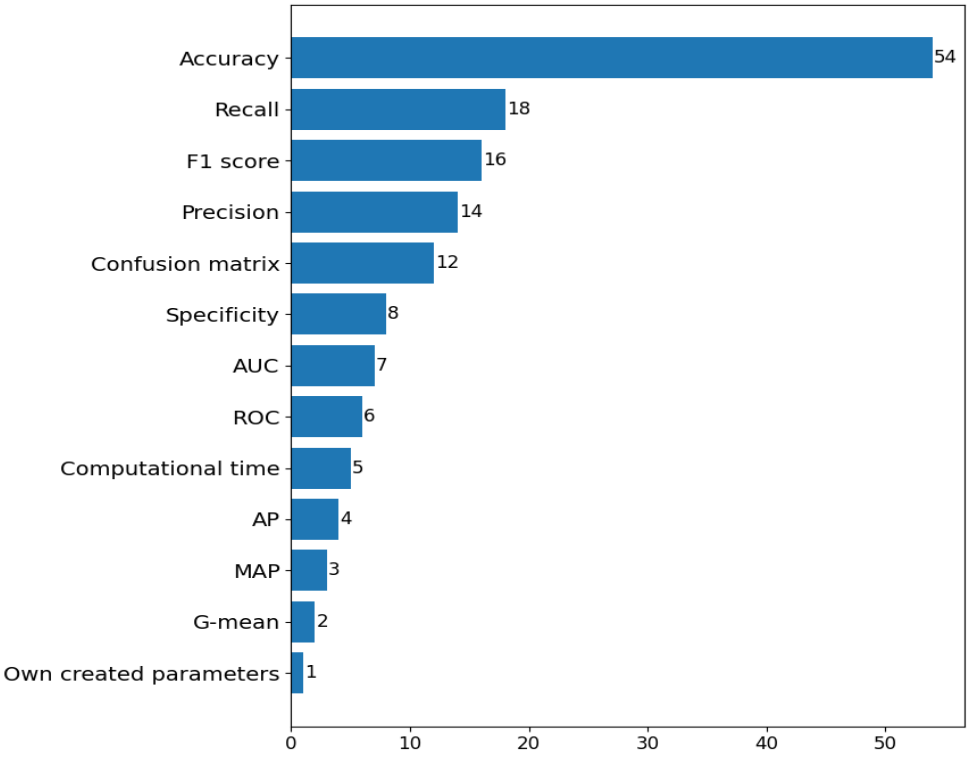

FIGURA 2.2: Conteo del uso de cada métrica de evaluación en los trabajos seleccionados.

detección de violencia. La Sección 2.3.1 analiza los datasets utilizados en los trabajos seleccionados. La Sección 2.3.2 trata los métodos de selección de fotogramas relevantes en los trabajos seleccionados. La Sección 2.3.3 presenta los tipos de datos de entrada que se introducen en los algoritmos de detección de violencia.

2.3.1. Datasets

Esta Sección analiza los datasets utilizados en los trabajos seleccionados, los cuales tratan la detección de violencia en vídeo mediante el uso de inteligencia artificial.

Las agresiones físicas son un evento difícil de detectar en vídeo debido a una serie de desafíos. Uno de estos retos es, sin duda, el hecho de que mientras las agresiones físicas ocurren en todo el mundo, no son tan comúnmente presenciadas o grabadas como otras actividades cotidianas, como personas practicando deportes, paseando o conversando. Por lo tanto, la recopilación de datasets públicos de calidad es esencial para el desarrollo de algoritmos óptimos de detección de agresiones físicas.

La Figura 2.3 muestra el recuento de citas de los datasets que han sido empleados en al menos dos de los trabajos seleccionados, dado que varios de ellos han sido

utilizados únicamente en un artículo. Los cinco datasets más utilizados en los trabajos seleccionados, en orden descendente, son: *Hockey Fights dataset* (Bermejo Nievas et al., 2011), *Action movies dataset* (Bermejo Nievas et al., 2011), *Violent Flow dataset* (Hassner et al., 2012), *RWF-2000* (Cheng et al., 2021a) y *Real Life Violent Situations dataset* (RLVS) (Soliman et al., 2019).

Algunos de estos datasets contienen escenas de agresiones en escenarios no reales, como escenas de películas de acción, lo que supone una limitación significativa para la *accuracy* de los algoritmos entrenados cuando deben detectar situaciones reales (Negre et al., 2024b). El uso extensivo de los datasets de peleas de hockey y películas de acción, creados en 2011, contienen escenas de violencia de *Hockey Fights dataset* y *Action movies dataset*. Aunque es comprensible usar estas grabaciones para comparar resultados con estudios de última generación, están lejos de reflejar situaciones reales debido a factores como el movimiento de cámara, iluminación, enfoque en la agresión, disfraces y calidad de imagen. Por eso, deben ser tratados con conocimiento académico en vista del entrenamiento real de modelos.

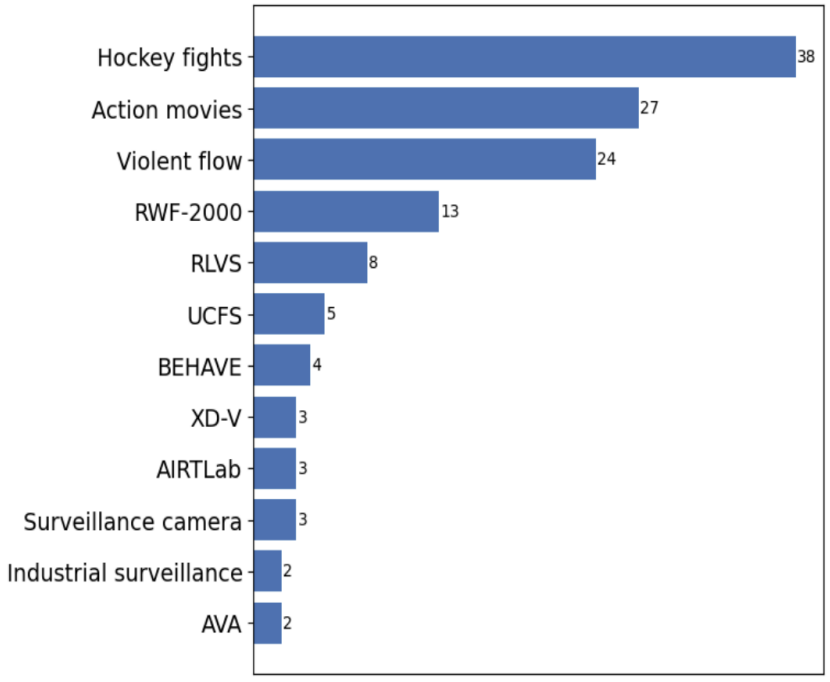

FIGURA 2.3: Recuento de citas de los datasets utilizados en los trabajos seleccionados (Negre et al., 2024b).

Uno de los criterios de exclusión para la selección de trabajos son aquellos que abordan otros tipos de agresión en lugar de violencia física. Cabe destacar que en la mayoría de los trabajos seleccionados los autores filtran los dataframes manteniendo solo los vídeos que

contienen agresiones físicas y no otros tipos de violencia como, por ejemplo, tiroteos. Por otro lado, no se consideran en este SMS datasets que contienen solo acciones no violentas, como el dataset *KTH* (Schuldt et al., 2004).

La Tabla 2.9 resume todos los datasets (28 en total) utilizados en los trabajos seleccionados. El valor *N.S.* en las celdas significa que la información no está especificada en el trabajo que presenta el dataframe. La columna *Nombre* contiene el nombre del dataframe; en los casos en que los creadores del dataframe no designan un nombre específico, se nombra como *Dataset Propio + artículo citado*. La columna *Año* contiene el año de creación del dataset. La columna *Solo A.F.* indica si el dataset contiene exclusivamente vídeos de agresiones físicas como forma de violencia. La columna *Número de Clips* contiene el número de vídeos en el dataframe. La columna *Vídeos Violentos* indica el número de vídeos violentos en el dataframe y la columna *Vídeos No Violentos* indica el número de vídeos no violentos en el dataframe. La amplia variedad de datasets, con diferentes tamaños y contenidos es muy positiva porque permite el estudio de la detección de agresiones físicas desde diferentes perspectivas. Se expone también a continuación qué clase de vídeos contiene cada dataset, así como datos relevantes a destacar.

El *Hockey fights dataset* (Bermejo Nievas et al., 2011) es el más utilizado en los trabajos seleccionados y contiene vídeos de acción de juegos de hockey sobre césped de la National Hockey League (NHL). El *Action movies dataset* (Bermejo Nievas et al., 2011) es el segundo más utilizado y está formado por escenas de películas de acción. Violent Flow (Hassner et al., 2012) o también llamado multitud violenta contiene vídeos de la vida real de YouTube con multitudes y audio. El *Real World Fight-2000 (RWF-2000) dataset* (Cheng et al., 2021a) comprende peleas de cámaras de vigilancia en escenas del mundo real. El *Real Life Violence Situations (RLVS)* (Soliman et al., 2019) consta de situaciones de violencia en la vida real obtenidas de vídeos de YouTube. El *UCF-Crime Selected (UCFS) dataset* (Sultani et al., 2018) recoge vídeos anómalos de vigilancia largos sin recortar con audio que cubren 13 categorías, incluyendo: abuso, arresto, incendio, asalto, accidente, robo, explosión, pelea, robo, accidente de carretera, disparo, robo y vandalismo.

El *BEHAVE dataset* (Blunsden & Fisher, 2010) contiene grabaciones de cámaras de vídeo fijas largas con cuadros delimitadores en eventos violentos y no violentos. El

Surveillance Camera dataset (Aktı et al., 2019) consta de escenas reales de vigilancia diurnas y nocturnas, con metraje en interiores y exteriores. El *XD-Violence (XD-V) dataset* (Wu et al., 2020) está formado por seis clases violentas (Abuso, Accidente Automovilístico, Explosión, Pelea, Disturbio y Disparo) de películas, deportes, juegos, cámaras de mano, CCTV, cámaras de automóviles, etc. El *AIRTLab dataset* (Sernani et al., 2021a) recoge 60 clips no violentos y 120 clips violentos de dos puntos posiciones de la cámara diferentes, respectivamente. El *Industrial Surveillance dataset* (Ullah et al., 2021) contiene escenas variadas de Google y YouTube grabadas en diferentes industrias, tiendas, oficinas y gasolineras vinculadas a industrias; donde la mayoría de las acciones están al margen del punto central de la cámara y los fps varían como en otros datasets de vigilancia.

El *Human Violence dataset* (Ji et al., 2020) recopila vídeos promocionales de películas con acciones violentas de YouTube, categorizados en combate, contacto físico, posesión de armas y sangre. El *Target dataset* (Srivastava et al., 2022) recoge vídeos de vigilancia con drones para vistas diversificadas y el entorno no restringido desde diferentes alturas. El dataset propio desarrollado por Naik y Gopalakrishna (2021) está formado por vídeos grabados propios con dos tipos de acciones violentas de una sola persona (puñetazos y patadas) realizadas dos veces por 20 personas diferentes variando: fondo, condiciones de iluminación, exteriores e interiores en diferentes ángulos, exteriores e interiores con diferentes vestimentas. El dataset propio desarrollado por Srivastava et al. (2021) contiene imágenes de drones grabadas desde 2-18 metros por actores de ambos géneros entre 17 y 30 años y diferentes alturas. El *HD dataset (High Definition)* (Srivastava et al., 2021) contiene vídeos de alta definición de 2 horas cada uno, con resolución de 1280×720 píxeles, e incluye marcas de tiempo para las escenas violentas, mientras que el resto muestra contenido no violento.

El *Conflict event dataset* (Cheng et al., 2021c) comprende vídeos divididos en 51 categorías de acción. El dataset propio desarrollado por Adithya et al. (2023) consta de vídeos de violencia diversos grabados con drones. El dataset propio desarrollado por Kumar et al. (2023) recoge vídeos del *RLVS dataset*, clips recopilados de YouTube y clips relevantes de GitHub. El *VSD2015 dataset* (Sjoberg et al., 2015) es una extensión del *LIRIS-ACCEDE dataset* compuesto por películas de acción. El *AVA-KINETICS dataset* (Li et al., 2020) está formado 80 acciones humanas diferentes. El *Media Eval VSD-2014 dataset* (Demarty et al., 2015) consta de 7 escenas de películas de acción que incluyen

peleas, sangre, fuego, armas de fuego, armas blancas, persecuciones de automóviles, escenas sangrientas, disparos y explosiones y gritos.

El dataset desarrollado por Mahalle y Rojatkar (2021a) recoge vídeos de YouTube convertidos a audio con un contenido variado de tipos de violencia de 31 categorías.El *SMFI* dataset (Aktı et al., 2022) está formado por imágenes y fotogramas de vídeo recopilados de Twitter, Google y NTU CCTV-Fights. En el caso del conjunto de datos propio desarrollado por Narynov et al. (2021) consta de vídeos reales de comportamiento agresivo y acoso en internet y redes sociales. El conjunto de datos propio desarrollado por Rachna et al. (2022) contiene vídeos de YouTube, películas de stock y vídeos violentos autograbados. El *Violent Clip Dataset (VCD)* (Shang et al., 2022) recoge 37 películas de Hollywood y 30 clips de vídeo de YouTube. Por último, el conjunto de datos propio desarrollado por Hung et al. (2021) se basa en vídeos de voluntarios que realizan juegos de rol de pacientes ancianos con discapacidad en las extremidades inferiores.

FIGURA 2.4: Ejemplos de *frames* de los principales conjuntos de datos de vídeos de violencia (Yao & Hu, 2023).

En conclusión, una variedad diversa de datasets con diferentes tamaños y contenidos está disponible públicamente. Esta variedad es muy ventajosa en el estudio de la detección de agresión física con requisitos específicos. Por último, la Figura 2.4 ejemplos de *frames* de los principales datasets empleados en los trabajos seleccionados.

Tabla 2.9: Conjuntos de datos de agresiones físicas utilizados en los trabajos seleccionados (Negre et al., 2024b).

Nombre	**Cita Original**	**Año**	**Solo A.F.**	**Clip No.**	**Media de *frames* o Duración**	***frames* por segundo (FPS)**	**Tamaño vídeo**	**Vídeos violentos**	**Vídeos No violentos**
Hockey Fights	Bermejo Nievas et al. (2011)	2011	✓	1000	50 *frames*	20-30	720×576	500	500
Action Movies	Bermejo Nievas et al. (2011)	2011	✓	200	49 *frames*	20-30	515×720	100	100
Violent Flow/Violent Crowd	Hassner et al. (2012)	2012	✓	246	N.S.	25	320×240	123	123
Real World Fight-2000 (RWF-2000)	Cheng et al. (2021a)	2021	✓	2000	5 s	30	Múltiple	1000	1000
Real Life Violence Situations (RLVS)	Soliman et al. (2019)	2019	✓	2000	5 s	20-30	480-720	1000	1000
UCF-Crime Selected (UCFS)	Sultani et al. (2018)	2018		1900	7247 *frames*	N.S.	N.S.	950	950
BEHAVE	Blunsden y Fisher (2010)	2010	✓	4	76800 *frames*	25	640×480	Continuo	Continuo
Surveillance Camera	Aktı et al. (2019)	2019	✓	300	2 s	Múltiple	Múltiple	150	150
XD-Violence Selected (XD-V)	Wu et al. (2020)	2020		4754	N.S.	N.S.	N.S.	2405	2349
AIRTLab	Sernani et al. (2021a)	2020		350	5.63 s	30	1920×1080	120	230
Industrial Surveillance	Ullah et al. (2021)	2021	✓	300	5 s	20-30	Múltiple	150	150

Human Violence	Ji et al. (2020)	2021		1930	30-100 s	30	1280×720	1930	0
Target	Srivastava et al. (2022)	2022	✓	150	5-10 s	60	1080p	N.S.	N.S.
Dataset propio	Naik y Gopalakrishna (2021)	2021	✓	273	N.S.	10	N.S.	180	93
Dataset propio	Srivastava et al. (2021)	2021	✓	N.S.	610 s	60	1080p	N.S.	N.S.
HD (High Definition)	Srivastava et al. (2021)	2022	✓	2	2 h	N.S.	1280×720	N.S.	N.S.
Conflict Event	Cheng et al. (2021c)	2021	✓	6849	N.S.	N.S.	N.S.	N.S.	N.S.
Dataset propio	Adithya et al. (2023)	2023	✓	20	N.S.	N.S.	N.S.	10	10
Dataset propio	Kumar et al. (2023)	2023	✓	1,112	N.S.	N.S.	N.S.	556	556
VSD2015	Sjoberg et al. (2015)	2015	✓	10,900	8-12 s	N.S.	N.S.	505	10395
AVA-Kinectics Dataset	Li et al. (2020)	2020		430/143	15 minutos	N.S.	N.S.	35	83
Media Eval VSD-2014	Demarty et al. (2015)	2014		N.S.	N.S.	25	N.S.	N.S.	N.S.
Dataset propio	Mahalle y Rojatkar (2021a)	2021		60	30 s	Audio	Audio	N.S.	N.S.
Social Media Fight Images (SMFI)	Aktı et al. (2022)	2022		5691	N.S.	N.S.	N.S.	2739	2952
Dataset propio	Narynov et al. (2021)	2021		2093	34.41 s	N.S.	N.S.	2093	0
Dataset propio	Rachna et al. (2022)	2022		106	N.S.	N.S.	N.S.	66	40
Violent Clip Dataset (VCD)	Shang et al. (2022)	2022		7279	8-12 s	N.S.	N.S.	3677	3609
Dataset propio	Hung et al. (2021)	2021		1500	3-6 s	13-15 s	N.S.	900	600

2.3.2. Selección de fotogramas relevantes

Esta Sección analiza los métodos de **selección de fotogramas relevantes** utilizados en los estudios revisados. Aunque la mayoría de los datasets presentados están recortados y etiquetados con escenas específicas de violencia y no violencia, el objetivo final es la detección en tiempo real. Como la agresión física es un evento anómalo, gran parte del tiempo de grabación no muestra violencia (Jaiswal & Mohod, 2021), lo que implica un alto costo computacional debido a la similitud entre fotogramas consecutivos (Ahmed et al., 2021).

A pesar de que la selección de fotogramas relevantes es vital para reducir el costo económico y computacional de un sistema de detección de agresiones físicas en tiempo real, no todos los artículos presentan métodos para abordar este problema. Los trabajos seleccionados se centran principalmente en el nivel de *accuracy* del método propuesto en el trabajo. De hecho, de entre los 63 trabajos seleccionados, solo 21 presentan métodos de extracción de fotogramas relevantes en sus documentos, es decir, el 33 %.

La Tabla 2.10 cita todos los trabajos seleccionados que utilizan un método de extracción de fotogramas relevantes. La columna *Extracción de Fotogramas Clave* contiene tres categorías que agrupan los diferentes enfoques de extracción de fotogramas: *variación de imagen*, *muestreo sistemático* y *detección de objetos*. La categoría *variación de imagen* se refiere a los métodos que se basan en la variación de la imagen para la extracción de fotogramas relevantes, como el uso de *optical flow*, variación espacio-temporal y variación en brillo o texturas. La categoría *Muestreo sistemático* se refiere a los métodos que seleccionan un cierto número de fotogramas, independientemente del vídeo. La categoría de *detección de objetos* se basa en la detección de personas en los fotogramas seleccionados. Por último, la columna *Método de Preprocesamiento* lista el método utilizado para la extracción de fotogramas relevantes.

En primer lugar, se presentan los métodos para seleccionar fotogramas característicos basados en la ***variación de imagen***. Mahmoodi et al. (2022) aplican un método de segmentación de imágenes llamado SSMI que solo selecciona fotogramas cuando hay variación en brillo, contraste o estructura. Magdy et al. (2023) emplean la técnica *Gunner Farneback*, que es un método de flujo óptico que estima el movimiento en una secuencia de imágenes comparando patrones locales de textura y brillo, generando un mapa de

Tabla 2.10: Categorización de los métodos de selección de fotogramas clave utilizados en los trabajos seleccionados (Negre et al., 2024b).

Cita Original	**Extracción de fotogramas clave**	**Método de preprocesamiento**
Mahmoodi et al. (2022)	Variación de la imagen	Índice de Similitud Estructural (SSIM)
Magdy et al. (2023)		Técnica de Gunner Farneback's
Bi et al. (2022)		Método de diferencia de fotogramas
Chen et al. (2021)		Segmentación de vídeo propia basada en la diferencia de brillo entre fotogramas de imagen
Ahmed et al. (2021)		Matriz de diferencia entre dos fotogramas con umbral
Jaiswal y Mohod (2021)		Error cuadrático medio (MSE)
Jahlan y Elrefaei (2021)	Muestreo sistemático	Selección aleatoria de fotogramas
Wintarti et al. (2022)		Selección aleatoria de fotogramas (20 *frames*/vídeo)
Asad et al. (2021)		Seleccionar fotogramas en intervalos regulares.
Kumar et al. (2023)		Selección de F número de cuadros por segundo con intervalo equidistante
Zheng et al. (2021)		Estrategia de muestreo disperso
Liang et al. (2021)	Detección de objetos	YoloV3
Freire-Obregón et al. (2022)		YoloV4
Rachna et al. (2022)		YoloV5
Kim et al. (2022), Zhang et al. (2022)		YoloV5
Ehsan et al. (2023)		Yolo + Farneback
Hung et al. (2021)		OpenPose
Singh et al. (2022)		OpenCV
Naik y Gopalakrishna (2021)		Mask-RCNN detección unipersonal
Ullah et al. (2022)		Mask-RCNN
Ullah et al. (2021)		CNN ligera

vectores de movimiento y extrayendo solo aquellos donde hay variación. Bi et al. (2022) implementan una técnica de selección de fotogramas característicos basada en el flujo óptico, reduciendo el número de fotogramas procesados en $1/k$. Además, según el número de fotogramas seleccionados, se filtran acciones no violentas como abrazos. Chen et al. (2021) desarrollan un método para dividir vídeos largos en vídeos más pequeños donde realizan la diferencia de brillo entre fotogramas adyacentes, lo que podría indicar un cambio en la acción del vídeo. El fotograma donde se encuentra la máxima variación de brillo es donde se segmenta el vídeo. Ahmed et al. (2021) realizan la diferencia entre dos matrices de fotogramas y marca como potencialmente violentos aquellos cuya diferencia de píxeles supera un valor de umbral predefinido. Jaiswal y Mohod (2021) implementan métodos de fragmentación geométrica, centrándose en la variación espacio-temporal de los fotogramas.

Se pasa a exponer aquellos métodos dentro de la categoría de ***muestreo sistemático*** de

selección de fotogramas característicos; estos métodos se basan en seleccionar un cierto número de fotogramas por unidad de tiempo, procesando así continuamente el vídeo entrante, aunque de una manera más ligera que analizando todos los fotogramas todo el tiempo. Wintarti et al. (2022) y Jahlan y Elrefaei (2021) seleccionan 20 fotogramas de cada secuencia de vídeo al azar. De forma similar, Zheng et al. (2021) seleccionan 16 fotogramas de los vídeos al azar. Por otro lado, Asad et al. (2021) toman entradas de dos fotogramas de vídeo en los tiempos t y $t+1$ a intervalos regulares. Además, Kumar et al. (2023) emplean F número de fotogramas por segundo con intervalos equidistantes.

Finalmente, se exponen los métodos en la categoría de ***detección de objetos***. Varios artículos en esta categoría recurren a Yolo (Ehsan et al., 2023; Kim et al., 2022; Liang et al., 2021; Rachna et al., 2022; Zhang et al., 2022), que es un algoritmo ampliamente ejecutado de detección de objetos en imágenes y vídeos. Rachna et al. (2022) y Freire-Obregón et al. (2022) combinan Yolo junto con DeepSort, que es un algoritmo de seguimiento de objetos de vídeo. Hung et al. (2021) aplican OpenPose y un conjunto de modelos de aprendizaje profundo para la detección de puntos clave del cuerpo humano. Singh et al. (2022) se sirven de OpenCV para realizar tareas como detección de objetos, seguimiento de objetos. Por otro lado, Naik y Gopalakrishna (2021) y Ullah et al. (2022) se valen de Mask-RCNN para la detección de objetos y la segmentación semántica en imágenes. Finalmente, Ullah et al. (2021) proponen una arquitectura CNN liviana para la detección de objetos y la selección de fotogramas.

En resumen, utilizar métodos de selección de fotogramas característicos que tengan potencial de contener violencia es crucial para hacer que los sistemas de detección de violencia en tiempo real sean eficientes y rentables; existiendo actualmente, como se ha expuesto, múltiples alternativas para llevar a cabo este proceso.

2.3.3. Tipo de entrada de los algoritmos de detección de violencia

Esta Sección trata las **entradas de los algoritmos de detección de violencia** utilizadas en los trabajos seleccionados. «Tipo de entrada» se refiere al tipo de dato entrante que recibe el algoritmo, así como los preprocesados que se hayan aplicar. En primer lugar, existen multitud de formas de alimentar a un algoritmo para la detección de violencia: vídeo RGB, vídeo en escala de grises, *optical flow*, entre otros. Asimismo, estas entradas o *inputs* se recopilan en la Tabla 2.11.

Tabla 2.11: Tipos de entrada de los algoritmos de detección de violencia empleada en los trabajos seleccionados (Negre et al., 2024b).

Tipo de entrada del Algoritmo	**Enfoque del algoritmo**	**Número de Citas**
RGB	Imagen	48
Vídeo en escala de grises		5
Desenfoque gaussiano		1
Parches de imagen		1
Operador Laplaciano		1
Desenfoque mediano		1
Matriz de rango 3 (xy/xt/yt)		1
Cálculo de límites		1
Flujo óptico	Movimiento	10
Diferencia entre fotogramas		2
Imagen de energía de movimiento separada		2
Fotogramas con fondo suprimido		1
Imagen dinámica		1
Flujo óptico con desenfoque gaussiano		1
Diferencia de fotogramas RGB con desenfoque gaussiano		1
Audio	Audio	5

La Tabla 2.11 está ordenada de forma descendente según el número de artículos que utilizan cada tipo de entrada del algoritmo, agrupado por el enfoque del algoritmo. Por otro lado, la columna *Enfoque del algoritmo* contiene tres categorías creadas para clasificar los diferentes tipos de entrada existentes: *imagen*, *movimiento* y *audio*. Se considera un tipo de entrada diferente a aquellas que aplican técnicas de suavizado de imagen, como el desenfoque gaussiano, ya que modifican el vídeo original.

A continuación, se expone en que consisten los tipos de entrada de los algoritmos empleados en los trabajos seleccionados, incluyendo sus correspondientes preprocesamientos de datos:

- **RGB** (Mumtaz et al., 2022b): esta entrada consiste en los canales de color Rojo, Verde y Azul de un vídeo. Es una representación estándar de imágenes y vídeos donde cada píxel tiene valores de intensidad en estos tres canales.
- **Flujo óptico** (Magdy et al., 2023): representa la velocidad y dirección del movimiento de objetos en un vídeo. Se obtiene calculando el desplazamiento de píxeles entre fotogramas consecutivos, permitiendo la detección de movimiento.

- **Audio** (Mahalle & Rojatkar, 2021b): el audio del vídeo puede ser una entrada importante para la detección de violencia, ya que puede contener pistas de sonido asociadas con eventos violentos como gritos.
- **Vídeos en escala de grises** (Kumar et al., 2023): un vídeo en escala de grises contiene solo información de luminancia y carece de datos de color. Es útil cuando el análisis se centra en cambios en la intensidad de luminancia en lugar de colores.
- **Diferencia entre fotogramas** (Islam et al., 2021b): esta entrada se basa en la diferencia de intensidad entre fotogramas sucesivos. Puede ser útil para detectar cambios abruptos o movimiento en un vídeo.
- **Imagen de energía de movimiento separada** (Mumtaz et al., 2022b): resalta áreas de energía de movimiento significativo en un vídeo, creada calculando y enfatizando regiones con cambios rápidos de intensidad de píxeles en los fotogramas. Ayuda a detectar y analizar el movimiento en los vídeos, facilitando tareas como la detección de movimiento y la identificación de eventos.
- **Fotogramas con fondo suprimido** (Islam et al., 2021a): fotogramas donde el fondo ha sido eliminado, dejando solo objetos en movimiento en primer plano.
- **RGB con Cálculo de límites** (Madhavan et al., 2021): implica determinar y resaltar bordes o límites dentro de un vídeo, ayudando en la identificación de contornos de objetos y estructuras espaciales. Ayuda a detectar y analizar formas y límites de objetos en fotogramas de vídeo.
- **Imagen dinámica** (Islam et al., 2021b): se centra en el movimiento de objetos destacados en el vídeo, combinando y promediando los píxeles de fondo con los patrones de movimiento mientras se retienen las cinéticas a largo plazo.
- **RGB con Desenfoque Gaussiano** (Singh et al., 2022): la entrada aplica un filtro de desenfoque gaussiano a los canales de color Rojo, Verde y Azul de un vídeo. Esta técnica de suavizado reduce el ruido y los detalles finos, lo que resulta en una versión suavizada y menos detallada de los fotogramas originales del vídeo.
- **Flujo óptico con desenfoque Gaussiano** (Singh et al., 2022): se aplica un filtro de desenfoque gaussiano al campo de flujo óptico en un vídeo. Esto ayuda a reducir el ruido y mejorar la suavidad de los vectores de flujo óptico, haciéndolo más

adecuado para tareas de análisis y detección de movimiento al reducir artefactos de movimiento espurios.

- **Diferencia de fotogramas RGB con desenfoque gaussiano** (Singh et al., 2022): implica calcular la diferencia en la información de color entre dos imágenes en un vídeo RGB y posteriormente aplicar un desenfoque gaussiano a esta diferencia. Esta técnica se utiliza para reducir el ruido y enfatizar cambios sustanciales en color u objetos entre fotogramas consecutivos, ayudando en la detección de alteraciones visuales significativas en vídeos.

- **Parches de imagen** (Aktı et al., 2022): divide una imagen o fotograma de vídeo en regiones más pequeñas y rectangulares conocidas como parches. Cada parche representa una porción localizada de la imagen.

- **RGB con Operador Laplaciano** (Chen et al., 2021): filtro matemático aplicado a una imagen para mejorar sus bordes y detalles finos. Este operador se usa en procesamiento de imágenes para la extracción de características, resaltando variaciones en la intensidad.

- **Desenfoque mediano** (Singh et al., 2022): técnica de procesamiento de imágenes que reemplaza el valor de cada píxel con el valor mediano dentro de un vecindario o ventana definida. El desenfoque mediano se usa comúnmente para el filtrado de ruido y mejorar la calidad de la imagen.

- **Matriz de rango tres** ($xy/xt/yt$) (Hu et al., 2022): representa típicamente información espacio-temporal en un vídeo. Es una matriz 3D que abarca tres dimensiones: espacial (x, y), temporal (t) y dominio de frecuencia (frecuencia temporal o flujo óptico).

En resumen, la variedad de tipos de entrada que se pueden introducir en el algoritmo es muy amplia, cada una con un enfoque diferente en la extracción de características que el algoritmo realizará; donde el tipo de entrada del algoritmo más utilizado es el vídeo RGB, seguido del *optical flow* con casi cinco veces menos citas.

2.4. Algoritmos de detección de violencia y evaluación de su exactitud

Esta Sección aborda los principales algoritmos utilizados en la detección de violencia en vídeo mediante inteligencia artificial, así como la evaluación de su exactitud en función de los datasets empleados. Dado que muchas arquitecturas combinan diferentes algoritmos para mejorar el rendimiento, el análisis se organiza en dos partes. La Sección 2.4.1 presenta los distintos tipos de algoritmos empleados en los trabajos seleccionados, destacando su estructura y funcionamiento dentro del proceso de detección de violencia. Se introduce la categorización utilizada en esta tesis para organizar los algoritmos en dos fases, según las ventajas que aporta cada tipo de modelo.

La Sección 2.4.2 analiza en detalle la combinación de redes neuronales convolucionales (CNN) y redes de memoria a largo plazo (LSTM), ya que esta estrategia ha demostrado ofrecer los mejores resultados en la detección de violencia. Se explica cómo las CNN extraen características espaciales de los fotogramas del vídeo, las cuales son procesadas posteriormente por las capas LSTM para capturar patrones temporales. Además, se presentan tablas con información detallada sobre las arquitecturas específicas utilizadas en los estudios analizados.

2.4.1. Algoritmos de detección de violencia mediante IA y su exactitud con los principales datasets

Esta Sección expone el tipo de algoritmos utilizados en los trabajos seleccionados para la detección de violencia en vídeo mediante IA. Esta constituye la etapa principal en el proceso de detección de violencia en vídeo, según los pasos expuestos en la Figura 2.1. Dado que una gran cantidad de arquitecturas utilizan dos algoritmos diferentes de forma combinada, se organiza el proceso de detección de violencia en dos fases: Fase 1 y Fase 2. Con el uso de dos algoritmos se pretende aprovechar las ventajas de predicción que presentan por separado. Por ejemplo, en el caso de la combinación de CNN y LSTM, la Fase 1 (CNN) se encarga de extraer características espaciales, estas características espaciales se introducen en la Fase 2 (capas LSTM) que se encargan de extraer patrones temporales. Las categorización por la que se agrupan los algoritmos a lo largo de toda esta Sección, ha sido creada en esta Tesis Doctoral y expuesta en la Sección 2.2.1

En la Tabla 2.12 se muestra para cada trabajo seleccionado los algoritmos empleados en la Fase 1 y Fase 2, la categoría a la que pertenecen y el tipo de clasificador utilizado. Además, se indica con un PT si el algoritmo había sido preentrenado previamente.

Tabla 2.12: Algoritmos y clasificadores utilizados para la detección de violencia en los trabajos seleccionados. El término *N* indica que el trabajo no utiliza un segundo algoritmo para realizar la detección de violencia.

Referencia	**Categoría Fase 1**	**Categoría Fase 2**	**Algoritmo Fase 1**	**Algoritmo Fase 2**	**Clasificador**
Ahmed et al. (2021)	CNN	N	CNN-V4	N	FCL(2; sigmoide)
Ehsan y Mohtavipour (2020)	CNN	N	CNN	N	FCL(2; SoftMax)
Jayasimhan y Pabitha (2022)	CNN	CNN	CNN-2D	CNN-3D	FCL(2; SoftMax)
Kim et al. (2022)	CNN	N	CNN-3D	N	UNK
Mahmoodi et al. (2022)	CNN	N	CNN-3D	N	FCL (2; SoftMax)
Monteiro y Durães (2022)	CNN	N	SlowfastNetwork	SlowfastNetwork + X3D Network	FCL (2; sigmoide)
Zhang et al. (2022)	CNN	N	CNN-3D	N	FCL (2; sigmoide)
Appavu et al. (2023)	CNN	N	CNN Optimizada	N	FCL (2; SoftMax)
Ji et al. (2020)	CNN	ML	CNN de doble flujo	N	Máquina de Aprendizaje de Clasificación
Adithya et al. (2023)	CNN PT	N	3D-CNN	N	FCL (2; SoftMax)
Bi et al. (2022)	CNN PT	N	ResNet-18 PT	N	FCL (2; UNK)
Chen et al. (2021)	CNN PT	N	Combinación ResNet + Inception-V1)	N	FCL (1; SoftMax)

Freire-Obregón et al. (2022)	CNN PT	N	Redes ConvNets 3D de Doble Flujo Infladas PT	N	Regresión logística (LR); XGBoost; Máquina de vectores de soporte lineal (LSVM)
Gkountakos et al. (2020)	CNN PT	N	CNN(3D-ResNet)	N	FCL (UNK; UNK)
Huszár et al. (2023)	CNN PT	N	X3D-M con Aprendizaje por Transferencia PT en ImageNet	N	FCL (2; sigmoide)
Jain y Vishwakarma (2020)	CNN PT	N	ResNet-V2 preentrenada con ImageNet	ResNetV2 & Fine-Tunning	FCL (1; SoftMax)
Liang et al. (2021)	CNN PT	N	ResNet-5 de Dos Canales [Ruta Lenta y Rápida]	N	FCL (1; UNK)
Mumtaz et al. (2022a)	CNN PT	CNN PT	Deep Multi-Net (Combinación de GoogleNet y AlexNet PT)	DMN	UNK
Qu et al. (2022)	CNN PT	CNN	C3D PT con el dataset THUMOS2014	DC3D	FCL (UNK; SoftMax)
Santos et al. (2021)	CNN PT	N	Red Neuronal X3D PT con Kinetics-400	Red Neuronal X3D PT con Kinetics-400 + Fine tunning	FCL (UNK; UNK)
Ullah et al. (2021)	LSTM	LSTM	Conv-LSTM	Red GRU	FCL (UNK; UNK)
Talha et al. (2022)	CNN	LSTM	CNN	LSTM	FCL (2; UNK)
Halder y Chatterjee (2020)	CNN	LSTM	CNN	Bi-LSTM	FCL(UNK; UNK)

Madhavan et al. (2021)	CNN	LSTM	Conv-2D	LSTM	FCL (UNK; UNK)
Ullah et al. (2022)	CNN	LSTM	Darknet + CNN de Flujo Óptico Residual	M-LSTM	FCL (UNK; SoftMax)
Vijeikis et al. (2022)	CNN	LSTM	U-Net distribuida en el tiempo	LSTM	FCL (2; UNK)
Islam et al. (2021b)	CNN	LSTM	MobileNet	LSTM Convolucional Separables	FCL (2; entropía binaria cruzada)
Sernani et al. (2021b)	CNN PT	LSTM	CNN-3D PT con Sports 1M	Conv-LSTM	SVM; FCL (SoftMax layers); FCL (SoftMax layers)
Mugunga et al. (2021b)	CNN PT	LSTM	VGG-16 PT con ImageNet	Bi-ConvLSTM	FCL (UNK; SoftMax)
Traoré y Akhloufi (2020a)	CNN PT	LSTM	VGG-16 PT con INRA person dataset	Bi-GRU	FCL (3; SoftMax)
Aarthy y Nithya (2022)	CNN PT	LSTM	VGG-16 PT	LSTM	FCL (3; sigmoide)
Asad et al. (2021)	CNN PT	LSTM	VGG-16	Bloques Residuales Densos Anchos (WDRB) + LSTM	FCL (1; SoftMax)
Contardo et al. (2023)	CNN PT	LSTM	MobileNet-V2	Bi-LSTM; ConvLSTM	FCL (2; ReLU/ Simgmoid)
Gupta y Ali (2022)	CNN PT	LSTM	VGG-16	LSTM/Bi-LSTM	Entropía cruzada binaria
Islam et al. (2021a)	CNN PT	LSTM	VGG-16 PT/VGG-19 PT	LSTM	FCL (1; SoftMax)

Jahlan y Elrefaei (2021)	CNN PT	LSTM	Búsqueda Automatizada de Arquitectura Neuronal Móvil (MNAS) PT	Conv-LSTM	Bosque Aleatorio, Máquina de Vectores de Soporte y K Vecinos Más Cercanos.
Mumtaz et al. (2022b)	CNN PT	LSTM	VGG-19	Bi-LSTM	FCL (UNK; SoftMax)
Sharma et al. (2021)	CNN PT	LSTM	Xception PT con ImageNet	LSTM	FCL (UNK; SoftMax)
Singh et al. (2022)	CNN PT	LSTM	DarkNet19 de dos canales preentrenado en ImageNet	LSTM	FCL (2; UNK)
Srivastava et al. (2022)	CNN PT	LSTM	CNN PT: VGG16/ VGG19/ InceptionV3/ DenseNet201/ ResNet101/ MobileNet/ NASNetLarge/ VGG16+VGG19/ ResNet50+ ResNet152V2/ InceptionV3/ ResNet101V2	LSTM	FCL (3; SoftMax)
Traoré y Akhloufi (2020b)	CNN PT	LSTM	EfficienNet-B0 de dos canales PT en ImageNet (uno para flujo óptico y otro para video RGB)	Bi-LSTM	FCL (1; sigmoide)
Hu et al. (2022)	Manual	Manual; CNN	TOP-ALCM	Características Manuales/CNN	SVM; CNN(light-CNN)

Jaiswal y Mohod (2021)	Manual	ML	Patrón Binario Local (LBP); Histograma Difuso de Orientaciones de Flujo Óptico	AdaBoost (Impulso Adaptativo)	Agrupación Ensemble RobustBoost
Lohithashva y Aradhya (2021)	Manual	ML	Patrón de Orientación Local (LOOP)	SVM	SVM
Mohtavipour et al. (2022)	Manual	CNN	Escala de grises; Vectores de flujo óptico; Imagen de Energía de Movimiento (MEI),	CNN de 3 flujos	FCL (2; SoftMax)
Wintarti et al. (2022)	Manual	ML	Análisis de Componentes Principales (PCA)/Transformada Discreta de Wavelet (DWT)	SVM	SVM
Zhou (2022)	Skeleton-based	Skeleton-based	CNN-3D + TokenPose	CNN-3D	FCL (1; sigmoide)
Hung et al. (2021)	Skeleton-based (DL)	ML	Número de esqueletos; Distancia entre esqueletos; cambios en la aceleración de la mano del esqueleto humano	SVM	SVM
Naik y Gopalakrishna (2021)	Skeleton-based (DL)	LSTM	DNN(DeepPose)	LSTM	FCL (UNK; SoftMax)
Narynov et al. (2021)	Skeleton-based (DL)	Skeleton-based (DL)	PoseNet pre-trained	PoseNet PT	FCL (UNK; Softmax)

Rachna et al. (2022)	Skeleton-based (DL)	Manual	PoseNet pre-trained	Distorsión Dinámica del Tiempo (DTW)	K-Nearest Neighbors; Random Forest; Naive Bayes
Srivastava et al. (2021)	Skeleton-based (DL)		CNN(two-branch multistage CNN)	N	SVM
Su et al. (2020)	Skeleton-based (Manual)	CNN	Algoritmo skeleton-based propio	SPIL	FCL (UNK; UNK)
Aktı et al. (2022)	Transformer	N	ViT	N	Fully connected layer(UNK; UNK)
Ehsan et al. (2023)	Transformer	N	Red STAT (Generador y Discriminador)	Red STAT (Generador) + VGG + Similitud por Coseno	
Kumar et al. (2023)	Transformer	N	Capa de Atención Self-Attention Multicanal (MSA)	N	FCL (1; sigmoide)
Mahalle y Rojatkar (2021b)	Audio (Manual)	ML	Tasa de Cruce por Cero (ZCR), Envolvente de Amplitud (AE), Energía a Corto Plazo (STE), Energía de Raíz Cuadrada Media (RMS), Flujo Espectral (SF), Ancho de Banda (BW), Relación de Energía de Banda (VER), Coeficiente Cepstral en la Frecuencia Mel (MFCC)	ELM	ELM

Wu et al. (2020)	Audio (Manual)	ML	Fórmulas propias	Red Holística Localizada (Rama Holística; Rama Localizada; Rama de Puntuación Dinámica)	FCL (2; ReLU)
Shang et al. (2022)	Audio (DL) + CNN PT	ML	TSM-50 (Módulo de Desplazamiento Temporal 50); ResNet-50; PANNs (Redes Neuronales de Audio PT)	MAF-Net	FCL (2; sigmoide)
Zheng et al. (2021)	Audio (DL) + CNN PT	LSTM	TEA network, VGGish	LSTM	FCL (1; UNK)
Cheng et al. (2021c)	Audio (DL) + CNN PT	CNN + P3D (Redes Neuronales Convolucionales Pseudo-3D PT) y AudioNet PT	Red de Razonamiento (2 FCL; SoftMax) + Red Predictiva (3 FCL; SoftMax)	Red de Razonamiento (2 FCL; SoftMax) + Red Predictiva (3 FCL; SoftMax)	

La columna *Classifier* nombra el clasificador que se ha empleado, donde aquellas celdas con un contenido con estructura *FCL (X, Y)*, indica el uso de capas completamente conectadas, donde cada neurona está conectada con todas las de la siguiente capa, con X número de capas e Y tipo de activación en la última capa.

En las columnas *Categoría Fase 1* y *Categoría Fase 2*, se utilizan códigos de color para facilitar la identificación del tipo de algoritmos. Si el término N aparece en la Tabla tanto en las columnas *Tipo Fase 2* como en *Algoritmo Fase 2*, significa que el trabajo no utiliza un segundo algoritmo para realizar la detección de violencia.

En base a la Tabla 2.12, resulta de interés analizar los tipos de algoritmos más empleados según las categorías establecidas al principio de esta Sección y utilizadas también en dicha Tabla. La Figura 2.5 representa el conteo de los tipos de algoritmos utilizados en la Fase 1. La Figura 2.6 representa un conteo de los tipos de algoritmos utilizados en la Fase 2. Por último, La Figura 2.7 representa el recuento de las combinaciones de algoritmos (Fase 1 + Fase 2) de algoritmos empleadas (Fase 1 + Fase 2).

En las tres Tablas —2.5, 2.6 y 2.7— se mantiene el mismo código de color para las categorías que el usado en la Tabla 2.12. Estas Tablas dividen las categorías utilizadas en los siguientes subgrupos: *CNN* se divide en *CNN* (CNN sin preentrenar) y *CNN PT* (preentrenamiento de CNN). *Audio* se divide en *audio (DL)* y *audio (manual)*, así como *skeleton-based* se divide en *skeleton-based (DL)* y *skeleton-based (manual)*, según si el tipo de algoritmo está basado en técnicas manuales o de DL.

En la Figura 2.5 se puede observar cómo la mayoría de los algoritmos utilizados en la primera fase son CNN preentrenadas y CNN. También se puede observar que la mayoría de los algoritmos basados en esqueleto se basan en técnicas de aprendizaje profundo.

En la Figura 2.6 se observa como las LSTM son evidentemente el tipo de algoritmo más utilizado, empleándose las CNN con menor frecuencia. Los algoritmos de aprendizaje automático (ML), que no se emplean en la Fase 1, son el segundo algoritmo más utilizado en la Fase 2.

Por último, se observa en la Figura 2.7 cómo las combinaciónes de algoritmos más empleada el uso de CNN junto con capas LSTM y el uso de CNN en exclusiva, predominando en ambos casos el uso de algoritmos preentrenados. La quinta combinación

más común es el uso de un algoritmo tradicional con una extracción manual de características y un algoritmo clásico de aprendizaje automático para el entrenamiento.

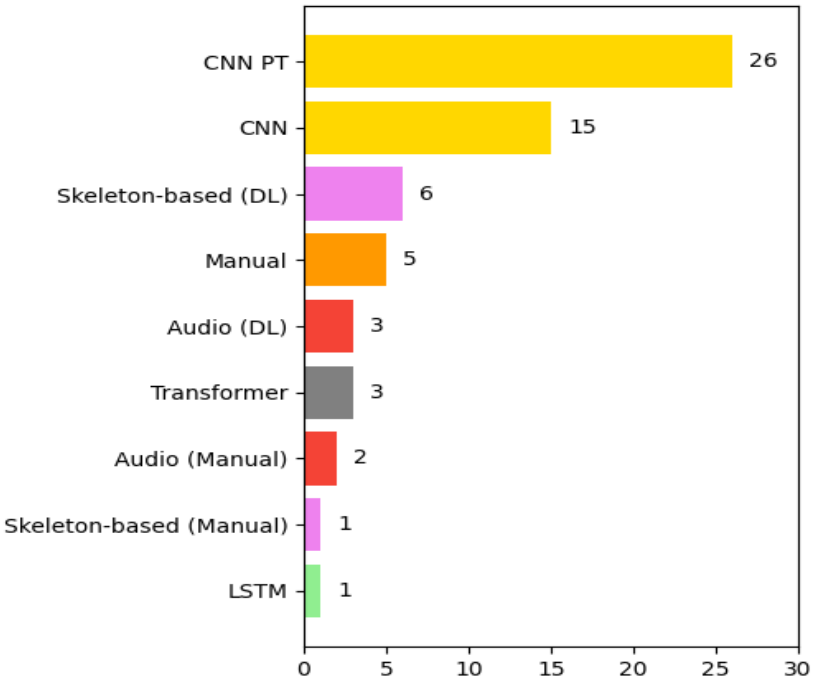

FIGURA 2.5: Recuento y categorización de tipos de algoritmos utilizados en la Fase 1 en los trabajos seleccionados (Negre et al., 2024b). Los trabajos analizados emplean habitualmente CNN en la Fase 1 para la detección de violencia en vídeo.

Tras las tareas de extracción de características, **el proceso finaliza con un clasificador** que determina si el vídeo es violento o no; siendo este es un proceso muy importante. Se muestra en la Tabla 2.12los clasificadores empleados en los trabajos seleccionados. En la Figura se utiliza un código de color para diferenciar aquellos que están basados en técnicas de aprendizaje profundo y aquellos que utilizan métodos tradicionales. Se entiende como métodos tradicionales aquellos clasificadores que son de origen matemático o están basados en técnicas de aprendizaje automático. Se puede observar en la Figura 2.8 que la técnica de clasificador más utilizada son las capas densamente conectadas. Esto es lógico dado que este clasificador está integrado en la salida de los métodos de aprendizaje profundo, que, como se ha visto, son los más ampliamente utilizados.

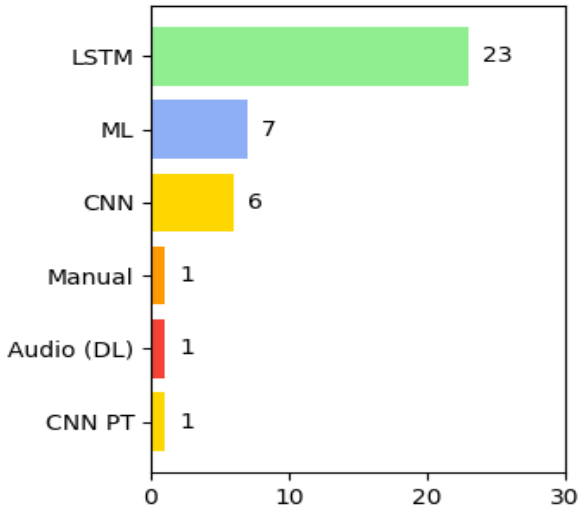

FIGURA 2.6: Recuento y categorización de tipos de algoritmos utilizados en la Fase 2 en los trabajos seleccionados (Negre et al., 2024b). Los trabajos analizados emplean habitualmente capas LSTM en la Fase 2 para la detección de violencia en vídeo.

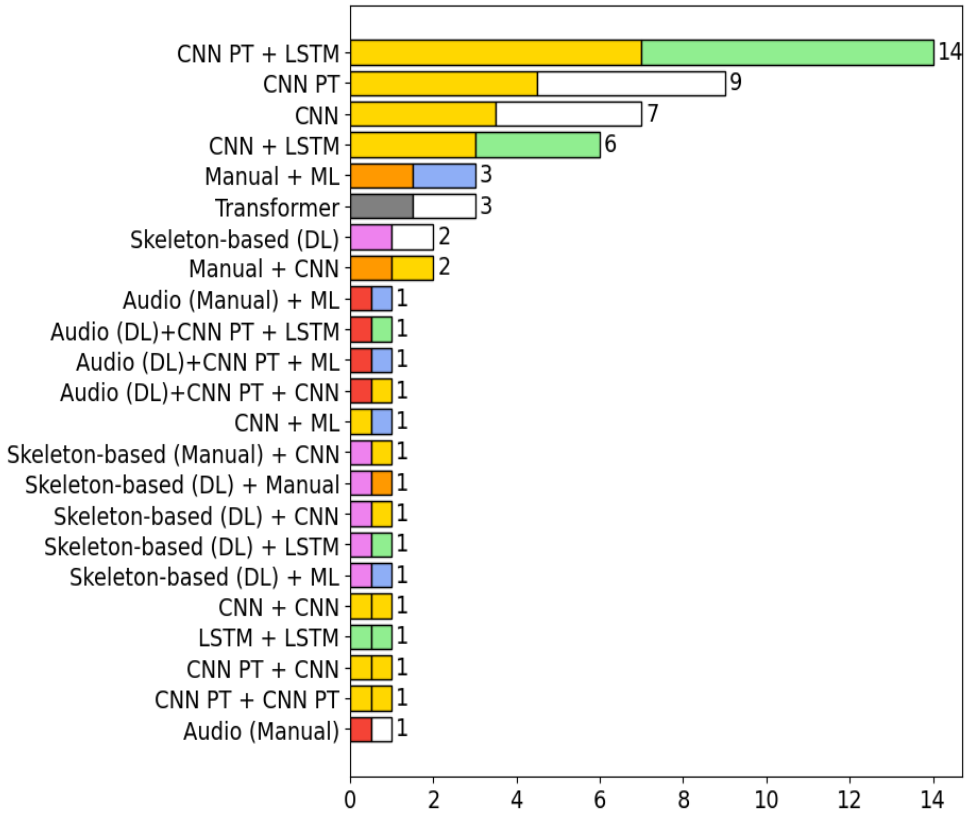

FIGURA 2.7: Recuento y categorización de combinaciones de tipos de algoritmos empleados en los trabajos seleccionados (Negre et al., 2024b). Los trabajos analizados combinan habitualmente CNN junto con LSTM para la detección de violencia en vídeo.

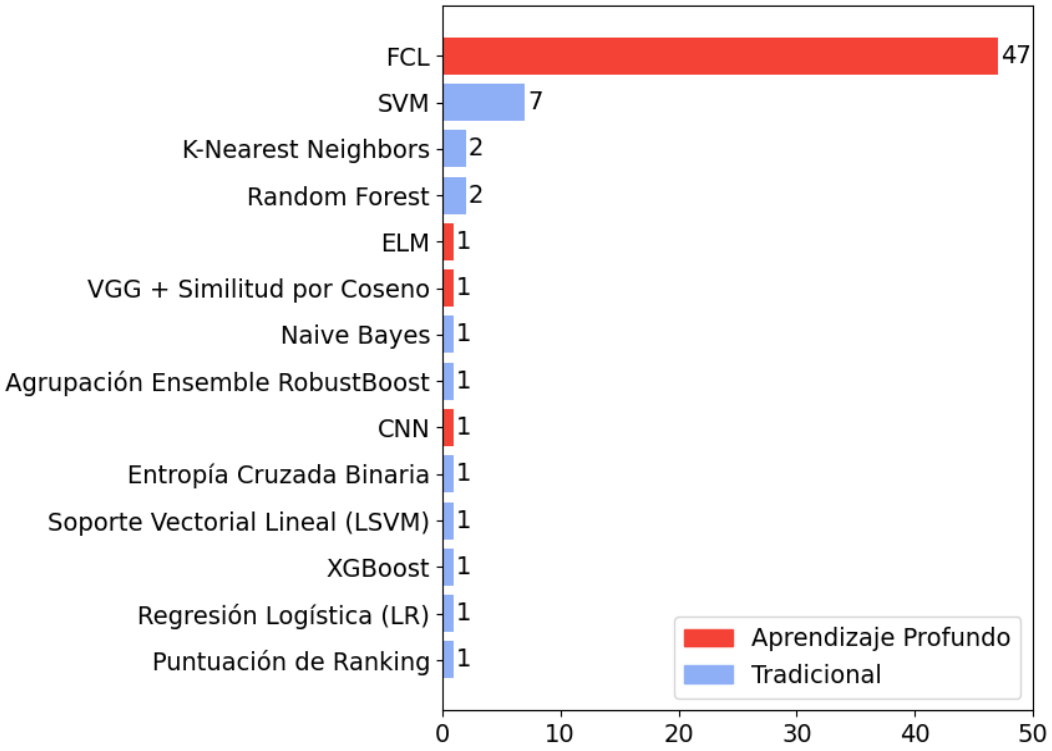

FIGURA 2.8: Recuento de los clasificadores utilizados en los trabajos seleccionados. Los trabajos analizados emplean habitualmente capas densamente conectadas como clasificador en la detección de violencia en vídeo.

Se emplean diferentes funciones de activación para las capas completamente conectadas, como se muestra en la Tabla 2.12. Las funciones de activación sigmoide, ReLU y SoftMax han sido utilizadas en los trabajos seleccionados. SoftMax no proporciona una salida binaria, por lo tanto, para obtener una clasificación binaria de *violencia* o *no violencia*, se debe establecer un valor umbral en el cual se considere que ocurre violencia.

Si bien la mejora en la detección de violencia se centra principalmente en la optimización de los algoritmos que realizan la extracción y el entrenamiento de las características, hay trabajos que han intentado estudiar cómo varían los resultados de los algoritmos al usar diferentes clasificadores, considerando la clasificación como un momento decisivo en el proceso.

Por ejemplo, Freire-Obregón et al. (2022) estudian la efectividad de varios clasificadores de aprendizaje automático, concluyendo que el uso de vídeos que han tenido el contexto de la escena eliminado sufre una pérdida de *accuracy* entre 2–5 %; además, este porcentaje es independiente del grado de contexto eliminado. Destaca que LSVM y LR son clasificadores efectivos para la detección de violencia en vídeos, superando a los enfoques basados en árboles de decisión. Por otro lado, Jahlan y Elrefaei (2021) utilizan

tres clasificadores diferentes para probar su efectividad. Islam et al. (2021b) analizan diferentes variaciones de las últimas capas antes de las capas densamente conectadas para observar su variación en los resultados obtenidos en la clasificación de violencia. Hu et al. (2022) proponen además del uso de un SVM como clasificador, usar una CNN ligera.

En resumen, el proceso de clasificación de violencia en vídeo se basa en la generación de un resultado, generalmente binario, donde el evento se clasifica como violento o no violento. Hay diferentes tipos de clasificadores que pueden dividirse principalmente entre técnicas basadas en aprendizaje profundo o aprendizaje automático. Finalmente, existen trabajos que han dado importancia a la selección del clasificador al estudiar cómo varían los resultados de la detección de violencia.

Se pasa, por último, al estudio de las ***accuracies* obtenidas por los trabajos seleccionados** en la implementación de los modelos propuestos. Este paso, representado en la Figura 2.1, consiste en evaluar cómo el algoritmo ya entrenado se adapta a los vídeos de prueba, con el objetivo de verificar su eficacia en la detección de violencia en vídeo. Algunos de los trabajos no utilizan la métrica de *accuracy*, si bien es, por mucho, la métrica más utilizada. Por ello, los resultados obtenidos por los trabajos seleccionados que no utilizan la *accuracy* como métrica no aparecen en esta compilación.

La Tabla 2.13 contiene los resultados de *accuracy* alcanzados en los trabajos seleccionados, ordenados por el conjunto de datos empleado y por el resultado de *accuracy* obtenido en orden descendente. La Tabla 2.13 no muestra aquellos datasets que contienen solo una cita; por lo tanto, solo se consideran los resultados cuyos artículos han utilizado los datasets mostrados en la Tabla 2.9. A continuación, la columna *Dataset* contiene el conjunto de datos utilizado. En cuanto a la columna *Referencia* contiene la referencia del artículo. Por otro lado, las columnas *Categoría Fase 1* y *Categoría Fase 2* contienen las subcategorías a las que pertenecen los algoritmos utilizados. Las columnas *Fase 1* y *Fase 2* contienen los algoritmos utilizados en la Fase 1 y Fase 2. Finalmente, la columna *Acc.* contiene la *accuracy* obtenida por un artículo dado; esta *accuracy* es propia de cada trabajo, de ahí las diferencias en sensibilidad entre ellos.

TABLA 2.13: *Accuracy* obtenida en la detección de violencia en los trabajos seleccionados, ordenados según el dataset empleado en el entrenamiento y en el testeo.

Dataset	**Referencia**	**Categoría Fase 1**	**Categoría Fase 2**	**Fase 1**	**Fase 2**	**Acc. (%)**
Hockey Fights	Appavu et al. (2023)	CNN	—	CNN optimizada	—	100
	Mohtavipour et al. (2022)	Manual	CNN	Escala de grises + Vectores de flujo óptico + Imagen de Energía de Movimiento (MEI)	CNN de 3 flujos	100
	Hu et al. (2022)	Manual	Manual/CNN	TOP-ALCM	Características manuales + CNN	99.9
	Mumtaz et al. (2022a)	CNN PT	CNN PT	Deep Multi-Net	DMN	99.82
	Islam et al. (2021b)	CNN	LSTM	MobileNet	Red Convolucional LSTM separable	99.5
	Vijeikis et al. (2022)	CNN	LSTM	U-Net distribuida en el tiempo	LSTM	99.5
	Freire-Obregón et al. (2022)	CNN PT	—	ConvNets 3D Infladas de dos flujos PT	—	99.45
	Mahmoodi et al. (2022)	CNN	—	CNN-3D	—	99.4
	Halder y Chatterjee (2020)	CNN	LSTM	CNN	Bi-LSTM	99.27
	Aarthy y Nithya (2022)	CNN PT	LSTM	VGG-16 PT	LSTM	99.1
	Mugunga et al. (2021b)	CNN PT	LSTM	VGG-16 PT with ImageNet	Bi-ConvLSTM	99.1

Traoré y Akhloufi (2020b)	CNN PT	LSTM	EfficientNet-B0 de dos canales PT en ImageNet	Bi-LSTM	99
Jahlan y Elrefaei (2021)	CNN PT	LSTM	Arquitectura Neural Móvil de Búsqueda PT	Conv-LSTM	99
Gupta y Ali (2022)	CNN PT	LSTM	VGG-16	Bi-LSTM	98.9
Asad et al. (2021)	CNN PT	LSTM	VGG-16	Bloques Residuales Anchos y Densos (WDRB) + LSTM	98.8
Sernani et al. (2021b)	CNN PT	LSTM	CNN-3D PT con Sports 1M	Conv-LSTM	98.76
Ullah et al. (2021)	LSTM	LSTM	Conv-LSTM	Red GRU	98.5
Bi et al. (2022)	CNN PT	—	ResNet-18 PT	—	98.5
Ahmed et al. (2021)	CNN	—	CNN-V4	—	98.11
Aktı et al. (2022)	Transformer	—	ViT	—	98
Su et al. (2020)	Skeleton-based (Manual)	CNN	Algoritmo skeleton-based propio	SPIL	98
Islam et al. (2021a)	CNN PT	LSTM	VGG-19 PT	LSTM	98
Traoré y Akhloufi (2020a)	CNN PT	LSTM	VGG-16 PT en INRA person dataset	Bi-GRU	98
Ehsan y Mohtavipour (2020)	CNN	—	CNN	—	98
Ullah et al. (2022)	CNN	LSTM	Darknet + CNN de Flujo Óptico Residual	M-LSTM	98
Gupta y Ali (2022)	CNN PT	LSTM	VGG-16	LSTM	97.6
Huszár et al. (2023)	CNN PT	—	X3D-M transferido y PT en ImageNet	—	97.5
Sharma et al. (2021)	CNN PT	LSTM	Xception PT con ImageNet	LSTM	96.55

	Singh et al. (2022)	CNN PT	LSTM	DarkNet19 de dos canales PT en ImageNet	LSTM	96.2
	Kumar et al. (2023)	Transformer	—	MSA	—	95.15
	Wintarti et al. (2022)	Manual	ML	Transformada Discreta de Wavelet (DWT)	SVM	95
	Talha et al. (2022)	CNN	LSTM	CNN	LSTM	94.9
	Chen et al. (2021)	CNN PT	—	ResNet+Inception-V1 combination	—	94.1
	Mahalle y Rojatkar (2021b)	Audio (Manual)	—	ZCR + AE + STE + RMS + SF + BW + VER + MFCC	ELM	93.5
	Jain y Vishwakarma (2020)	CNN PT	—	ResNet-V2 pre-trained with ImageNet	ResNetV2 + Fine-Tunning	93.33
	Lohithashva y Aradhya (2021)	Manual	ML	Local Orientation Pattern (LOOP)	SVM	92.25
	Mumtaz et al. (2022b)	CNN PT	LSTM	VGG-19	Bi-LSTM	91.29
	Jaiswal y Mohod (2021)	Manual	ML	Local Binary Pattern (LBP) + Fuzzy Histogram of Optical Flow Orientations	AdaBoost	88.66
	Mahmoodi et al. (2022)	CNN	—	CNN-3D	—	100
	Appavu et al. (2023)	CNN	—	Optimized CNN	—	100
	Freire-Obregón et al. (2022)	CNN PT	—	Pre-trained Two-stream inflated 3D ConvNets	—	100

Huszár et al. (2023)	CNN PT	—	Transfer-Learned X3D-M pre-trained on ImageNet	—	100
Jain y Vishwakarma (2020)	CNN PT	—	ResNet-V2 pre-trained with ImageNet	ResNetV2 + Fine-Tunning	100
Mumtaz et al. (2022a)	CNN PT	CNN PT	Deep Multi-Net	DMN	100
Halder y Chatterjee (2020)	CNN	LSTM	CNN	Bi-LSTM	100
Islam et al. (2021b)	CNN	LSTM	MobileNet	Separable Convolutional LSTM	100
Mugunga et al. (2021b)	CNN PT	LSTM	VGG-16 pre-trained with ImageNet	Bi-ConvLSTM	100
Hu et al. (2022)	Manual	Manual/CNN	TOP-ALCM	HandCraftedFeatures+CNN	100
Mohtavipour et al. (2022)	Manual	CNN	Grayscale + Optical flow vectors + Motion Energy Image (MEI)	3-stream CNN	100
Wintarti et al. (2022)	Manual	ML	Discrete Wavelet Transform (DWT)	SVM	100
Aktı et al. (2022)	Transformer	—	ViT	—	100
Islam et al. (2021a)	CNN PT	LSTM	VGG-19 pre-trained	LSTM	99.9
Ahmed et al. (2021)	CNN	—	CNN-V4	—	99.03
Ehsan y Mohtavipour (2020)	CNN	—	CNN	—	99
Asad et al. (2021)	CNN PT	LSTM	VGG-16	Wide Dense Residual Blocks (WDRB) + LSTM	98.99

Action movies	Su et al. (2020)	Skeleton-based (Manual)	CNN	Own skeleton-based algorithm	SPIL	98.5
	Sharma et al. (2021)	CNN PT	LSTM	Xception pre-trained with ImageNet	LSTM	98.32
	Singh et al. (2022)	CNN PT	LSTM	Two channel DarkNet19 pretrained on ImageNet	LSTM	97.41
	Vijeikis et al. (2022)	CNN	LSTM	Time-distributed U-Net	LSTM	96.1
	Jahlan y Elrefaei (2021)	CNN PT	LSTM	Mobile Neural Architecture Search (MNAS) Pre-trained Automated	Conv-LSTM	96
	Mahalle y Rojatkar (2021b)	Audio (Manual)	—	ZCR + AE + STE + RMS + SF + BW + VER + MFCC	ELM	94.32
	Talha et al. (2022)	CNN	LSTM	CNN	LSTM	92.2
	Sernani et al. (2021b)	CNN PT	LSTM	CNN-3D trained with Sports 1M	Conv-LSTM	100
	Jahlan y Elrefaei (2021)	CNN PT	LSTM	Mobile Neural Architecture Search (MNAS) Pre-trained Automated	Conv-LSTM	100
	Mohtavipour et al. (2022)	Manual	CNN	Grayscale+Optical flow vectors+Motion Energy Image (MEI)	3-stream CNN	100
	Appavu et al. (2023)	CNN	—	Optimized CNN	—	99.68
	Freire-Obregón et al. (2022)	CNN PT	—	Pre-trained Two-stream inflated 3D ConvNets	—	99.45

Violent flow	Gkountakos et al. (2020)	CNN PT	—	ResNet-3D	—	99.31
	Halder y Chatterjee (2020)	CNN	LSTM	CNN	Bi-LSTM	98.64
	Ullah et al. (2022)	CNN	LSTM	Darknet + Residual Optical Flow CNN	M-LSTM	98.21
	Aktı et al. (2022)	Transformer	—	ViT	—	98
	Ahmed et al. (2021)	CNN	—	CNN-V4	—	97.65
	Mahmoodi et al. (2022)	CNN	—	CNN-3D	—	97.49
	Asad et al. (2021)	CNN PT	LSTM	VGG-16	Wide Dense Residual Blocks (WDRB) + LSTM	97.1
	Asad et al. (2021)	CNN PT	LSTM	VGG-16	Wide Dense Residual Blocks (WDRB) + LSTM	97.1
	Asad et al. (2021)	CNN PT	LSTM	VGG-16	Wide Dense Residual Blocks (WDRB) + LSTM	97.1
	Traoré y Akhloufi (2020a)	CNN PT	LSTM	VGG-16 PT en INRA person dataset	Bi-GRU	95.5
	Gupta y Ali (2022)	CNN PT	LSTM	VGG-16	Bi-LSTM	95.4
	Su et al. (2020)	Skeleton-based (Manual)	CNN	Algoritmo skeleton-based propio	SPIL	94.5
	Ehsan y Mohtavipour (2020)	CNN	—	CNN	—	94
	Traoré y Akhloufi (2020b)	CNN PT	LSTM	EfficientNet-B0 de dos canales preentrenado en ImageNet	Bi-LSTM	93.75

	Gupta y Ali (2022)	CNN PT	LSTM	VGG-16	Bi-LSTM	92.2
	Huszár et al. (2023)	CNN PT	—	X3D-M transferido y PT en ImageNet	—	92
	Lohithashva y Aradhya (2021)	Manual	ML	Patrón de Orientación Local (LOOP)	SVM	91.54
	Hu et al. (2022)	Manual	Manual/CNN	TOP-ALCM	Características Manuales + CNN	91
	Mumtaz et al. (2022b)	CNN PT	LSTM	VGG-19	Bi-LSTM	89.63
	Talha et al. (2022)	CNN	LSTM	CNN	LSTM	77.31
	Bi et al. (2022)	CNN PT	—	ResNet-18 PT	—	94.6
	Mugunga et al. (2021b)	CNN PT	LSTM	VGG-16 PT con ImageNet	Bi-ConvLSTM	92.4
	Mumtaz et al. (2022b)	CNN PT	LSTM	VGG-19	Bi-LSTM	90.47
	Islam et al. (2021b)	CNN	LSTM	MobileNet	Red Convolutional LSTM separable	89.75
	Zhou (2022)	Skeleton-based (DL)	CNN	TokenPose+HRNetW32	CNN-3D	89.45
	Su et al. (2020)	Skeleton-based (Manual)	CNN	Own skeleton-based algorithm	SPIL	89.3
	Ullah et al. (2021)	LSTM	LSTM	Conv-LSTM	Red GRU	88.2
	Huszár et al. (2023)	CNN PT	—	X3D-M transferido y preentrenado en ImageNet	—	85

RWF-2000	Santos et al. (2021)	CNN PT	—	X3D PT en Kinetics-400	X3D PT con Kinetics-400 + Fine tunning	84.75
	Vijeikis et al. (2022)	CNN	LSTM	U-Net distribuida en el tiempo	LSTM	82.0
	Chen et al. (2021)	CNN PT	—	Combinación ResNet+Inception-V1	—	72
RLVS	Santos et al. (2021)	CNN PT	—	X3D PT en Kinetics-400	X3D PT en Kinetics-400 + Fine tunning	98
	Traoré y Akhloufi (2020b)	CNN PT	LSTM	EfficienNet-B0 de dos canales PT en ImageNet	Bi-LSTM	96.74
	Huszár et al. (2023)	CNN PT	—	Transfer-Learned X3D-M PT en ImageNet	—	95.2
	Kumar et al. (2023)	Transformer	—	MSA	—	95.12
	Bi et al. (2022)	CNN PT	—	ResNet-18 PT	—	95
	Traoré y Akhloufi (2020a)	CNN PT	LSTM	VGG-16 PT en INRA person dataset	Bi-GRU	90.25
	Jain y Vishwakarma (2020)	CNN PT	—	ResNet-V2 PT en ImageNet	ResNetV2 + Fine-Tunning	86.79
	Jayasimhan y Pabitha (2022)	CNN	CNN	CNN-2D	CNN-3D	84.5
UFC crime	Adithya et al. (2023)	CNN PT	—	3D-CNN	—	99.57
	Mugunga et al. (2021b)	CNN PT	LSTM	VGG-16 PT en ImageNet	Bi-ConvLSTM	99.1
	Huszár et al. (2023)	CNN PT	—	Transfer-Learned X3D-M PT en ImageNet	—	84.2

BEHAVE	Mugunga et al. (2021b)	CNN PT	LSTM	VGG-16 PT en ImageNet	Bi-ConvLSTM	99.3
	Asad et al. (2021)	CNN PT	LSTM	VGG-16	Bloques Residuales Anchos y Densos (WDRB) + LSTM	95.9
	Qu et al. (2022)	CNN PT	CNN	C3D PT en THUMOS2014 dataset	DC3D	73.8
Surveillance Camera	Aktı et al. (2022)	Transformer	—	ViT	—	84.6
	Ullah et al. (2021)	LSTM	LSTM	Conv-LSTM	Red GRU	75.9
	Ullah et al. (2022)	CNN	LSTM	Darknet + CNN de Flujo Óptico Residual	M-LSTM	74
AIRTLab	Sernani et al. (2021b)	CNN PT	LSTM	CNN-3D PT en Sports 1M	Conv-LSTM	97.32
Industrial Surveillance	Mumtaz et al. (2022b)	CNN PT	LSTM	VGG-19	Bi-LSTM	81.22
	Ullah et al. (2021)	LSTM	LSTM	Conv-LSTM	Red GRU	80
AVA	Monteiro y Durães (2022)	CNN	—	SlowfastNetwork	SlowfastNetworw + red X3D	84.5
XD-V	Huszár et al. (2023)	CNN PT	—	Transfer-Learned X3D-M PT en ImageNet	—	89.31

Se observa que el uso exclusivo de CNN y la combinación de CNN con capas LSTM obtienen excelentes resultados, alcanzando, por ejemplo, el 99 % y el 100 % en el dataset *Hockey Fights*. Así mismo, otros trabajos como el de Hu et al. (2022) y Mohtavipour et al. (2022) basados en métodos de extracción manual de características en combinación con CNN, así como el modelo desarrollado por Aktı et al. (2022) basado en el uso de *transformers*, también obtienen muy buenos resultados por encima del 98 % en el mismo conjunto de datos. En general, se observa que la elección del dataset afecta a los resultados obtenidos, ya que los datasets con escenas más variadas, calidad de imagen más pobre, más personas en la escena, entre otras características, realizan peores detecciones.

En conclusión, esta Sección muestra y analiza los tipos de algoritmos y clasificadores empleados en los trabajos seleccionados. Este paso, ilustrado en la Figura 2.1, consiste analizar cómo el modelo entrenado se adapta a los vídeos de prueba para verificar su eficacia en la detección de violencia en vídeo. La Sección incluye la Tabla 4.15, que recoge las *accuracies* obtenidas, organizadas según el dataset utilizado.

2.4.2. Análisis de la combinación de CNN y LSTM en la detección de violencia y su exactitud con los principales datasets

Esta Sección analiza las arquitecturas de los trabajos seleccionados que utilizan la combinación de CNN (Fase 1) y LSTM (Fase 2) para la detección de violencia en vídeo, dado que es la combinación que mejores resultados ofrece según lo estudiado en la Sección 2.4.1. Dicha combinación se basa en que la CNN extrae características espaciales de los cuadros que componen el vídeo. Estas características espaciales extraídas alimentan a las capas LSTM para extraer características temporales de ellas (Mumtaz et al., 2022b; Vijeikis et al., 2022). A este respecto, las Tablas 2.14 y 2.15 presentan información detallada sobre las arquitecturas utilizadas en estos trabajos.

A continuación, se presentan las columnas que conforman la Tabla 2.14. En aquellos casos en que sus columnas tengan valores separados por barras «/», significa que el artículo ha entrenado diferentes combinaciones de arquitectura o hiperparámetros. Por otro lado, el valor «N.S.» significa que el valor no está claramente especificado en el trabajo.

- La columna ***CNN*** contiene las CNN empleadas para la extracción de las características espaciales de los *frames* de vídeo. Se puede observar que 11 de los 16 trabajos (68 %) utilizan redes preentrenadas, por lo que el uso de *transfer-learning* es ampliamente utilizado. Vale la pena señalar que de los trabajos seleccionados en la Tabla 2.14, hasta donde llega el conocimiento del autor de esta Tesis Doctoral, ninguno de ellos comprueba cuánto mejora la predicción de violencia el uso de LSTM, a diferencia del uso solo de CNN.

- La columna ***LSTM*** contiene el tipo de RNN utilizado que extrae las características temporales a partir de las caracterísitcas espaciales de cada *frame* extraídas por la CNN. Un 50 % utiliza LSTM, un 33 % Bi-LSTM, un 11 % Conv-LSTM y tan solo un 5.5 % GRU.

  Las capas LSTM son un tipo de unidad recurrente diseñada para superar el problema del desvanecimiento del gradiente en redes neuronales recurrentes; tiene una estructura celular que permite el almacenamiento y la recuperación de información a largo plazo, lo que lo hace efectivo para modelar secuencias temporales (Asad et al., 2021; Sharma et al., 2021; Vijeikis et al., 2022).

  Por otro lado, las capas Bi-LSTM son una variante de LSTM que procesan la secuencia de entrada en dos direcciones, combinando información contextual de ambos pasos temporales pasados y futuros, lo que puede ayudar a capturar patrones bidireccionales (Halder & Chatterjee, 2020). Finalmente, las Conv-LSTM combinan la estructura de celdas LSTM con la capacidad de procesamiento espacial de las capas convolucionales (Contardo et al., 2023; Jahlan & Elrefaei, 2021).

  Solo uno de los trabajos seleccionados contiene una comparación entre capas LSTM y capas Bi-LSTM (Gupta & Ali, 2022), donde se concluye que la combinación de VGG-16 y Bi-LSTM supera a la combinación de VGG-16 y LSTM. Se observa que empleando el dataset de *Hockey Fights*, el uso de Bi-LSTM con respecto a LSTM representa una *accuracy* del 97.6 % en comparación con el 98.8 % (aumento del 1.2 %) y en el *Violent Flow* dataset el uso de Bi-LSTM versus LSTM representa una *accuracy* del 92.2 % en comparación con el 95.1 % (aumento del 2.9 %).

- La columna ***Clasificación*** contiene el clasificador aplicado en los trabajos. En todos los casos, se utilizan las capas completamente conectadas (*fully connected layers* - FCL), excepto en el trabajo de Jahlan y Elrefaei (2021). Jahlan y Elrefaei

(2021) optan por utilizar varios clasificadores de Machine Learning clásicos para probar cómo varía la *accuracy* obtenida. Además, la función de activación *SoftMax* es la más empleada para la última capa densa, si bien se trata de una función de activación multi-clase, aun siendo la detección de violencia un problema binario.

- La columna ***Entrena Capas P.T.*** indica si los algoritmos que utilizan CNN PT entrenan algunas capas con los datasets de violencia. Ello sirve para ajustar los pesos preentrenados a esta nueva tarea de clasificación. Solo dos de los trabajos seleccionados realizan esto, siendo la investigación de Sharma et al. (2021) el único que especifica el número de capas que se descongelan para el entrenamiento con los datos de violencia (las cuatro últimas capas).

- La columna ***Capas LSTM*** contiene el número de capas LSTM imlpementadas; en varios de los trabajos su número no está claramente especificado, en ocho de ellos se utiliza una sola capa LSTM, y solo en dos de ellos se utilizan dos (Mumtaz et al., 2022b) y tres (Mugunga et al., 2021b) capas LSTM (argumentando que un mayor número de capas permite extraer patrones temporales más complejos).

- La columna ***Neuronas LSTM*** contiene el número de neuronas en las capas LSTM. Aunque existen varios trabajos que no especifican estos valores, generalmente se consideran típicas las cifras de 64, 128 y 512.

- La columna ***Capas densas*** contiene el número de capas densas que actúan como clasificadores; los valores típicos son el uso de entre dos y cuatro capas (incluida la capa de salida que debe contener dos neuronas, dado que el problema está dividido entre violencia y no violencia), donde un mayor número de capas densas implica una mayor capacidad de aprendizaje, pero también un mayor costo computacional y un mayor riesgo de sobreajuste (Mugunga et al., 2021b; Srivastava et al., 2022).

- La columna ***Neuronas densas*** contiene el número de neuronas que se implementan para las capas densamente conectadas, la cuales actúan como clasificador. Los valores típicos son 128, 256, y 512.

Por otro lado, se exponen a continuación las columnas que conforman los contenidos de la Tabla 2.15, la cual es la continuación de la Tabla 2.14, mostrando los mismos trabajos pero con nuevos parámetros en columnas adicionales:

- La columna ***Red P.T. usa FCL***, cuyo acrónimo significa «Red Preentrenada usa Fully Connected Layer», indica si los algoritmos que utilizan CNN preentrenadas mantienen las capas densas al realizar *transfer-learning* e implementar su arquitectura del algoritmo de detección de violencia. Como ya se discute anteriormente, el *transfer-learning* aprovecha las redes neuronales preentrenadas en datasets grandes para otros propósitos. Después de una serie de capas convolucionales, las CNN preentrenadas tienen una serie de capas densas que actúan como clasificador, con la última capa teniendo tantas neuronas como clases pueda predecir la CNN. Esta columna indica si el algoritmo de detección de violencia utiliza solo las capas convolucionales (en cuyo caso la columna tiene el valor «N» o No) o si contiene las capas convolucionales y las capas densas excluyendo la última capa (en cuyo caso la columna tiene el valor «S»). Si la columna tiene el valor *No P.T.*, significa que el algoritmo utiliza una CNN no preentrenada. En total, seis de los diez trabajos que utilizan CNN preentrenadas no incluyen las capas densamente conectadas para la extracción de características espaciales; los otros cuatro sí las mantienen.

- La columna ***Conexión CNN-LSTM*** indica cómo se transmiten las características espaciales extraídas de la CNN a la LSTM, aunque ninguno de los trabajos menciona explícitamente este aspecto al desarrollar la arquitectura del modelo. Este paso es relevante para la aplicación de algoritmos de inteligencia artificial explicables (XAI). Las CNN preentrenadas utilizadas en los trabajos seleccionados están entrenadas en datasets de imágenes, por lo que su estructura está diseñada para recibir una imagen, extraer información espacial de ella y, después de capas conectadas densamente, dar un resultado sobre la probabilidad de pertenencia a una clase específica. En general, las CNN (preentrenadas o no) suelen haber sido utilizadas para el análisis de imágenes, no de vídeo (ya que estos contienen información temporal), por lo que se debe encontrar una manera de hacer que la CNN termine de calcular todas las características espaciales contenidas en los *frames* de un vídeo para luego pasarlas como entrada a la LSTM. Las tres formas utilizadas en los trabajos mostrados en la Tabla 2.15 son las siguientes:

    - **Conv-LSTM**: las estructuras convolucionales-LSTM permiten que, mientras las CNN están extrayendo características espaciales de los *frames* de vídeo, simultáneamente se estén analizando para cada salto de tiempo las relaciones

temporales (Contardo et al., 2023; Jahlan & Elrefaei, 2021). Por lo tanto, esta arquitectura tiene su propia solución.

- **Concatenación de características espaciales**: esta solución consiste en entrenar e implementar los dos modelos por separado, por un lado la CNN y por otro la LSTM. De esta manera, cuando la CNN termine de analizar todos los *frames* de los vídeos, las características espaciales extraídas de todos los *frames* se agrupan de manera ordenada en un solo elemento (un array). Esto implica que los dos algoritmos (CNN y LSTM) se entrenen por separado. En caso de que la CNN esté preentrenada, es posible que no se la entrene nuevamente con los datos de violencia. Por otro lado, las capas LSTM y las capas densamente conectadas se entrenan de forma conjunta. Sin embargo, este proceso implica que el entrenamiento no se realiza en un solo bloque, si bien no se menciona en los trabajos citados que esto pueda afectar la *accuracy* del proceso completo (Mumtaz et al., 2022b; Srivastava et al., 2022).
- **Capa Distribuida en el Tiempo (*Time Distributed*)**: las capas *Time Distributed* de Keras permiten la ejecución de una capa de una red neuronal, varias capas o incluso un modelo completo para cada secuencia temporal (por cada frame, por ejemplo). Esto significa que es posible introducir un vídeo pre-procesado a una CNN, que la CNN procese cada uno de los *frames* o secuencias temporales y que hasta que todos estén terminados y encapsulados en el mismo elemento, la salida no se transmita a la siguiente capa —una LSTM, en este caso— (Sharma et al., 2021; Vijeikis et al., 2022). Este es un gran cambio con respecto al enfoque de *Concatenación de características espaciales*, ya que permite tener un único modelo entrenado de manera conjunta, realizándose todo el proceso unificado desde la entrada del vídeo preprocesado hasta la salida bi-clase que indica si la escena es violenta o no.

- La columna ***Learning Rate*** contiene el valor de la tasa de aprendizaje implementada en el proceso de entrenamiento del algoritmo. Los valores utilizados van desde 10^{-2} hasta 10^{-4}.

- Las columnas ***Hockey Fights***, ***Action Movies***, ***Violent Flow***, ***RWF-2000*** y ***RLVS*** contienen la *accuracy* obtenida durante el proceso de testeo por los trabajos que realizan el proceso de entrenamiento y testeo con estos datasets. Estos datasets

han sido seleccionados porque son los cinco datasets de agresiones físicas en vídeo más utilizados en la literatura desde junio de 2020 hasta junio de 2023 (Negre et al., 2024b), con las columnas ordenadas desde el más utilizado (*Hockey Fights*) hasta el menos utilizado (*Real Life Violence Situations*).

Se puede observar en la Tabla 2.15 que, aunque la mayoría de los trabajos han entrenado y testeado su modelo propuesto con el *Hockey Fights* dataset, el resto de los datasets son utilizados con mucha menor frecuencia. Por otro lado, el *Hockey Fights* dataset y el *Action movies* dataset, que son los dos datasets más utilizados, se basan en partidos de hockey y películas de acción, donde el enfoque y la iluminación de las personas son muy buenos. Esto implica que las calidad de las detecciones obtenidas por los trabajos seleccionados son muy altas en esos casos, pero más bajas en datasets con escenas reales, con cámaras de seguridad en diferentes ángulos y posiciones y condiciones de iluminación inferiores como las de los datasets de *RWF-2000* y RLVS.

Tabla 2.14: Trabajos seleccionados de detección de violencia que utilizan una combinación de CNN y LSTM. Parte 1.

Referencia	**CNN**	**LSTM**	**Clasificación**	**Entrena Capas P.T.**	**Capas LSTM**	**Neuronas LSTM**	**Capas Densas**	**Neuronas Densas**
Vijeikis et al. (2022)	MobileNet V2	LSTM	FCL (N.S.)	N	1	128	2	32, 2
Halder y Chatterjee (2020)	CNN de creación propia	Bi-LSTM	FCL (N.S.)	No P.T.	N.S.	N.S.	N.S.	N.S.
Aarthy y Nithya (2022)	PT VGG-16 (ImageNet)	LSTM	FCL (Sigmoid)	N	1	50	4	64, 64, 64, 2
Mugunga et al. (2021b)	PT VGG-16 (ImageNet)	Bi-Conv-LSTM	FCL (SoftMax)	N	3	256	4	1000, 256, 10, 2
Jahlan y Elrefaei (2021)	PT MNAS CNN	Conv-LSTM	Random Forest, SVM, K nearest neighbour	N	N.S.	N.S.	N FCL	N FCL
Traoré y Akhloufi (2020b)	Dos EfficienNet-B0 PT en ImageNet	Bi-LSTM	FCL (Sigmoid)	N	N.S.	N.S.	3	N.S.
Asad et al. (2021)	Dos VGG-16 PT (ImageNet) + Wide Dense Residual Blocks (WDRB)	LSTM	FCL (SoftMax)	N	1	512	1	2
Gupta y Ali (2022)	VGG-16 PT (ImageNet)	LSTM/Bi-LSTM	FCL (N.S.)	N	1	64	N.S.	N.S.
Ullah et al. (2022)	Darknet + Flujo Óptico Residual	M-LSTM	FCL (SoftMax)	No P.T.	N.S.	N.S.	N.S.	N.S.
Traoré y Akhloufi (2020a)	VGG-16 PT (INRA)	Bi-GRU	FCL (SoftMax)	N	N.S.	N.S.	3	512, 256, 2
Islam et al. (2021a)	CNN propia	LSTM	FCL (SoftMax)	No P.T.	1	128	3	256, 16, 2
Sharma et al. (2021)	Xception PT en ImageNet	LSTM	FCL (SoftMax)	S (4 last layers)	1	512	3	128, 32, 2
Talha et al. (2022)	CNN	LSTM	FCL	No PT	N.S.	N.S.	N.S.	N.S.
Mumtaz et al. (2022b)	VGG-19 PT en ImageNet + capas convolucionales extra	Bi-LSTM	FCL (SoftMax)	S (N.S.)	2	128, 64	N.S.	N.S.
Srivastava et al. (2022)	PT: VGG-16/ VGG-19/ InceptionV3...	LSTM	FCL (SoftMax)	N	1	512	3	1024, 512, 2
Contardo et al. (2023)	MobileNetV2 PT en ImageNet	Bi-LSTM/ ConvLSTM	FCL (SoftMax)	N	1/1	128/64	2	128/256

Tabla 2.15: Trabajos seleccionados de detección de violencia que utilizan una combinación de CNN y LSTM. Parte 2.

Referencia	**Red P.T. usa FCL**	**Conexión CNN-LSTM**	**Learning Rate**	**Hockey Fight**	**Action Movies**	**Violent Flow**	**RWF-2000**	**RLVS**
Vijeikis et al. (2022)	No P.T.	T.D.L	N.S.	99.5	96.1	–	82	–
Halder y Chatterjee (2020)	No P.T.	Concatenación	10^{-2}	99.27	100	98.6	–	–
Aarthy y Nithya (2022)	S	Concatenación	10^{-3}	99.1	–	–	–	–
Mugunga et al. (2021b)	N	No agregación (Conv-LSTM)	10^{-4}	99.1	100	98.4	92.4	–
Jahlan y Elrefaei (2021)	No FCL	No agregación (Conv-LSTM)	N.S.	99	96	100	–	–
Traoré y Akhloufi (2020b)	N	Concatenación	10^{-3}	99	–	93.8	–	96.7
Asad et al. (2021)	N	Concatenación	$10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$	98.8	98.99	97.1	–	–
Gupta y Ali (2022)	S	Concatenación	N.S.	97.6/98.8	–	92.2/95.1	–	–
Ullah et al. (2022)	No P.T.	Concatenación	10^{-4}	98	–	98.21	–	–
Traoré y Akhloufi (2020a)	N	Concatenación	10^{-4}	98	–	95.5	–	90.25
Islam et al. (2021a)	No P.T.	Concatenación	10^{-4}	97	100	–	–	–
Sharma et al. (2021)	N.S.	T.D.L	N.S.	96.55	98.32	–	–	–
Talha et al. (2022)	N	Concatenación	10^{-2}	94.9	92.2	77.31	–	–
Mumtaz et al. (2022b)	S	Concatenación, Flatten	10^{-2}	91.29	–	90.5	90.5	–
Srivastava et al. (2022)	S	Concatenación	10^{-2}	–	–	–	–	–
Contardo et al., 2023	N	T.D.L, Flatten	N.S.	–	–	–	–	–

Con todo, a pesar de la gran cantidad de estudios existentes, no parece haber un consenso claro sobre una combinación óptima de hiperparámetros para implementar algoritmos de detección de violencia. Factores como la tasa de aprendizaje, el número de neuronas en las capas LSTM o en las capas densas continúan siendo objeto de exploración y ajuste en la literatura. Por otro lado, solo el trabajo de (Mumtaz et al., 2022b) ha abordado específicamente la ventaja potencial del uso de capas Bi-LSTM en comparación con las capas LSTM convencionales.

2.5. Uso de IA fiable en la detección de violencia

Esta Sección analiza si los algoritmos utilizados en los trabajos seleccionados emplean técnicas de inteligencia artificial fiable o, en su defecto, técnicas basadas en alguno de sus pilares. En la introducción del Capítulo 2 se destaca la importancia de desarrollar algoritmos de IA basados en técnicas de inteligencia artificial fiable, de modo que su funcionamiento y resultados puedan ser entendidos, en contraposición a algoritmos opacos y no interpretables. Junto con documentos elaborados por otras importantes organizaciones, el informe de la Comisión Europea (European Commission, 2019) define que la inteligencia artificial fiable debe estar asentada sobre tres pilares: *licitud, eticidad y robustez.*

En primer lugar, se buscan **menciones de palabras clave que conforman la inteligencia artificial fiable** según el informe de la Comisión Europea (European Commission, 2019): `trustworthy, lícita, ethical, robust, and explainable` (como pilar de la eticidad). Posteriormente, se analizan aquellos que, mencionando o no estos términos, incluyen la fiabilidad de alguna manera. De los 63 trabajos seleccionados en total, 33 artículos (52 %) no hacen uso ni referencia a ninguno de los términos. Por otro lado, del total de trabajos, 30 artículos (47 %) sí nombran alguno de los términos, si bien ninguno realiza una implementación de técnicas de detección de violencia basado en inteligencia artificial fiable o en alguno de sus pilares.

La Tabla 2.16 presenta dónde se encuentran localizadas en los trabajos seleccionados las menciones de estas palabras clave, ordenadas de mayor a menor número de menciones.

- La columna *Componentes de Fiabilidad* contiene los elementos mencionados que conforman la inteligencia artificial fiable.

Tabla 2.16: Mención o utilización de IA Fiable en los trabajos seleccionados.

Componentes de Fiabilidad	**Localización de la Mención**	**Conteo de Citas**
Robustez	Mención	15
	Modelo	10
Eticidad	Dataset	1
Explicabilidad	Modelo	1

- La columna *Localización de la mención* indica en qué parte del artículo se menciona ese término.
  - El término *Mención* en dicha columna significa que la palabra clave se menciona en la introducción, el trabajo relacionado o la conclusión, pero este termino no se refiere al modelo propuesto en el trabajo.
  - El término en la columna *Modelo*, significa que la palabra clave se refiere al modelo desarrollado en el trabajo.
  - El término en la columna *Dataset*, significa que se hace referencia al dataset empleado en el trabajo.
- La columna *Conteo de Citas* contiene el número de los trabajos seleccionados que mencionan esos términos.

El término *Mención* en dicha columna significa que la palabra clave se menciona en la introducción, el trabajo relacionado o la conclusión, pero este término no se refiere al modelo propuesto en el trabajo. Si el término en la columna es *Modelo*, significa que la palabra clave se refiere al modelo desarrollado en el trabajo. Por último, si el término en la columna es *Dataset*, significa que se hace referencia al dataset empleado en el trabajo. Finalmente, la columna *Conteo de Citas* contiene el número de los trabajos seleccionados que mencionan esos términos.

Se puede observar cómo en la Tabla 2.16 el término más citado es *Robustez*, tanto como mención en las secciones introductorias/conclusivas como en referencia a los modelos propuestos en dichos trabajos. Esto se debe a que la detección de violencia es principalmente un problema binario, es decir, existe violencia en la escena o no. Se discute en la Sección 2.2 cómo un método ampliamente utilizado para evaluar los resultados es el uso de una matriz de confusión binaria, lo que permite comprobar cuán

robusto es el algoritmo frente a otras acciones que aparecen en el vídeo. Ello se debe a que la detección de violencia es compleja (Omarov et al., 2022), dado que puede confundirse con acciones no violentas que pueden parecer agresiones físicas, como un abrazo. Por lo tanto, la robustez es un término ampliamente citado debido tanto a la naturaleza binaria del problema como a su ambigüedad con respecto a otras acciones.

El concepto de *ética* solo es mencionado por Nadeem et al. (2023), quienes abordan los problemas éticos involucrados en el etiquetado de escenas violentas para la generación de datasets enfrentando a las personas involucradas en los vídeos. Por lo tanto, los autores proponen el desarrollo de un algoritmo de etiquetado automático con garantías éticas. La mención de *explicabilidad*, como elemento que conforma el pilar de la eticidad, también es mucho menor. Hu et al. (2022) abordan el problema de los algoritmos de aprendizaje profundo de «caja negra» o «*black-box*», en inglés, y señalan la cuantificación de la incertidumbre (UQ) para medir el nivel de incertidumbre de estos algoritmos, utilizando el método de *Monte Carlo* y la Validación Cruzada Jerárquica. No se encuentra una mención directa del término *fiable* o *lícita*.

En resumen, no se hace uso de técnicas de inteligencia artificial fiable en los trabajos seleccionados con la intención de realizar más ético, robusto o lícita el proceso de detección de violencia en vídeo, si bien sí se hace mención a ciertos términos, como se expone en la Tabla 2.16. Sin embargo, **sí se observan en los trabajos seleccionados técnicas empleadas en las arquitecturas presentadas que aportan de forma intrínsica cierta explicabilidad al proceso**. Se exponen a cotinuación dichas técnicas, asociadas al paso de la detección de violencia en las que se implementan.

Como uno de los pilares que conforman la eticidad, la XAI se centra en aportar comprensión y explicación sobre el proceso de toma de decisiones de un modelo de inteligencia artificial. Ello resulta clave, ya que la opacidad en los modelos de IA puede ser un desafío para la aceptación y adopción, especialmente en aplicaciones donde comprender el razonamiento detrás de las decisiones automatizadas es fundamental (Chamola et al., 2023), como es el caso de la detección de violencia en vídeo.

Con base en los pasos fundamentales para la detección de violencia en vídeo, ilustrados en la Figura 2.1, se expone la explicabilidad intrínseca de los algoritmos implementados en los trabajos seleccionados para cada uno de estos pasos:

En el proceso de **extracción de fotogramas característicos**, que se aborda en la Sección 2.3, se categorizan los métodos de extracción de fotogramas potenciales de contener violencia en la Tabla 2.10 según su funcionamiento. Los algoritmos de la categoría de *detección de objetos* se especializan en identificar personas u objetos en vídeo. Un ejemplo destacado de esta clase de algoritmos es *Yolo* (Diwan et al., 2023). Estos métodos contribuyen a la explicabilidad del sistema, ya que permiten entender fácilmente por qué ciertos fotogramas son seleccionados como potenciales indicadores de violencia, al detectar un número significativo de personas en la escena registrada.

Por otro lado, también se abordan en la Sección 2.3 y, específicamente, en la Tabla 2.11, las entradas de los algoritmos empleados, categorizadas según si estaban centradas en la imagen, centradas en el movimiento o centradas en el audio. Las entradas de los algoritmos centrados en el movimiento pueden proporcionar una mayor explicabilidad, como en el caso de la *imagen de energía de movimiento separada* (Mumtaz et al., 2022b), que enfatiza las áreas con cambios rápidos de intensidad de píxeles en los cuadros. Estas entradas pueden ayudar a comprender las áreas más destacadas que el algoritmo probablemente analizará para realizar la clasificación de la escena.

Finalmente, en el proceso de **detección de violencia**, el uso de algoritmos como los *transformers*, CNN y LSTM implica una comprensión nula o extremadamente baja sobre el proceso de toma de decisión. Por otro lado, los algoritmos de aprendizaje automático, así como la extracción manual de características, tienen una lógica más simple que puede ser algo más comprensible. Sin embargo, la elección ideal sería el uso de algoritmos complejos que sean al mismo tiempo altamente explicables. En este sentido, los algoritmos basados en esqueletos son altamente comprensibles, ya que su propósito es detectar las articulaciones de las personas presentes en el vídeo y su evolución a lo largo del tiempo (Hung et al., 2021; Zhou, 2022). Por lo tanto, uno de los algoritmos en los trabajos seleccionados que tiene un mayor grado de explicabilidad es el desarrollado por Naik y Gopalakrishna (2021). En ese artículo, los autores utilizan mapas de calor para resaltar a las personas detectadas en el vídeo en el preprocesamiento, además de marcar las articulaciones en código de colores *verde, amarillo y rojo*, que indicaban no violencia, violencia potencial y violencia definitiva.

En general, la preocupación de los trabajos seleccionados por el desarrollo de una inteligencia artificial fiable para la detección de agresiones físicas en vídeo es

prácticamente nula. Se observa que el concepto de robustez se utiliza para evaluar los buenos resultados de los algoritmos desarrollados, dado que el problema es binario y se mide utilizando una matriz de confusión. Sin embargo, este concepto de robustez no se está utilizando desde la perspectiva de pilar de IA fiable. Se identifican en esta Sección algunos algoritmos que pueden mejorar la comprensión y explicabilidad del proceso, aunque el objetivo principal de los autores no fuese aumentar la explicabilidad, sino hacer más clara la explicación para el lector. Esto refleja la necesidad de investigar en técnicas que mejoren la explicabilidad en la detección de violencia en vídeo. Por ello, esta Tesis Doctoral incluye en sus hipótesis y objetivos, expuestos en la Sección , la investigación en este ámbito, y el Capítulo 5 diseña e implementa una arquitectura explicable de detección de violencia en vídeo para abordar esta problemática.

2.6. Discusión

La investigación realizada en este Capítulo ofrece una visión detallada del estado actual en la detección de violencia en vídeo mediante el uso de inteligencia artificial, enfocándose en la evolución de los métodos y técnicas aplicadas a esta problemática, así como en la integración de principios de IA fiable. Este análisis exhaustivo permite extraer conclusiones clave sobre la problemática existente y las carencias del estado del arte para esta Tesis Doctoral, así como sugerir direcciones para futuras investigaciones en este campo.

- **La detección de violencia en vídeo conforma un área de investigación crucial que busca abordar las agresiones físicas a través del análisis automatizado de imágenes y vídeos**. Las agresiones físicas son una preocupación generalizada en la sociedad, que afecta a individuos globalmente e impacta casi todos los aspectos de la vida (Yao & Hu, 2023). Esta influencia se extiende no solo a las víctimas inmediatas, sino también a sus familias y la comunidad en general. La detección de violencia basada en inteligencia artificial es la última barrera para defender a las víctimas de sus ataques y puede realizar vigilancia a gran escala constantemente con el tiempo (Mugunga et al., 2021b). *En este sentido, la introducción de este Capítulo expone el problema que suponen las agresiones físicas para la sociedad en su conjunto, así como las propuestas con*

enfoques temporales a largo, medio y corto plazo que se han propuesto y estudiado con el tiempo (Martínez-González et al., 2021; Vomfell et al., 2018).

- **No existe una revisión del estado del arte que analice de manera integral todas las etapas del proceso de detección de violencia en vídeo, desde el preprocesamiento hasta la toma de decisiones, ni que aborde los retos actuales desde la perspectiva de la IA fiable** (Negre et al., 2024a). La detección de violencia implica múltiples fases tanto en el entrenamiento de los modelos como en su implementación en tiempo real, por lo que un análisis actualizado y completo es esencial para evaluar la aplicabilidad y fiabilidad de las soluciones actuales. *De acuerdo con esta observación, la Sección 2.1 analiza los trabajos de revisión publicados entre julio de 2020 y julio de 2023, además de planificar y desarrollar un estudio sistemático de la literatura que cubre las necesidades de una revisión actualizada y completa (Negre et al., 2024b).*

- **El análisis realizado destaca los principales retos presentes en los trabajos recientes sobre la detección de violencia en vídeo mediante inteligencia artificial, subrayando, además, la necesidad de desarrollar una categorización exhaustiva de los algoritmos utilizados en este campo.** (Negre et al., 2024a). Los retos más mencionados en los trabajos seleccionados incluyen la *oclusión de elementos*, *variación en la iluminación*, el *costo de procesamiento en tiempo real*, la *posición de la cámara* y el *fondo no estacionario*. Estos desafíos son cruciales para el desarrollo de sistemas de detección más precisos y eficientes en entornos reales.

  Por otro lado, se ha propuesto una categorización de los algoritmos utilizados en la detección de violencia en vídeo, que incluye: algoritmos manuales, redes neuronales convolucionales (CNN), redes neuronales recurrentes LSTM, modelos basados en esqueletos, Transformer y algoritmos basados en audio. Esta clasificación permite entender mejor la diversidad de enfoques existentes y sus aplicaciones en función de las características específicas de cada problema.

  En este sentido, la Sección 2.2 (Negre et al., 2024b) profundiza en estos aspectos, presentando un análisis detallado de los algoritmos y los desafíos asociados a la detección de violencia mediante inteligencia artificial.

- **La selección adecuada de datasets y entradas es fundamental para el desarrollo eficiente de algoritmos de detección de violencia en vídeo, dado que las agresiones físicas son eventos poco comunes y difíciles de identificar en las grabaciones**. Existen diversos datasets públicos de calidad para la detección de agresiones físicas, pero la naturaleza poco frecuente de estos eventos plantea desafíos para obtener datos representativos (Jaiswal & Mohod, 2021). Esta variabilidad es útil para abordar requisitos específicos, pero complica la recopilación de datos de alta calidad para entrenar modelos de IA confiables.

  Un reto clave en la detección en tiempo real es la selección eficiente de fotogramas relevantes. Debido a que gran parte de la grabación no muestra violencia (Ahmed et al., 2021), identificar solo los fotogramas que contienen eventos violentos es esencial para reducir costos computacionales. Sin embargo, solo el 33 % de los trabajos revisados presentan métodos específicos para esta tarea.

  Respecto a las entradas de los algoritmos, se utilizan diversos tipos de datos, como vídeo RGB, escala de grises y *optical flow*. El vídeo RGB es el más común, seguido por el *optical flow*, que aparece en significativamente menos trabajos.

  A este respecto, la Sección 2.3 subraya la importancia de contar con una diversidad adecuada de datasets y de emplear métodos eficientes de selección de fotogramas para avanzar en la detección de agresiones físicas en vídeo. Además, la Sección aborda en detalle los enfoques utilizados para mejorar la selección de fotogramas y las entradas del algoritmo, contribuyendo al desarrollo de soluciones más efectivas y eficientes en este campo.

- **El diseño e implementación de arquitecturas basadas en CNN combinadas con capas LSTM demuestran ser una aproximación prometedora, a pesar de que quedan áreas del estado del arte por investigar**. Algunos aspectos del estado del arte todavía no abordados son la mejora que supone el uso de CNN junto con LSTM con respecto al uso de CNN por sí solas, al igual que la mejora que supone la utilización de capas LSTM con respecto a capas Bi-LSTM. Por otro lado, hasta donde el autor de esta Tesis Doctoral tiene concocimiento, no se ha estudiado si existe un cierto número de neuronas en las capas LSTM y capas densas con las que se obtenga una mejor *accuracy. Alienado con este respecto, el Capítulo 3, 4 y 5 se dedica al desarrollo y análisis exhaustivo de estas arquitecturas, con el objetivo de explorar en mayor*

profundidad los aspectos que aún no han sido completamente abordados en el estado del arte.

- **La incorporación de técnicas de IA fiable en la detección de violencia en vídeo es esencial para mejorar la transparencia, interpretabilidad y robustez de los modelos, sin embargo, hasta la fecha, no se han implementado explícitamente en las arquitecturas existentes**. Los algoritmos de aprendizaje profundo son en su mayoría cajas negras que no proveen fiabilidad sobre su proceso de toma de decisiones (Chamola et al., 2023), lo que ha impulsado el desarrollo de enfoques centrados en la inteligencia artificial fiable. En esta línea, organismos como la Comisión Europea han definido estándares que incluyen la eticidad, la lícita y la robustez como pilares fundamentales (European Commission, 2019).

  A pesar de que 15 trabajos revisados mencionan la robustez de sus modelos, ninguno aplica explícitamente técnicas de IA fiable con el objetivo de hacerlos más confiables (Negre et al., 2024b). No obstante, algunos trabajos emplean estrategias que pueden contribuir a la explicabilidad del proceso de detección, como el uso de algoritmos de detección de objetos para la extracción de *frames* característicos o la implementación de modelos basados en esqueletos para analizar el movimiento (Speith, 2022).

  En respuesta a esta necesidad, la Sección 2.5 analiza los algoritmos utilizados en los trabajos seleccionados que pueden facilitar la comprensión del proceso de detección de violencia en vídeo, reforzando la importancia de desarrollar enfoques que alineen estos modelos con los principios de la IA fiable. Asimismo, el Capítulo 5 diseña e implementa una arquitectura de detección de violencia en vídeo basada en la red VGG-19 preentrenada combinada con capas LSTM/Bi-LSTM, integrando técnicas de explicabilidad como GradCAM para la generación de mapas de saliencia.

Con todo, a partir de lo concluido en el Capítulo 2, se diseñan y desarrollan nuevas arquitecturas propuestas en esta Tesis Doctoral en los Capítulos 3, 4 y 5, cuyo contenido se muestra resumido en la Figura 2.9. El diseño, implementación y estudio de dichas arquitecturas abordan específicamente los problemas identificados en el uso combinado de CNN con capas LSTM y la implementación de algoritmos de XAI en la detección de violencia en vídeo.

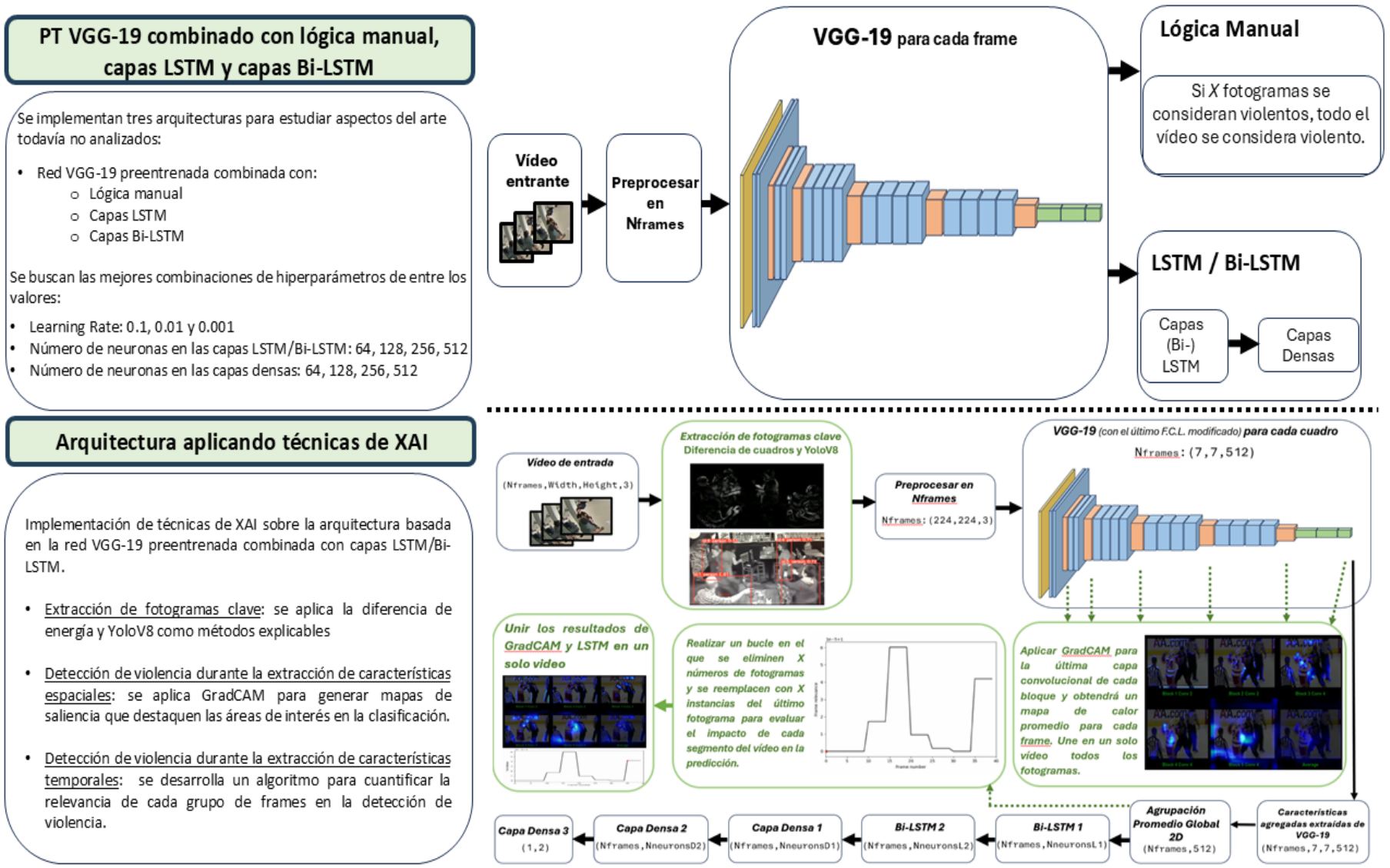

FIGURA 2.9: Arquitecturas propuestas para la detección de violencia en vídeo en el uso de técnicas que combinan CNN y capas LSTM aplicando posteriormente técnicas de XAI.

Capítulo 3

CNN y su combinación con un número mínimo de fotogramas violentos

Capítulo 3

CNN y su combinación con un número mínimo de fotogramas violentos

Índice

Las agresiones físicas representan un problema significativo en nuestra sociedad, afectando a individuos a nivel global (Long et al., 2021). La detección de violencia en vídeos mediante inteligencia artificial (IA) se ha convertido en un área de investigación relevante, especialmente para la protección de las víctimas. Entre los diversos enfoques disponibles, la combinación de Redes Neuronales Convolucionales (CNN) y Redes de Memoria a Corto y Largo Plazo (LSTM) ha demostrado ser particularmente efectiva, logrando excelentes resultados (Negre et al., 2024a). Sin embargo, se considera relevante cuantificar el impacto que supone combinar consecutivamente las CNN con capas LSTM, en comparación con el uso exclusivo de CNN, pudiendo mejorar la comprensión sobre la detección de violencia fotograma a fotograma y su combinación con otro tipo de arquitecturas.

Por ello, el *objetivo general* del Capítulo es estudiar la relevancia de las características extraídas exclusivamente por CNN frente a arquitecturas que combinan CNN con LSTM. Como se expone en el Capítulo 2, a diferencia de otros algoritmos basados en técnicas

de *Machine Learning* tradicional o heurísticas para la extracción de características manuales, las CNN tienen una gran capacidad para la extracción de características espaciales, siendo grandes candidatas para la detección de violencia en vídeo. Hasta donde se tiene conocimiento, ninguno de los estudios analizados ha demostrado de manera concluyente si el uso de LSTM mejora significativamente la precisión (*accuracy*) obtenida con respecto al uso exclusivo de CNN. Ello se considera relevante para cuantificar la capacidad individual de las CNN en la detección de violencia así como la futura propuesta de otras arquitecturas. *Todo ello en concordancia con los objetivos* ***OB2*** *y* ***OB4*** *establecidos en esta Tesis Doctoral en la Sección 2.5.*

En este contexto, el Capítulo aborda tres **objetivos específicos**, teniendo en cuenta que los Capítulos 4 y 5 continuan con el desarrollo de las soluciones propuestas en esta Tesis Doctoral:

1. Entrenar y testear las CNN preentrenadas en *ImageNet* VGG-16 y VGG-19 para clasificar fotogramas como violentos o no violentos en los datasets seleccionados para esta Tesis Doctoral.

2. Desarrollar y analizar una arquitectura combinada, donde las CNN preentrenadas VGG-16 y VGG-19 preentrenada se integren con una lógica manual. Esta última considera un vídeo como violento solo si un número mínimo de fotogramas es clasificado como violento por la red.

3. Estudiar de entre VGG-16 y VGG-19 cual obtiene mejores resultados tanto en la detección fotograma a fotograma de vídeos violentos, como con el uso de un mínimo número de fotogramas violentos, con el fin de seleccionar dicha CNN para su uso en los Capítulos 4 y 5.

Los **resultados obtenidos** demuestran cómo en la implementación de las arquitecturas propuestas, la red VGG-19 preentrenada obtiene mejores resultados que VGG-16, tanto fotograma a fotograma, como al establecer un número mínimo de fotogramas violentos, con un 95 % de *accuracy* en el dataset *Hockey Fights*, un 96 % de *accuracy* usando el dataset *Violent Flow* y un 96 % de *accuracy* usando el dataset *RWF-2000*. Por su parte, VGG-16 obtiene el mismo resultado en el dataset *Hockey Fights*, pero un 5 % menos en el dataset *Violent Flow* y un 10 % menor en el dataset *RWF-2000*.

Este Capítulo se divide en cinco Secciones. La Sección 3.1 justifica la elección de la CNN utilizada en el desarrollo e implementación de las arquitecturas presentadas en

esta Tesis doctoral. Además, describe en detalle las arquitecturas de CNN candidatas, VGG-16 y VGG-19, y su integración con un número mínimo de fotogramas detectados como violentos. La Sección 3.2 expone el entrenamiento de las redes CNN seleccionadas. La Sección 3.3 contiene los resultados del testeo de las redes CNN seleccionadas, así como los de su combinación con un número mínimo de fotogramas detectados como violentos. La Sección 3.4 analiza los resultados obtenidos por los modelos desarrollados en esta Tesis Doctoral entre sí y con otros trabajos seleccionados que implementen arquitecturas similares. La Sección 3.5 aborda las conclusiones del Capítulo.

3.1. Arquitectura basada en CNN con número mínimo de fotogramas violentos

Esta Sección muestra las arquitecturas de las CNN seleccionadas para su estudio en la detección de violencia en este Capítulo, tanto fotograma a fotograma, como con el uso combinado de un número mínimo de fotogramas violentos para considerar la escena como violenta. La Sección 3.1 analiza las arquitecturas de las CNN seleccionadas, las redes preentrenadas VGG-16 y VGG-19, con el objetivo de determinar cuál de ellas ofrece un mejor rendimiento en la detección de fotogramas violentos dentro de los datasets seleccionados a lo largo del Capítulo. La Sección 3.1.2 expone la estructura de una CNN preentrenada y su combinación con un número mínimo de fotogramas clasificados como violentos.

3.1.1. Selección de la arquitectura CNN: arquitectura de VGG-16 y VGG-19

Esta Sección presenta la justificación de la elección de las arquitecturas VGG-16 y VGG-19 preentrenadas como CNN a emplear para el diseño e implementación de las arquitecturas de esta Tesis Doctoral.

Tal y como se analiza en el Capítulo 2, la combinación de redes convolucionales (CNN) con capas LSTM es una estrategia ampliamente utilizada y particularmente efectiva en comparación con otras soluciones. En la Tabla 2.13, que recoge las *accuracies* obtenidas en la detección de violencia en vídeo por los trabajos seleccionados —ordenados según el conjunto de datos empleado en el entrenamiento—, se observa una notable diversidad de CNN utilizadas en el estado del arte, siendo VGG-16 la más empleada.

En todos los casos en los que se utiliza VGG-16, esta se emplea como red preentrenada; es decir, no se entrena desde cero, sino que se aplican técnicas de ajuste fino (*fine-tuning*), reentrenando la red con vídeos violentos para adaptar sus pesos a la tarea específica de detección de violencia en vídeo. La red VGG-16 fue desarrollada por el *Visual Geometry Group* de la Universidad de Oxford, que también desarrolló la red VGG-19 Simonyan y Zisserman (2014). Esta última se compone de dieciséis capas convolucionales y tres capas densamente conectadas, frente a las trece capas convolucionales que posee la VGG-16.

Diversos estudios han demostrado que un mayor número de capas convolucionales permite extraer características más complejas de las imágenes. Ello queda patente entre las redes VGG-16 y VGG-19 al ser preentrenadas en el dataset de imágenes ImageNet, en el que la red VGG-19 obtiene mejores resultados al clasificar imágenes que la red VGG-16 Simonyan y Zisserman (2014). No obstante, a pesar de que la detección de violencia en vídeo constituye una tarea compleja y altamente variable, VGG-19 solo ha sido utilizada en dos de los trabajos seleccionados en el Capítulo 2, concretamente por Mumtaz et al. (2022b) e Islam et al. (2021a). Los resultados obtenidos por estos estudios muestran que VGG-16 alcanza rendimientos iguales o superiores a VGG-19 en datasets como *Hockey Fights*, *Action Fights* y *Violent Flow*, superando en ambos casos a otras CNN empleadas en el estado del arte.

Por todo ello, se considera relevante y de interés estudiar las redes VGG-16 y VGG-19 preentrenadas como CNN potencial a empelear en las arquitecturas a diseñadas e implementadas en esta Tesis Doctoral, con el objetivo de verificar si el mayor número de capas convolucionales permite mejorar los resultados obtenidos por VGG-16 en trabajos previos del estado del arte.

Pasando ahora al **análisis de las arquitecturas de VGG-16 y VGG-19**, se tratan redes neuronales convolucionales profundas que consta de 16 y 19 capas, respectivamente, conocidas por su simplicidad y eficacia, utilizando únicamente filtros convolucionales de tamaño 3x3 y *max-pooling* para reducir la dimensionalidad (Simonyan & Zisserman, 2014). Tanto VGG-16 como VGG-19 se han destacado por su alto rendimiento en tareas de clasificación de imágenes y son ampliamente utilizadas en aplicaciones de visión por computadora.

En la Figura 3.1 se representa la estructura de VGG-19 donde se puede observar cómo espera recibir una matriz de forma (224, 224, 3), lo que indica 224 píxeles de ancho y alto para cada canal de color RGB. Luego pasa por una serie de bloques convolucionales,

que están compuestos por capas convolucionales consecutivas; las capas convolucionales utilizan filtros o núcleos que consisten en una serie de pesos que se aplican en la matriz de píxeles que es la imagen, a partir de la cual la red neuronal aprende patrones cada vez más complejos.

Entre cada bloque convolucional se intercalan capas de *Max Pooling* que reducen la dimensión de la matriz de píxeles que es la imagen, seleccionando un valor máximo de cada región. Una vez que la imagen ha pasado por todos los bloques convolucionales, se introduce a tres capas densamente conectadas que actúan como clasificador. En la Figura 3.1 se puede ver cómo la última capa completamente conectada genera una matriz con la forma (1, 1, 1000). Esto se debe a que cuando VGG-19 se entrena en el dataset de imágenes *ImageNet*, debe clasificar entre un total de 1000 clases. Una vez que se genera la salida de la última capa, se aplica la función *SoftMax* para normalizar la salida generada, asegurando que la suma de las probabilidades sea igual a 1.

En la Figura 3.2 se representa la estructura de VGG-16, cuya estructura es muy similar a la de VGG-19 con la principal diferencia en el número de capas convolucionales: VGG-16 posee 13 capas convolucionales en lugar de las 16 que conforman VGG-19. Ambas arquitecturas comparten la misma filosofía de diseño, utilizando filtros pequeños de tamaño 3×3, pasos de convolución (strides) de 1 y padding para conservar la resolución espacial. La reducción progresiva del tamaño espacial se logra a través de las capas de *Max Pooling*, y al final, ambas redes emplean tres capas completamente conectadas para la clasificación. La diferencia principal radica en la profundidad de la red, lo que implica que VGG-19 es ligeramente más costosa computacionalmente y puede capturar patrones de mayor complejidad, mientras que VGG-16 ofrece un mejor compromiso entre rendimiento y eficiencia en muchos escenarios prácticos.

Para poder emplear VGG-16 y VGG-19 preentrenadas en *ImageNet* para la detección de violencia, se vuelve necesario alterar la última capa completamente conectada, que originalmente contiene 1000 neuronas, a una sola neurona, manteniendo los pesos de la red preentrenada. Este ajuste es necesario debido a que la detección de violencia es un problema binario (violencia o no violencia). Dicha capa con una neurona deberá entrenarse con el dataset de violencia para poder actuar como clasificador correctamente.

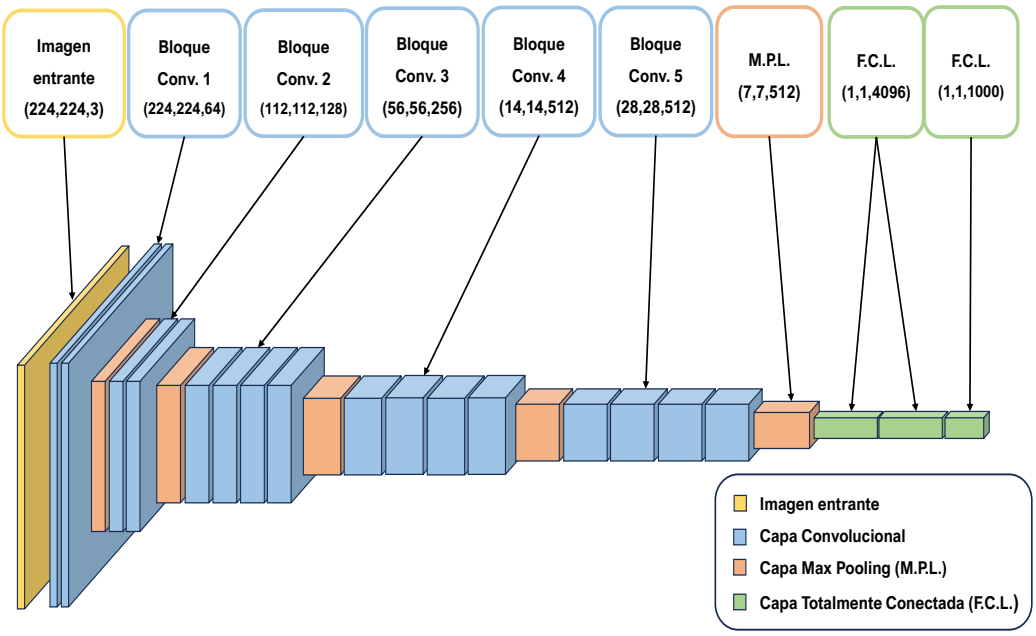

FIGURA 3.1: Estructura de la red VGG-19.

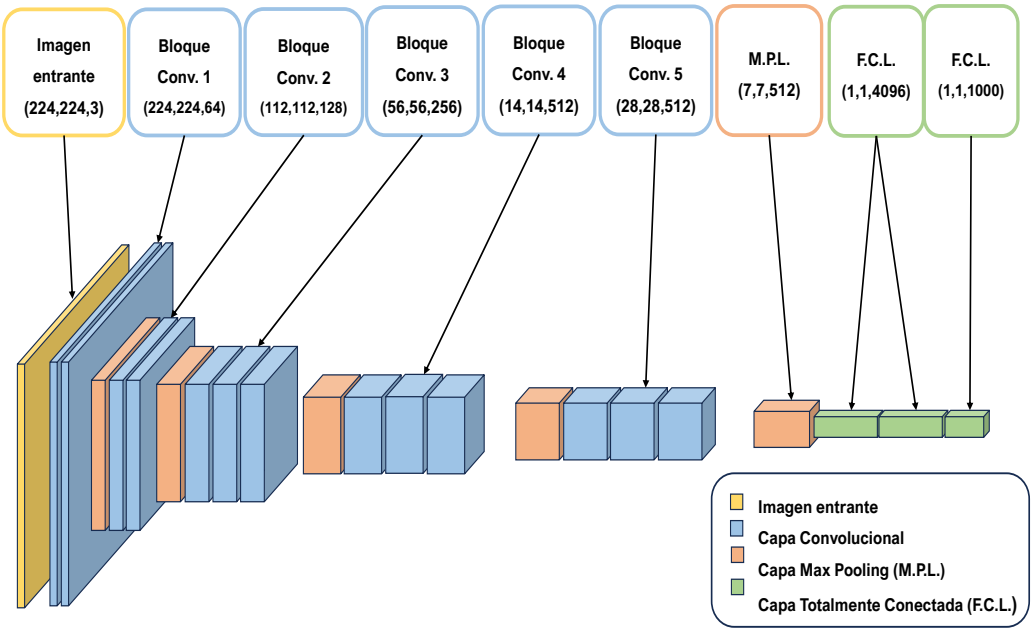

FIGURA 3.2: Estructura de la red VGG-16.

3.1.2. Arquitectura basada en CNN con número mínimo de fotogramas violentos

Esta Sección expone la arquitectura basada en CNN combinada con un número mínimo de fotogramas clasificados como violentos. Esta arquitectura se propone frente a la

propuesta de realizar la detección de violencia mediante el uso combinado de CNN junto con capas LSTM (o cualquier otro tipo de RNN). Esta propuesta es relevante con el fin de cuantificar la verdadera relevancia que suponen las CNN por sí mismas respecto al uso combinado con capas LSTM. Ello también es de importancia de cara a que el análisis fotograma a fotograma con un umbral mínimo de valores para considerar la escena violenta, facilita la detección de violencia en tiempo real, al no tener que almacenar en memoria los fotogramas que van siendo recibidos en tiempo real, si no únicamente los valores de las predicciones. Por ello, se proponen dos métodos destinados a escoger cuál de las redes CNN preentrenadas, VGG-16 o VGG-19, resulta más adecuada para su uso en tareas de detección de fotogramas violentos. Estos métodos permiten, además, cuantificar la relevancia de cada arquitectura por sí sola, sin recurrir a la combinación con capas LSTM.

- El primer método consiste en **analizar el número de fotogramas que la CNN detecta como violentos y no violentos**.
- El segundo método consiste en **establecer un límite en el número de fotogramas detectados como violentos, más allá del cual se considera el vídeo como violento**, con el fin de analizar conjuntamente un vídeo después de la predicción realizada por la CNN fotograma a fotograma. Es relevante señalar que en el análisis realizado, no se ha encontrado ninguna solución como la propuesta.

Se pasa a exponer la arquitectura que combina CNN con un mínimo de fotogramas clasificados como violentos, mostrada en la Figura 3.3. El proceso comienza con la entrada de un vídeo, cuyos fotogramas son preprocesados y luego introducidos individualmente en una red CNN. Esta red ha sido modificada, eliminando sus capas completamente conectadas originales y sustituyendo la última por una capa densa con una neurona, entrenada específicamente sobre datasets de violencia. Cada fotograma se clasifica como «violento» y «no violento» mediante esta red. Posteriormente, se aplica una lógica de decisión basada en umbrales: si un número X de fotogramas supera cierto umbral de probabilidad de violencia, se clasifica todo el vídeo como violento.

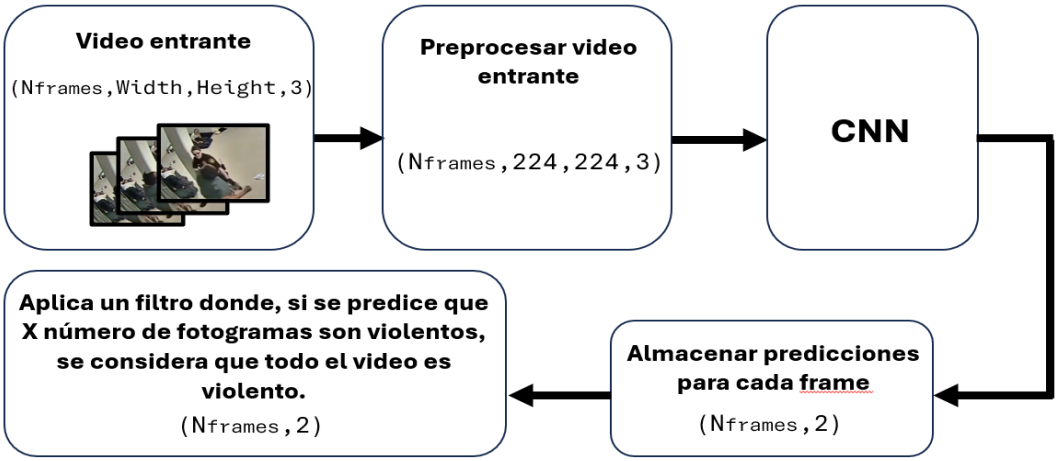

FIGURA 3.3: Modelo de detección de violencia utilizando una red CNN combinada con un número mínimo de fotogramas violentos.

3.2. Entrenamiento de CNN con número mínimo de fotogramas violentos

Esta Sección trata los datasets empleados para el entrenamiento y testeo de las arquitecturas de este Capítulo, además de todas las arquitecturas que se desarrollan en esta Tesis Doctoral, así como el entrenamiento de las CNN seleccionadas VGG-16 y VGG-19 preentrenadas. La Sección 3.2.1 presenta los datasets empleados durante todos los entrenamientos y testeos de las arquitecturas propuestas en esta Tesis Doctoral. La Sección 3.2.2 expone los resultados obtenidos durante la fase de entrenamiento de la arquitectura de las CNN VGG-16 y VGG-19 preentrenadas en el dataset de *ImageNet*.

3.2.1. Datasets utilizados en esta Tesis Doctoral

En esta Tesis Doctoral se emplearán los datasets *Hockey Fights*, *Violent Flow* y *Real World Fight 2000* debido a los siguientes motivos: *(i)* por un lado, entrenar y testear algoritmos con datasets ampliamente citados favorece la facilidad de comparar resultados con otros trabajos *(ii)* por otro, el uso de datasets que contengan escenas reales asegura la capacidad de detección realista de los algoritmos propuestos.

Por un lado, el dataset *Hockey Fights* (Bermejo Nievas et al., 2011) es elegido por ser el dataset más empleado en la literatura reciente si tenemos en cuenta el número de citas (Negre et al., 2024b). Este contiene vídeos de acción de juegos de hockey de la Liga Nacional de Hockey (NHL). Aunque las peleas son reales, los vídeos a menudo contienen primeros planos, están bien iluminados y no representan una situación de

Tabla 3.1: Características de los datasets seleccionados.

Nombre	**Año**	**Número de vídeos**	**Número Medio de *Frames***	**Frames por segundo (fps)**
Hockey fights	2011	1000	50	20-30
Violent Flow/Crowd	2012	246	107	25
RWF-2000	2021	2000	150	30

la vida real donde la persona agredida corre en busca de ayuda. Sin embargo, poder comparar resultados con un gran número de trabajos es una razón convincente para su selección. Este dataset aborda condiciones de alta visibilidad de la escena violenta.

En segundo lugar, como tercer dataset más citado en el mismo periodo, el dataset *Violent Flow* (Hassner et al., 2012) contiene vídeos reales de YouTube concurridos con audio. Este dataset aborda condiciones de alta densidad poblacional en la escena violenta. Por último, como cuarto dataset más citado en el mismo periodo, el dataset *Real World Fight 2000* contiene escenas de violencia grabadas por cámaras de seguridad. Posiblemente sea el dataset más completo debido a su variado y extenso contenido de escenas violentas, que abarca diversas situaciones, tipos de violencia y dos mil vídeos. Este dataset aborda condiciones de baja visibilidad de la escena violenta.

Se decide no escoger el dataset *Action Movies*, que, si bien, es el segundo dataset más utilizado en dicho periodo, contiene escenas de películas de acción con primeros planos y buena iluminación, en última instancia, escenas irreales. La Tabla 3.1 contiene información sobre los datasets seleccionados. Los tres datasets contienen el mismo número de vídeos violentos que no violentos. Para todos los datasets, se utiliza el 70 % de los vídeos para entrenamiento, el 10 % para validación y el 20 % para el testeo.

El dataset *Violent Flow* es el único de los tres datasets seleccionados que contiene vídeos con diferente número de *frames*. Por lo tanto, aquellos vídeos que cuentan con menos *frames* que la mediana del número de *frames* (ciento siete) tienen el último *frame* insertado al final del vídeo tantas veces como sea necesario para alcanzar los ciento siete *frames*; mientras que si el vídeo es más largo que la duración mediana, se recortan los *frames* al final del vídeo. Esto se debe a que las arquitecturas propuestas requieren que la duración de los vídeos entrantes sea constante.

3.2.2. Entrenamiento de CNN

Esta Sección expone el entrenamiento de las CNN seleccionadas VGG-16 y VGG-19 preentrenadas en el dataset *ImageNet*, el cual consiste, como se expone en la Sección 3.1.1, en sustituir y entrenar la última capa densamente conectada de 1000 neuronas por una capa de una sola neurona última capa densamente conectada para que ésta aprenda a clasificar fotogramas como violentos o no violentos los fotogramas de los vídeos. Cabe recalcar, además, que todas las computaciones de esta Tesis Doctoral se realizan en un servidor con un procesador Intel(R) Core(TM) i9-10940X con 188 gigabytes de RAM y una GPU NVIDIA GeForce RTX 3090 con 24 gigabytes de memoria. Se aplica una tasa de aprendizaje de 10^{-3} y 50 épocas, lo cual se considera suficiente para la última capa densamente conectada, la cual actúa como clasificador binario. El optimizador utilizado es *Adam*, elegido por su eficiencia computacional y su capacidad para adaptarse dinámicamente al ritmo de aprendizaje de cada parámetro, y la función de pérdida *Entropía cruzada binaria*, dado que se trata de un problema de clasificación binaria en el que se busca predecir la probabilidad de que un *frame* pertenezca a una de dos clases mutuamente excluyentes.

Para realizar el **proceso de entrenamiento**, todos los *frames* de los vídeos de entrenamiento se redimensionan a la forma (224, 224, 3), que corresponde a la altura, anchura y número de canales de color (RGB) requeridos por la arquitectura VGG-16 y VGG-19. De este modo, cada vídeo tiene un tamaño de (N_{frames}, 224, 224, 3). Dado que VGG-16 y VGG-19 procesan imágenes individuales (y no vídeos), estas se procesan en un orden específico para que más tarde se puedan asociar, en secuencia, los *frames* de cada vídeo dentro de una única matriz con forma ($N_{train\ videos} \cdot N_{frames}$, 224, 224, 3). Por último, se crea una matriz de forma ($N_{train\ videos} \cdot N_{frames}$, 2) que contiene un valor 0 en caso de que los *frames* del vídeo al que corresponden sean no violentos, y 1 en caso de que sean violentos. Cabe destacar que es posible que un pequeño número de *frames*, al principio y/o al final del vídeo, sean *frames* que correspondan a antes o después de la acción violenta, añadiendo imágenes al proceso de entrenamiento que no están dentro de la acción violenta, pero indicando que sí lo están por pertenecer a un vídeo etiquetado como violento. Dado que la última capa a entrenar únicamente consta de una neurona, su entrenamiento no requiere de una cantidad grande de épocas, por ello se establece un máximo de 10 épocas durante el entrenamiento.

Tabla 3.2: Resultados del entrenamiento de VGG-16 con los datasets seleccionados.

Dataset	**Tasa de Aprendizaje**	**Épocas**	***Validation Accuracy***
Hockey Fights	0.001	10	0.99
RWF-2000			0.98
Violent Flow			0.99

Tabla 3.3: Resultados del entrenamiento de VGG-19 con los datasets seleccionados.

Dataset	**Tasa de Aprendizaje**	**Épocas**	***Validation Accuracy***
Hockey Fights	0.001	10	1.00
RWF-2000			0.99
Violent Flow			1.00

La Tabla 3.2 y la Tabla 3.3 resumen la información de entrenamiento de VGG-16 aplicando los tres datasets seleccionados e indicando en las columnas *training accuracy* y *validation accuracy* los valores obtenidos en la última época. Cada uno de los tres datasets seleccionados es dividido en un 80 % para entrenamiento (*training*), un 20 % para validación (*validation*) y un 20 % para testeo (*test*).

Con todo, se puede observar cómo tanto VGG-16 como VGG-19 tienen unos buenos resultados durante la etapa de entrenamiento, ajustando la última capa con una única neurona de clasificación a la tarea de detección de violencia.

3.3. Resultados del testeo de CNN con número mínimo de fotogramas violentos

Esta Sección contiene los resultados del testeo de las arquitecturas expuestas en las Secciones 3.1.1 y 3.1.2. La Sección 3.3.1 expone los resultados de testeo de las CNN. La Sección 3.3.2 analiza los resultados de la arquitectura que combina CNN junto con una característica manual.

3.3.1. Testeo de CNN

Esta Sección expone los resultados obtenidos durante el proceso testeo de las CNN candidatas a emplear en esta Tesis Doctoral, VGG-16 y VGG-19, con su última capa densa compuesta por una neurona entrenada en la Sección 3.2.2 con los tres datasets

seleccionados. El **proceso de testeo** consiste en realizar predicciones con las CNN seleccionadas, utilizando el 20 % de los vídeos que no se utilizaron durante el proceso de entrenamiento, donde se pueden dar cuatro posibles situaciones: *Verdadero Positivo (VP)*, *Verdadero Negativo (VN)* o *Falso Positivo (FP)*, *Falso Negativo (FN)*. Estas cuatro posibilidades son las clásicas de cualquier problema de clasificación binaria supervisada, donde las imágenes han sido previamente etiquetadas como violentas o no violentas.

- *Verdadero Positivo* indica que el modelo predice que la imagen violenta es violenta.
- *Verdadero Negativo* indica que el modelo predice que la imagen no violenta es no violenta.
- *Falso Positivo* indica que el modelo predice que la imagen no violenta es violenta.
- *Falso Negativo* indica que el modelo predice que la imagen violenta es no violenta.

El objetivo es maximizar tanto el *Verdadero Positivo* como el *Verdadero Negativo*, minimizando el *Falso Positivo* y el *Falso Negativo*. Un *Falso Negativo* es peor que un *Falso Positivo*, ya que si una agresión está ocurriendo y el modelo no la detecta como tal, la víctima no sería rescatada.

Comenzando con los resultados obtenidos por parte de **VGG-16**, la Tabla 3.4 contiene las matrices de confusión obtenidas del testeo de VGG-16 para los datasets: *Hockey Fights* (escenas de alta visibilidad), *RWF-2000* (escenas de baja visibilidad) y *Violent Flow* (escenas de alta densidad poblacional); respectivamente. Los resultados obtenidos muestran como la *accuracy* obtenida en los datasets *Hockey Fights* y *Violent Flow* alcanza un 92 %, 64 % y un 86 %, respectivamente. Los resultados más bajos en lo que respecta al dataset *RWF-2000* son debidos a la gran variedad de escenas que contiene, en general con una iluminación y claridad de imagen baja.

Prosiguiendo con los resultados obtenidos por parte de **VGG-19**, la Tabla 3.5 los resultados obtenidos muestran como la *accuracy* obtenida en los datasets *Hockey Fights*, *Violent Flow* y *RWF-2000* alcanza un 92 %, 64 % y un 86 %, respectivamente. Los resultados más bajos en lo que respecta al dataset *RWF-2000* son debidos a la gran variedad de escenas que contiene, en general con una iluminación y claridad de imagen baja.

Tabla 3.4: Métricas de los resultados de test utilizando VGG-16 preentrenada en los datasets utilizados.

Dataset	**Numero *test videos***	***Accuracy***	***Precision***	***Recall***	***F1 Score***	***Specificity***	***AUC***
Hockey Fights	200	0.92	0.97	0.91	0.94	0.95	0.98
RWF-2000	400	0.64	0.65	0.55	0.61	0.72	0.64
Violent Flow	50	0.86	0.81	0.92	0.87	0.79	0.90

Tabla 3.5: Métricas de los resultados de test utilizando VGG-19 preentrenada en los datasets utilizados.

Dataset	**Numero *test videos***	***Accuracy***	***Precision***	***Recall***	***F1 Score***	***Specificity***	***AUC***
Hockey Fights	200	0.94	0.93	0.95	0.94	0.93	0.92
RWF-2000	400	0.64	0.65	0.63	0.64	0.66	0.71
Violent Flow	50	0.92	0.87	0.99	0.93	0.86	0.97

Centrado en la métrica de *accuracy* como la más utilizada, *Hockey fights* obtiene un 94 %, *RWF-2000* un 64 % y *Violent Flow* un 92 %. Se puede observar como los valores obtenidos por parte de VGG-19 son superiores en los datatasets *Hockey Fights* y *Violent Flow* a los obtenidos por parte de VGG-16, si bien la dificultad a la hora de detectar violencia en el dataset *RWF-2000* se mantiene con la misma *accuracy*.

3.3.2. Testeo de CNN con número mínimo de fotogramas violentos

Esta Sección tiene como objetivo entrenar y testear el modelo de detección de violencia expuesto en la Figura 3.3, empleando como CNN las redes seleccionadas como candidatas VGG-16 y VGG-19. Además de ser un método que combina CNN y una característica manual, lo cual es una aportación de esta Tesis Doctoral, además de resultar de interés en comparación con arquitecturas que combinan CNN con el uso combinado de capas LSTM.

Para determinar si un vídeo debe considerarse violento o no, se define un umbral: un número mínimo de *frames* que deben ser detectados como violentos dentro del vídeo. Si el número de *frames* violentos supera ese umbral, el vídeo completo se clasifica como violento. En este trabajo, se evalúan diferentes valores para ese umbral: 10, 20, 40, 80 y 120 *frames*. La Tabla 3.6 muestra estos valores junto con el porcentaje que representan sobre la duración total de los vídeos en cada conjunto de datos. Dado que algunos vídeos tienen una duración limitada, no todos los valores de umbral se aplican en todos los casos.

Tabla 3.6: Número mínimo de *frames* para considerar un vídeo como violento y el porcentaje de *frames* que representa con respecto al número total de *frames* de los vídeos de cada *frame* de datos.

Número mínimo de *frames* para considerar un vídeo como violento	***Hockey Fights*** **(%)**	***Violent Flow*** **(%)**	***RWF-2000*** **(%)**
10	25.0	9.3	6.7
20	50.0	18.7	13.3
40	100.0	37.4	26.7
80	-	74.8	53.3
120	-	-	80.0

Tabla 3.7: Resultados de Test de VGG-16 y un mínimo número de *frames* violentos utilizando *Hockey fights*.

Mínimo número de *frames*	***Accuracy***	**VP**	**VN**	**FP**	**FN**	**VP %**	**VN %**	**FP %**	**FN %**
10	0.95	127	63	4	6	63.5	31.5	2.0	3.0
20	0.92	122	63	4	11	61.1	31.5	2.0	5.5
40	0.84	0.86	106	66	0	28	53.1	33.2	0.0

Por ejemplo, los vídeos del conjunto de datos *Hockey Fights* tienen solo 40 *frames*, por lo que no se prueban umbrales mayores a 40. En el caso de *Violent Flow*, cuya mediana de duración es de 107 *frames* por vídeo, no se considera el umbral de 120. Solo en el conjunto de datos *RWF-2000*, cuyos vídeos contienen 150 *frames*, se evalúan todos los valores propuestos.

Testeo de VGG-16 combinada con un mínimo número de fotogramas violentos

Se exponen a continuación los resultados del testeo de VGG-16 combinado con un mínimo número de fotogramas violentos en los mismos datasets seleccionados en los que han sido entrenados.

La Tabla 3.7 muestra los resultados obtenidos para valores mínimos de *frames* de 10, 20 y 40 en el dataset *Hockey fights*, incluyendo la *accuracy* y las matrices de confusión en valor absoluto correspondientes a cada configuración. Se resalta en color verde la mejor *accuracy* alcanzada, la cual se obtiene utilizando un número reducido de *frames*, concretamente 10 *frames*.

Tabla 3.8: Resultados de Test de VGG-16 y un mínimo número de *frames* violentos utilizando el dataset Violent Flow.

Mínimo Número de *frames*	***Accuracy***	**VP**	**VN**	**FP**	**FN**	**VP %**	**VN %**	**FP %**	**FN %**
10	0.78	23	16	11	0	46.0	32.0	22.0	0.0
20	0.80	23	17	10	0	46.0	34.0	20.0	0.0
40	0.86	23	20	7	0	46.0	40.0	14.0	0.0
80	0.86	21	22	5	2	42.0	44.0	10.0	4.0

Tabla 3.9: Resultados de Test de VGG16 y un mínimo número de *frames* violentos utilizando *RWF-2000*.

Mínimo Número de *frames*	***Accuracy***	**VP**	**VN**	**FP**	**FN**	**VP %**	**VN %**	**FP %**	**FN %**
10	0.71	169	96	56	31	48.0	27.3	15.9	8.8
20	0.68	157	97	55	43	44.6	27.6	15.6	12.2
40	0.64	140	103	49	60	39.8	29.3	13.9	17.0
80	0.58	108	110	42	92	30.7	31.3	11.9	26.1
120	0.48	34	136	16	166	9.7	38.6	4.5	47.2

La Tabla 3.8 presenta los resultados obtenidos para valores mínimos de 10, 20, 40 y 80 *frames* en el dataset *Violent Flow*, mostrando tanto la *accuracy* como las matrices de confusión en valor absoluto para cada configuración. Se destaca en verde que la mejor *accuracy* se alcanza utilizando un número elevado de *frames* —80 en este caso—, siendo el valor máximo considerado para este conjunto de datos; en contraste con los datasets de *Hockey Fights* y *RWF-2000*, donde el mejor desempeño se logró con el menor número de *frames*.

La Tabla 3.9 muestra los resultados correspondientes a los valores mínimos de 10, 20, 40, 80 y 120 *frames* en el dataset *RWF-2000*, incluyendo la *accuracy* y las matrices de confusión en valor absoluto para cada uno. Se resalta en verde que la mejor *accuracy* se alcanza utilizando el número mínimo de *frames* más bajo evaluado, 10, al igual que en el caso del dataset *Hockey fights*.

Finalmente, en la Tabla 3.10 se presentan, para cada uno de los tres conjuntos de datos seleccionados, las combinaciones de modelo y número mínimo de *frames* que logran la mayor *accuracy*. También se incluyen las métricas más habituales de clasificación binaria para evaluar el desempeño de dichas combinaciones.

Tabla 3.10: Métricas completas para el mejor modelo VGG-16 con número mínimo de *frames* violentos por dataset.

Dataset	**Mínimo número de *frames***	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Specificity**	**AUC**
Hockey Fights	10	0.95	0.97	0.95	0.96	0.9403	0.9476
RWF-2000	10	0.71	0.75	0.85	0.80	0.63	0.74
Violent Flow	80	0.86	0.80	0.91	0.86	0.81	0.86

Tabla 3.11: Resultados de Test de VGG19 y un mínimo número de *frames* violentos utilizando *Hockey fights*.

Mínimo número de *frames*	***Accuracy***	**VP**	**VN**	**FP**	**FN**	**VP %**	**VN %**	**FP %**	**FN %**
10	0.95	98	91	9	2	49.0	45.5	4.5	1.0
20	0.94	97	91	9	3	48.5	45.5	4.5	1.5
40	0.84	70	97	3	30	35.0	48.5	1.5	15.0

Se aprecia que el modelo alcanza una *Precision* del 0.97 % y un *F1 score* del 0.96 % en el dataset *Hockey Fights*, mientras que en *Violent Flow* obtiene una *precision* de 0.80 % y un *F1 score* de 0.86 %, reflejando un buen equilibrio entre *Precision* y *Recall*. Por otro lado, en el dataset *RWF-2000*, las métricas son considerablemente más bajas, con una *precision* de 0.71 % y una *specificity* de apenas 0.63 %, lo que indica que el modelo enfrenta mayores dificultades para distinguir entre clases en este conjunto respecto a los otros dos.

Testeo de VGG-19 combinada con un mínimo nímero de fotogramas violentos

A continuación se presentan los resultados de la prueba de VGG-19 utilizando un número reducido de fotogramas violentos, sobre los mismos datasets que se emplearon durante el entrenamiento.

La Tabla 3.11 contiene los resultados para el número mínimo de *frames* con valores 10, 20 y 40 en el dataset *Hockey fights*, mostrando la *accuracy* y las matrices de confusión en valor absoluto para cada caso. Se recalca en color verde la mejor *accuracy* obtenida, la cual ocurre con un número mínimo de *frames* bajo, como 10 *frames*.

La Tabla 3.12 contiene los resultados para el número mínimo de *frames* con valores de 10, 20, 40 y 80 *frames* en el dataset *Violent Flow*, mostrando la *accuracy* y las matrices de confusión en valor absoluto para cada caso. Se puede observar en verde que la mejor *accuracy* se obtiene para un número de *frames* elevado —80 en este caso— el máximo

Tabla 3.12: Resultados de Test de VGG-19 y un mínimo número de *frames* violentos utilizando el dataset Violent Flow.

Mínimo Número de *frames*	***Accuracy***	**VP**	**VN**	**FP**	**FN**	**VP %**	**VN %**	**FP %**	**FN %**
10	0.84	25	17	8	0	50.0	34.0	16.0	0.0
20	0.88	25	19	6	0	50.0	38.0	12.0	0.0
40	0.94	25	22	3	0	50.0	44.0	6.0	0.0
80	0.96	25	23	2	0	50.0	46.0	4.0	0.0

Tabla 3.13: Resultados de Test de VGG19 y un mínimo número de *frames* violentos utilizando *RWF-2000*.

Mínimo Número de *frames*	***Accuracy***	**VP**	**VN**	**FP**	**FN**	**VP %**	**VN %**	**FP %**	**FN %**
10	0.76	177	90	62	23	50.3	25.6	17.6	6.5
20	0.74	167	95	57	33	47.4	27.0	16.2	9.4
40	0.73	153	103	49	47	43.5	29.3	13.9	13.4
80	0.63	104	118	34	96	29.5	33.5	9.7	27.3
120	0.50	25	152	0	175	7.1	43.2	0.0	49.7

valor posible de los seleccionados para este dataset; al contrario de los vídeos de *Hockey Fights* y *RWF-2000*, cuyo valor óptimo ha sido el menor posible.

La Tabla 3.13 contiene los resultados para el número mínimo de *frames* con valores de 10, 20, 40, 80 y 120 en el dataset *RWF-2000*, mostrando la *accuracy* y las matrices de confusión en valor absoluto para cada caso. Se puede observar en verde que la mejor *accuracy* se obtiene con el valor más bajo de número mínimo de *frames* implementado, 10, en este caso, al igual que con el dataset *Hockey fights*.

Finalmente, en la Tabla 3.14 se muestran, para cada uno de los tres conjuntos de datos seleccionados, las combinaciones de modelo y número mínimo de *frames* que obtienen la mayor *accuracy*. Además, se incluyen las métricas más comunes en clasificación binaria para evaluar el rendimiento de dichas combinaciones.

Se puede observar que el modelo logra una *Precision* de 0.92 % y un *F1 score* de 0.95 % en el dataset *Hockey Fights*, mientras que en *Violent Flow* se alcanzan valores aún más altos con una precisión de 0.93 % y un *F1 score* de 0.96 %, lo que indica un buen equilibrio entre las métricas de *Precision* y *Recall*. Sin embargo, en el dataset *RWF-2000*, las métricas son notablemente más bajas, con una *precision* de 0.74 % y una *specificity*

TABLA 3.14: Métricas para el mejor modelo VGG19 con número mínimo de *frames* para cada uno de los datasets seleccionados.

Dataset	**Mínimo Número de *frames***	***Accuracy***	***Precision***	***Recall***	***F1 Score***	***Specificity***
Hockey Fights	10	0.95	0.92	0.98	0.95	0.91
RWF-2000	10	0.76	0.74	0.89	0.89	0.6
Violent Flow	80	0.96	0.93	1	0.96	0.92

de tan solo el 0.6 %, lo que sugiere que el modelo tiene más dificultad para diferenciar las clases en este conjunto de datos en comparación con los otros dos.

Con todo, la aplicación de un mínimo número de *frames* detectados como violentos por parte de VGG-19 para considerar el vídeo como violento o no violento obtiene muy buenos resultados en el dataset de *Hockey Fights* y *Violent Flow*. Si bien en los datasets *Hockey Fights* y *RWF-2000* parece haber una tendencia a una mejor detección con un número bajo de mínimo número de *frames* detectados como violentos, esto no se cumple en el caso del dataset *Violent Flow*, cuyo valor óptimo es de ochenta *frames*.

3.4. Análisis comparativo de las arquitecturas y resultados

Esta Sección tiene como objetivo analizar las arquitecturas propuestas y comparar los resultados obtenidos con otros trabajos del estado del arte. En la Sección 3.4.1 se estudian las arquitecturas desarrolladas, mientras que en la Sección 3.4.2 se presentan y comparan los resultados experimentales.

3.4.1. Análisis de las arquitecturas propuestas

Esta Sección aborda el análisis de las arquitecturas propuestas en este Capítulo: la primera, que emplea una CNN para detectar violencia en vídeo *frame* a *frame*, siendo las CNN candidatas en esta Tesis Doctoral VGG-16 y VGG-19. La segunda, que emplea una CNN junto con un umbral mínimo de *frames* violentos para clasificar un vídeo como violento, de nuevo siendo VGG-16 y VGG-19 las CNN candidatas en esta Tesis Doctoral.

Comenzando con la primera arquitectura, que emplea **CNN para detectar la violencia en vídeo *frame* a *frame***, los resultados obtenidos en este estudio sugieren que las CNN, por sí solas, son capaces de ofrecer una clasificación precisa y con un

menor coste computacional, tanto aportando una mayor ligereza a los modelos, ya que no emplean capas RNN para la extracción de patrones temporales, como en el tiempo de inferencia, donde los resultados obtenidos por parte de VGG-16 y VGG-19 muestran cómo el **tiempo medio de inferencia** se realiza en 1.56 milisegundos y 1.64 milisegundos, respectivamente. Esto implica que VGG-16 infiere un 5 % más rápido que VGG-19, si bien, esto, por supuesto, depende del servidor donde se ejecuta. Este hallazgo es especialmente relevante cuando se consideran los resultados obtenidos en los datasets de *Hockey Fights* y *Violent Flow*, donde las CNN muestran un rendimiento sobresaliente en comparación con enfoques que combinan CNN con capas LSTM y Bi-LSTM. Sin embargo, este enfoque no contiene un concepto de vídeo en su conjunto, lo que supone que en un dispositivo en tiempo real se daría una alerta de violencia en una escena al detectarla como violenta con respecto a un único fotograma.

Pasando ahora al modelo propuesto, que emplea la **CNN preentrenada combinada con un umbral mínimo de fotogramas** violentos para clasificar el vídeo como violento, se ha demostrado especialmente eficaz. Este enfoque se basa en un número mínimo de fotogramas violentos para clasificar un vídeo como violento, ya que, si solo un único fotograma es detectado como positivo, se incrementa el número de alertas de escenas violentas, lo que a su vez eleva la cantidad de falsos positivos. De esta manera, al considerar un conjunto de fotogramas como violentos, se mejora la precisión de la clasificación, reduciendo la probabilidad de generar alertas incorrectas. Por esta razón, considerar un conjunto de fotogramas como violentos aprovecha la ligereza de los modelos y la rapidez de inferencia, además de permitir una clasificación más precisa y fiable.

Otro aspecto importante a destacar es que, en un dispositivo en tiempo real, almacenar los *frames* a medida que llegan para su posterior inferencia consume una cantidad considerable de memoria. Sin embargo, con un modelo que combina CNN con un mínimo número de fotogramas detectados como violentos, las inferencias se realizan conforme se procesan los fotogramas, lo que no solo reduce la cantidad de memoria requerida, sino que también permite disponer de modelos más ligeros al no necesitar capas LSTM para analizar las características temporales.

Estos resultados refuerzan la idea de que la violencia en un vídeo puede ser detectada utilizando principalmente la información visual, sin depender necesariamente del análisis explícito de la relación temporal entre las imágenes. A pesar de que un acto violento

es, por naturaleza, un fenómeno que se desarrolla a lo largo del tiempo, los resultados obtenidos indican que las CNN, que extraen características espaciales, tienen un papel clave en la detección de violencia, incluso sin la necesidad de combinarse con LSTM.

En resumen, la arquitectura propuesta, basada en CNN con un umbral mínimo de *frames*, ha mostrado ser una solución efectiva y eficiente para la detección de violencia en vídeos, superando en algunos casos a enfoques más complejos que incluyen LSTM, como se expone en el Capítulo 4. Además, al eliminar la necesidad de almacenar los *frames* completos en memoria y al hacer el modelo más ligero, este enfoque es más escalable y adecuado para su implementación en dispositivos con recursos limitados.

3.4.2. Análisis y comparación de los resultados obtenidos

Esta Sección tiene como objetivo analizar los resultados obtenidos por los modelos expuestos en este Capítulo, empleando como CNN candidatas VGG-16 y VGG-19, y relacionarlos con los resultados de otros trabajos seleccionados. Se exponen en la Tabla 3.15 y en la Tabla los resultados de los modelos propuestos empleando como CNN VGG-16 y VGG-19, respectivamente. Por otro lado, se representan en las Tablas 4.16, 4.17 y 4.18 los resultados obtenidos mediante el uso de VGG-16 y VGG-19, así como otros resultados de modelos implementados en el estado del arte, empleando los datasets *Hockey Fights*, *Violent Flow* y *RWF-2000*. Se representan con color verde las celdas de los modelos propuestos en esta Tesis Doctoral, en rojo aquellas que utilizan, como CNN, bien VGG-16 o VGG-19, y en blanco aquellos trabajos seleccionados que no utilizan VGG-16 o VGG-19. En base a estos resultados, se exponen una serie de conclusiones.

Al analizar los resultados obtenidos con el dataset *Hockey fights*, los modelos propuestos superan la *accuracy* del trabajo de Mumtaz et al. (2022b), que alcanza un 91.29 % utilizando VGG-19 preentrenada con Bi-LSTM. Sin embargo, ninguno logra resultados superiores a los basados en VGG-16 preentrenada, que alcanzan hasta un 99.1 %, como en los trabajos de Aarthy y Nithya (2022) y Mugunga et al. (2021b). Esto indica que los modelos propuestos no superan la capacidad de predicción sigue siendo más eficaz en este contexto. Además, el uso de características manuales o análisis *frame* a *frame* no supera a los modelos de aprendizaje profundo, destacando la superioridad de arquitecturas que aprenden características de manera autónoma.

Tabla 3.15: *Accuracy* de los modelos propuestos, basados en VGG-16, en el Capítulo 3 de esta Tesis Doctoral.

Dataset	**VGG-16**	
	frame* a *frame	+ **Número mínimo de fotogramas**
Hockey Fights	92 %	95 %
RWF-2000	64 %	71 %
Violent Flow	86 %	86 %

Tabla 3.16: *Accuracy* de los modelos propuestos, basados en VGG-19, en el Capítulo 3 de esta Tesis Doctoral.

Dataset	**VGG-19 preentrenada**	
	frame* a *frame	+ **Característica Manual**
Hockey Fights	94 %	95 %
RWF-2000	64 %	76 %
Violent Flow	92 %	96 %

Respecto a los resultados obtenidos con el uso del dataset *Violent Flow*, el trabajo de Mumtaz et al. (2022b) alcanza una *accuracy* del 90.5 % utilizando VGG-19 preentrenada junto con Bi-LSTM, sin embargo, el modelo propuesto en esta Tesis Doctoral que combina VGG-19 con características manuales logra un mejor rendimiento, alcanzando un 96.0 % de *accuracy*. Además, este resultado supera al obtenido por otros trabajos basados en VGG-16 preentrenada, como el de Gupta y Ali (2022), que logra entre un 92.2 % y 95.1 % con la combinación de VGG-16 y LSTM/Bi-LSTM, y el de Asad et al. (2021), que obtiene un 97.1 %. Este resultado demuestra la eficacia de la combinación de VGG-19 con características manuales en la mejora del rendimiento en la detección de violencia.

Finalmente, en el caso del dataset *RWF-2000*, ninguno de los modelos propuestos en este Capítulo, alcanza resultados cercanos a los obtenidos por los tres trabajos que también utilizan este conjunto de datos. En particular, el modelo de Mugunga et al. (2021b), que combina VGG-16 preentrenada con Bi-Conv-LSTM, logra un 92.4 % de *accuracy*, mientras que el modelo de Mumtaz et al. (2022b), basado en VGG-19 preentrenada con capas CNN adicionales y Bi-LSTM, obtiene un 90.5 %. Los modelos propuestos en este Capítulo alcanzan *accuracies* más bajas, variando desde un 76.0 % con VGG-19 y características manuales, hasta un 64.0 % con el enfoque *frame* a *frame*. Esto demuestra que *RWF-2000* es el dataset con el que los modelos propuestos tienen mayores dificultades para detectar violencia en vídeo, probablemente debido a la complejidad y variedad de las escenas presentes en este conjunto de datos.

TABLA 3.17: *Accuracy* de los modelos propuestos en el Capítulo 3 de esta Tesis Doctoral y de otros trabajos seleccionados que utilizan el dataset *Hockey Fights*.

Referencia	**CNN**	**Combinación**	***Hockey Fights***
Vijeikis et al. (2022)	MobileNet V2	LSTM	99.5 %
Halder y Chatterjee (2020)	CNN propia	Bi-LSTM	99.3 %
Mugunga et al. (2021b)	VGG-16 PT (ImageNet)	Bi-Conv-LSTM	99.1 %
Aarthy y Nithya (2022)	VGG-16 PT (ImageNet)	LSTM	99.1 %
Jahlan y Elrefaei (2021)	MNAS CNN PT	Conv-LSTM	99.0 %
Traoré y Akhloufi (2020b)	Dos EfficienNet-B0 PT en ImageNet	Bi-LSTM	99 %
Gupta y Ali (2022)	VGG-16 PT en ImageNet	LSTM/Bi-LSTM	97.6 %/98.8 %
Asad et al. (2021)	Dos VGG-16 PT en ImageNet + Bloques Residuales Densos y Anchos (WDRB)	LSTM	98.8 %
Ullah et al. (2022)	Darknet + Flujo óptico residual	M-LSTM	98.0 %
Traoré y Akhloufi (2020a)	VGG-16 PT en INRA	Bi-GRU	98.0 %
Islam et al. (2021a)	CNN propia	LSTM	97.0 %
Sharma et al. (2021)	Xception PT en ImageNet	LSTM	96.6 %
Modelo propuesto	VGG19 PT en ImageNet	Característica manual	95.0 %
Modelo propuesto	VGG16 PT en ImageNet	Característica manual	95.0 %
Talha et al. (2022)	CNN	LSTM	94.9 %
Modelo propuesto	VGG-19 PT en ImageNet (frame by frame)	N	94 %
Modelo propuesto	VGG-16 PT en ImageNet (frame by frame)	N	92 %
Mumtaz et al. (2022b)	VGG-19 PT en ImageNet + capas CNN extra	Bi-LSTM	91.29 %

3.5. Discusión

El objetivo de este Capítulo ha sido el estudio de la relevancia de las características extraídas exclusivamente por CNN frente a arquitecturas que combinan CNN con LSTM. Por ello, se desarrollan dos modelos:

- El primer modelo implica únicamente un análisis de *frame* a *frame* de las detecciones realizadas mediante el uso de una CNN, si bien se pierden las relaciones temporales entre los *frames*, proporcionando información sobre la eficacia del análisis de acciones a través de imágenes. Se han escogido como CNN a estudiar VGG-16 y VGG-19 preentrenadas.
- El segundo modelo incorpora una CNN con implementación de lógica manual, de modo que un vídeo no se considera violento a menos que un cierto número de

Tabla 3.18: *Accuracy* de los modelos propuestos en en el Capítulo 3 de esta Tesis Doctoral y de otros trabajos seleccionados que utilizan el dataset *Violent Flow*.

Referencia	**CNN**	**LSTM**	***Violent Flow***
Gupta y Ali (2022)	VGG-16 PT en ImageNet	LSTM/Bi-LSTM	92.2 %/95.1 %
Jahlan y Elrefaei (2021)	MNAS CNN PT	Conv-LSTM	100.0 %
Halder y Chatterjee (2020)	CNN propia	Bi-LSTM	98.6 %
Mugunga et al. (2021b)	VGG-16 PT en ImageNet	Bi-Conv-LSTM	98.4 %
Ullah et al. (2022)	Darknet + Flujo Óptico Residual	M-LSTM	98.2 %
Asad et al. (2021)	Dos VGG-16 PT en ImageNet + Bloques Residuales Densos y Anchos (WDRB)	LSTM	97.1 %
Modelo propuesto	VGG-19 PT en ImageNet	Característica Manual	96.0 %
Traoré y Akhloufi (2020a)	VGG-16 PT en INRA	Bi-GRU	95.5 %
Traoré y Akhloufi (2020b)	Dos EfficienNet-B0 PT en ImageNet	Bi-LSTM	93.7 %
Modelo propuesto	VGG-19 PT en ImageNet (frame a frame)	N	92.0 %
Mumtaz et al. (2022b)	VGG-19 PT en ImageNet + capas CNN extra	Bi-LSTM	90.5 %
Modelo propuesto	VGG-16 PT en ImageNet	Característica Manual	86.0 %
Modelo propuesto	VGG-16 PT en ImageNet (frame a frame)	N	86.0 %
Talha et al. (2022)	CNN	LSTM	77.3 %

Tabla 3.19: *Accuracy* de los modelos propuestos en en el Capítulo 3 de esta Tesis Doctoral y de otros trabajos del estado del arte utilizando el dataset *RWF-2000*.

Referencia	**CNN**	**LSTM**	***RWF-2000***
Mugunga et al. (2021b)	VGG-16 PT en ImageNet	Bi-Conv-LSTM	92.4 %
Mumtaz et al. (2022b)	PT VGG-19 (ImageNet) + extra CNN layers	Bi-LSTM	90.5 %
Vijeikis et al. (2022)	MobileNet V2	LSTM	82.0 %
Modelo propuesto	VGG-19 PT en ImageNet	Característica Manual	76.0 %
Modelo propuesto	VGG-16 PT en ImageNet	Característica Manual	71.0 %
Modelo propuesto	VGG-19 PT en ImageNet (frame a frame)	N	64.0 %
Modelo propuesto	VGG-16 PT en ImageNet (frame a frame)	N	64.0 %

frames sean detectados como violentos por parte de la CNN. De nuevo, se han escodigo como CNN a estudiar VGG-16 y VGG-19 preentrenada.

A partir de los modelos diseñados e implementados, se ha llegado a la conclusión de que

- La clasificación basada exclusivamente en redes convolucionales (CNN) resulta suficiente para lograr una detección efectiva de violencia en vídeo. La incorporación de un umbral mínimo de *frames* violentos permite tomar decisiones en tiempo real de forma escalable, sin necesidad de almacenar múltiples *frames* para una inferencia posterior. En su lugar, solo se guardan los resultados de las inferencias individuales. Además, al prescindir del uso de capas LSTM, se logra un modelo más ligero y rápido, lo cual favorece su implementación en entornos con recursos limitados. Los resultados de las arquitecturas que emplean VGG-19 preentrenada y VGG-19 preentrenada combinado con características manuales alcanzan respectivamente una *accuracy* del 94 % y 95 % en *Hockey Fights*, un 64 % y un 76 % en *RWF-2000*, y un 92 % y un 96 % en *Violent Flow*, los cuales son superiores a los obtenidos por VGG-16.

- Como CNN a escoger, se decide seleccionar VGG-19 frente a VGG-16 debido a que se han obtenido resultados entre un 2 % y un 10 % de mejora en la *accuracy* de testeo obtenida en los resultados de la experimentación expuesta en este Capítulo. Por otro lado, el uso de VGG-19 frente a VGG-16 supone una disminución de un 5 % en la velocidad de inferencia en el servidor empleado, pasando de un tiempo medio de inferencia de 1.64 milisegundos con VGG-16, a 1.56 milisegundos con VGG-19. Con todo, dados los mejores resultados obtenidos de VGG-19 frente a VGG-19, se escoge VGG-19 como red CNN a utilizar en los Capítulos 4 y 5 de esta Tesis Doctoral.

- Si bien los modelos desarrollados en este Capítulo de esta Tesis Doctoral no han superado algunos de los resultados obtenidos por parte de otros trabajos seleccionados, hay algunos que sí son destacables. Por ejemplo, en el caso del dataset *Violent Flow*, el modelo desarrollado en esta Tesis Doctoral que combina VGG-19 junto con un mínimo número de *frames* para detectar un vídeo como violento, obtiene un 96 % de *accuracy*, que es inferior al 98.4 % registrado por VGG-16 con Bi-Conv-LSTM (Mugunga et al., 2021b).

En base a las conclusiones extraídas en este Capítulo, se observa que las arquitecturas diseñadas e implementadas funcionan correctamente, logrando valores de exactitud relevantes. En el Capítulo 4 se diseñan e implementan arquitecturas que integran el modelo VGG-19 preentrenada con capas LSTM y Bi-LSTM. Posteriormente, se comparan los resultados obtenidos en este Capítulo con los del Capítulo 4.

Capítulo 4

VGG-19 combinada con capas RNN

Capítulo 4

VGG-19 combinada con capas RNN

Índice

La violencia física representa un desafío importante en nuestra sociedad, afectando a personas en todo el mundo (Long et al., 2021). Tal y como se establece en el Capítulo 2, se exploran múltiples enfoques para abordar la detección de agresiones físicas en las sociedades de todo el mundo, siendo la detección en tiempo real de la violencia la última barrera para la protección de las víctimas frente a la agresión. Al final del Capítulo 2 se exponen las carencias del estado del

En cuanto a los algoritmos utilizados para identificar violencia en vídeo mediante inteligencia artificial, la combinación más frecuente en los últimos dos años ha sido la integración de Redes Neuronales Convolucionales y Redes de Memoria a Corto y Largo Plazo (Negre et al., 2024b), obteniendo resultados destacados.

En base a los resultados obtenidos en el Capítulo 3 de esta Tesis Dcotoral, y considerando que las imágenes de vídeo poseen una relación temporal entre fotogramas, se decide emplear redes LSTM y Bi-LSTM después de la etapa previa de la arquitectura CNN, ya que estas redes son especialmente adecuadas para capturar patrones temporales. Además, la experimentación expuesta en el Capítulo 3 muestra que la mejor CNN para la tarea fue VGG-19, comparada con VGG-16, lo que refuerza la elección de esta arquitectura. En base a todo ello, se exponen a continuación los **objetivos para este**

Capítulo 4, los cuales están alineados con los objetivos de esta Tesis Doctoral expuestos en el Capítulo 1:

- Desarrollar e implementar modelos de detección de violencia basados en VGG-19 preentrenada y analizar sus resultados al combinarlos con características manuales, en comparación con arquitecturas que integran capas RNN. Este enfoque busca incorporar la información sobre la relación temporal entre los fotogramas del vídeo, utilizando capas RNN, en particular LSTM y Bi-LSTM, para capturar patrones secuenciales, en consonancia con los objetivos **OB3** y **OB5** establecidos en la Sección 2.5 y alineado con los objetivos del Capítulo 3.

- Analizar la mejora que supone el uso de capas Bi-LSTM sobre capas LSTM en las arquitecturas desarrolladas cuando se utilizan en conjunto con VGG-19. *Alineado con los objetivos **OB3** y **OB5** de esta Tesis Doctoral.*

- Entrenar la arquitectura que combina VGG-19 con capas RNN, en particular LSTM y Bi-LSTM, con una amplia gama de hiperparámetros como: tasa de aprendizaje, el número de neuronas de las capas RNN, así como el número de neuronas de las capas densamente conectadas. Con ello se espera poder analizar y observar si un ciertos valores o combinación de valores de hiperparámetros obtienen mejores resultados. *En sintonía con los objetivos **OB3** y **OB5** de la presente Tesis Doctoral.*

- Demostrar la idoneidad de la arquitectura propuesta basada en VGG-19 preentrenada en el dataset de *ImageNet* combinada con capas RNN. En el Capítulo 3 VGG-19 obtiene mejores resultados que VGG-16 en la detección de las escenas de violencia tanto fotograma a fotograma como implementando un número mínimo de fotogramas detectados como violentos para considerar el vídeo como violento. En este Capítulo se pretende comparar si la combinación de VGG-19 con capas RNN supera la combinación de VGG-16 combinado con capas RNN de los trabajos seleccionados del estado del arte. *Conforme a los objetivos **OB3** y **OB5** planteados en esta Tesis Doctoral.*

Con todo, este Capítulo se divide en las siguientes subsecciones. La Sección 4.1 describe la arquitectura propuesta para la combinación de la red VGG-19 preentrenada con capas RNN, concretamente LSTM/Bi-LSTM. La Sección 4.2 expone los resultados obtenidos

durante la fase de entrenamiento de la arquitectura que combina el uso de VGG-19 preentrenada y capas RNN, siendo estas LSTM/Bi-LSTM. La Sección 4.3 presenta los resultados obtenidos durante el proceso de testeo de la arquitectura que combina VGG-19 preentrenada combinada con capas, en concreto LSTM/Bi-LSTM. La Sección 4.4 trata los resultados obtenidos por las arquitecturas propuestas en este Capítulo. La Sección 5.5 expone las conclusiones del Capítulo.

4.1. Arquitectura VGG-19 combinada con capas RNN

Esta Sección describe la arquitectura propuesta para la combinación de VGG-19 preentrenada y capas RNN. La arquitectura consiste en la combinación de la red preentrenada VGG-19 en *ImageNet* combinada con capas RNN, donde VGG-19 extrae características espaciales que son introducidas en las capas RNN para la extracción de características temporales, siendo candidatas específicamente las capas LSTM/Bi-LSTM para esta Tesis Doctoral.

Las redes LSTM (Long Short-Term Memory) son un tipo especializado de redes neuronales recurrentes (RNN) diseñadas para modelar secuencias y dependencias a largo plazo, resolviendo el problema del desvanecimiento del gradiente característico de las RNN tradicionales Sharma et al. (2021). A nivel funcional, las LSTM están compuestas por celdas de memoria que regulan el flujo de información mediante tres compuertas principales: la compuerta de olvido, la compuerta de entrada y la compuerta de salida. Estas compuertas utilizan funciones de activación sigmoide y tanh para controlar qué información se descarta, qué nueva información se almacena y qué parte del estado interno se propaga hacia la salida Vidhya y Uthra (2021). En una arquitectura Bi-LSTM (Bidirectional LSTM), se introducen dos secuencias de LSTM que procesan la información en direcciones opuestas (hacia adelante y hacia atrás), lo cual permite capturar contexto tanto pasado como futuro en las secuencias Halder y Chatterjee (2020).

La **primera arquitectura** diseñada que combina el uso de VGG-19 preentrenada junto con capas RNN, en particular LSTM/Bi-LSTM, emplea una **capa *Time Distributed*** para encapsular la red VGG-19, como se puede observar en la Figura 4.1. Esta arquitectura tiene la ventaja de que el entrenamiento del modelo se puede realizar como un único algoritmo, ya que el uso de las capas *Time Distributed* permite que hasta que se termine la extracción de las características de todos los fotogramas del vídeo por

parte de la red VGG-19, el proceso no continúe hacia el uso de Bi-LSTM y las capas densamente conectadas.

Sin embargo, esta arquitectura implica un menor control sobre el entrenamiento particular de la CNN, así como el entrenamiento de las capas LSTM y las capas densamente conectadas que actúan como clasificador, al entrenarse todo como un único bloque. También supone un inconveniente al implementar algoritmos de inteligencia artificial explicables con esta arquitectura. Para aplicar GradCAM (como un algoritmo de IA Explicable ampliamente utilizado) se debe crear un submodelo a partir del modelo entrenado para obtener dos salidas: el resultado de la clasificación y el resultado de la capa convolucional a analizar. Sin embargo, estas capas hacen que sea realmente complejo acceder a las capas de un modelo que, a su vez, está dentro de otro modelo, como es el caso de VGG-19 en la arquitectura diseñada.

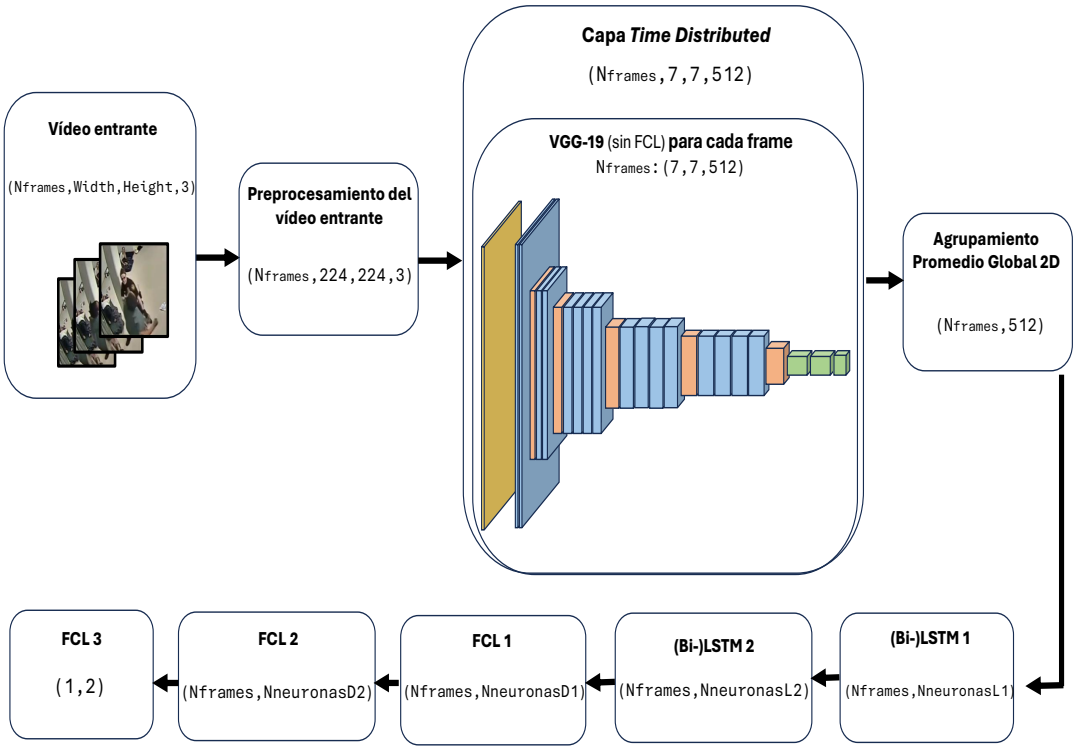

FIGURA 4.1: Arquitectura de detección de violencia combinando VGG-19 preentrenada junto con capas LSTM/Bi-LSTM utilizando una capa *Time Distributed*.

Por todo ello, se propone una **segunda arquitectura**, que es la implementada en esta Tesis Doctoral, que divide el proceso en dos partes distintas **basada en la concatenación de características espaciales** entre VGG-19 y las capas RNN. La arquitectura se expone en la Figura 4.2. Esta forma de realizar la detección de violencia

en combinación usando CNN y RNN consiste en extraer las características espaciales de los *frames* de un vídeo por un lado y, una vez que se obtienen todas las características espaciales, se concatenan en un solo elemento (un array) para ser introducidas en las capas LSTM que extraen características temporales con las que alimentar las capas densas, que actúan como clasificadores. A continuación, se describen las dos partes principales del algoritmo representado en la Figura 4.2:

- **Primera parte del algoritmo:** Esta etapa se encarga del preprocesamiento y extracción de características espaciales:
  - Se recibe un vídeo de entrada, que se divide en N_{frames}.
  - Cada *frame* se redimensiona al tamaño $(224, 224, 3)$ para ser compatible con la arquitectura VGG-19.
  - Cada uno de estos *frames* es introducido de manera individual en la red VGG-19 preentrenada (sin las capas completamente conectadas).
  - La salida de VGG-19 para cada *frame* es un mapa de características de tamaño $(7, 7, 512)$.
  - Estas características se agrupan a lo largo del eje temporal, resultando en una matriz de dimensiones $(N_{frames}, 7, 7, 512)$.
  - Posteriormente, se aplica una capa de *Global Average Pooling 2D*, la cual reduce las dimensiones espaciales, produciendo una representación compacta de cada *frame* en la forma $(N_{frames}, 512)$.
- **Segunda parte del algoritmo:** Esta etapa se encarga de la extracción de características temporales y la clasificación:
  - La secuencia de vectores $(N_{frames}, 512)$ se introduce en una o dos capas LSTM o Bi-LSTM, que permiten capturar las dependencias temporales entre los *frames*.
  - Las salidas de las capas LSTM/Bi-LSTM tienen dimensiones $(N_{frames}, N_{neuronsL1})$ y $(N_{frames}, N_{neuronsL2})$, dependiendo del número de neuronas utilizadas en cada capa.
  - A continuación, las salidas temporales se procesan mediante tres capas completamente conectadas (FCL), cuya dimensión final es $(1, 2)$ para la clasificación binaria (violento/no violento).

- Las capas densas utilizan funciones de activación *Sigmoid*, adecuadas para tareas de clasificación binaria.

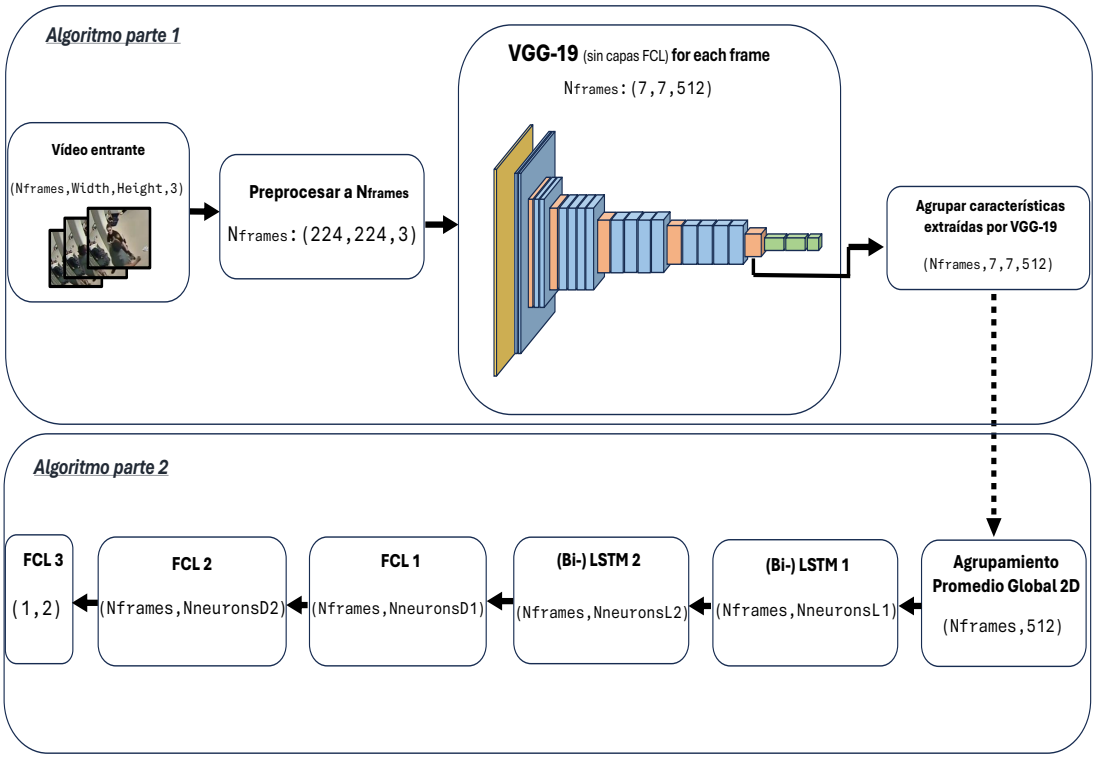

FIGURA 4.2: Arquitectura de detección de violencia combinando VGG-19 preentrenada junto con capas LSTM/Bi-LSTM utilizando concatenación de características espaciales.

4.2. Entrenamiento de VGG-19 con capas RNN

Esta Sección expone los resultados obtenidos durante la fase de entrenamiento de la arquitectura que combina el uso de VGG-19 preentrenada y capas LSTM/Bi-LSTM, con la finalidad de estudiar cómo estas capas RNN ayudan a extraer patrones temporales entre los fotogramas que conforman los vídeos; todo ello aplicado a distintos tipos de escenarios según el dataset seleccionado. A partir lo de realizado en la Sección 3.2.2 del Capítulo 3, la red VGG-19 preentrenada en el dataset de *ImageNet* ha sido modificada sustituyéndola por una capa densa con una neurona y entrenando esta última nueva capa en los datasets de violencia seleccionados y expuestos en la Sección 3.2.1 del Capítulo 3. La Sección 4.2.1 trata el entrenamiento de VGG-19 preentrenada con capas LSTM/Bi-LSTM. La Sección 4.2.2 compara el rendimiento de arquitecturas con capas LSTM y Bi-LSTM en diferentes contextos de entrenamiento, agrupados por tipo de

escena (alta visibilidad, alta densidad poblacional y baja visibilidad), y concluye con un análisis en entornos cambiantes mediante métricas de exactitud y precisión.

Cada dataset, descrito con mayor detalle en el Capítulo 2, ha sido seleccionado para el entrenamiento y testeo de las arquitecturas propuestas, abordando un tipo específico de escena con condiciones particulares:

- El dataset *Hockey Fights* (dataset 1) incorpora escenas con condiciones de alta visibilidad.
- El dataset *Violent Flow* (dataset 2) contiene escenas en condiciones de alta densidad poblacional.
- El dataset *RWF-2000* (dataset 3) incluye escenas con condiciones de alta baja visibilidad.

4.2.1. Entrenamiento de VGG-19 con capas LSTM/Bi-LSTM en distintos escenarios

Esta Sección presenta el entrenamiento de la arquitectura que combina el uso de la red VGG-19 preentrenada en el dataset *ImageNet* junto con capas LSTM/Bi-LSTM, mostrada en la Figura 4.2 de la Sección 4.1.

Durante el entrenamiento, se lleva a cabo un proceso de ajuste de hiperparámetros mediante una técnica de sintonización automática, partiendo del conjunto de valores más utilizados en los trabajos seleccionados en detección de violencia en vídeo, presentados en el Capítulo 2, Tabla 2.14. Para ello, se utiliza *Keras Tuner*, una herramienta que facilita la búsqueda de la combinación óptima de hiperparámetros mediante técnicas avanzadas de optimización. En particular, se exploran diferentes combinaciones de tasa de aprendizaje, número de neuronas en las capas LSTM y número de neuronas en las capas completamente conectadas, con el objetivo de identificar la configuración que proporciona el mejor rendimiento del modelo.

Los valores típicos utilizados en los trabajos seleccionados compilados en la Tabla 2.14 para la tasa de aprendizaje son 10^{-2}, 10^{-3}, 10^{-4}, los cuales controlan la magnitud del ajuste de los pesos en cada iteración del entrenamiento; para el número de neuronas de

las capas LSTM/Bi-LSTM y las capas completamente conectadas se emplean valores como 64, 128, 256, 512 y 1024, que determinan la capacidad de representación del modelo. Se permite en la búsqueda de combinaciones que la primera y segunda capa de LSTM(Bi-LSTM y las capas completamente conectadas adopten números diferentes de neuronas, ya que se desea estudiar cómo afecta el aumentar, disminuir o mantener el mismo número de neuronas afecta a los resultados obtenidos. Al igual que en el entrenamiento de la red VGG-19, se utiliza el optimizador *Adam*, que garantiza un ajuste eficiente y estable de los parámetros, mientras que la función de pérdida *entropía cruzada binaria* es apropiada para la clasificación binaria, lo que maximiza la precisión del modelo (Negre et al., 2024c). Teniendo en cuenta un número tan elevado de hiperparámetros posibles, el número de combinaciones es muy elevado. Se implementa *Keras-Tuner* para probar un total de 200 combinaciones posibles, lo que representa aproximadamente el 10 % del total de 1875.

Los resultados se muestran agrupados en Tablas según el tipo de escena que abordan para las arquitecturas que emplean tanto capas LSTM como capas Bi-LSTM. Dado que durante el entrenamiento el objetivo es maximizar la métrica de *validation accuracy*, se exponen los tres mejores modelos obtenidos por *Keras-Tuner*, para cada tipo de escena y tipo de arquitectura, en su búsqueda de la mejor combinación de hiperparámetros, con el objetivo de no almacenar únicamente el que mejor se ajuste a los datos de validación.

I) Escenas en condiciones de con condiciones de alta visibilidad

En primer lugar, la Tabla 4.1 expone los resultados de las arquitecturas entrenadas en condiciones de alta densidad poblacional.

Tabla 4.1: Comparación de los tres mejores modelos con capas LSTM y Bi-LSTM en condiciones de alta visibilidad (dataset *Hockey Fights)*.

Tipo	**Modelo**	***Validation Accuracy***	***Learning Rate***	**Capa 1**	**Capa 2**	**FCL 1**	**FCL 2**
LSTM	1.1	0.96	0.0001	256	128	64	64
	1.2	0.96	0.01	512	512	128	64
	1.3	0.96	0.001	64	1024	512	64
Bi-LSTM	1.1	0.97	0.01	64	64	64	512
	1.2	0.97	0.01	512	64	1024	512
	1.3	0.97	0.01	64	256	64	256

II) Escenas en condiciones de alta densidad poblacional

Se muestran en segundo lugar en la Tabla 4.2 los resultados de las arquitecturas entrenadas en condiciones de alta densidad poblacional.

Tabla 4.2: Comparación de los tres mejores modelos con capas LSTM y Bi-LSTM para el dataset *Violent Flow*.

Tipo	**Modelo**	***Validation Accuracy***	***Learning Rate***	**Capa 1**	**Capa 2**	**FCL 1**	**FCL 2**
LSTM	2.1	0.96	0.01	64	512	64	128
	2.2	0.96	0.01	128	64	128	64
	2.3	0.96	0.0001	512	512	128	1024
Bi-LSTM	2.1	0.96	0.001	1024	512	64	256
	2.2	0.96	0.01	256	256	256	128
	2.3	0.96	0.01	128	512	128	256

III) Escenas en condiciones de baja visibilidad

Finalmente, la Tabla 4.11 muestra en tercer lugar los resultados de las arquitecturas entrenadas en condiciones de baja visibilidad.

Tabla 4.3: Comparación de los tres mejores modelos con capas LSTM y Bi-LSTM para el dataset *RWF-2000*.

Tipo	**Modelo**	***Validation Accuracy***	***Learning Rate***	**Capa 1**	**Capa 2**	**FCL 1**	**FCL 2**
LSTM	3.1	0.87	0.001	1024	1024	256	512
	3.2	0.87	0.0001	512	64	128	128
	3.3	0.87	0.001	128	512	128	256
Bi-LSTM	3.1	0.87	0.01	512	64	256	64
	3.2	0.87	0.001	512	256	64	256
	3.3	0.87	0.001	128	128	64	128

Como análisis preliminar del entrenamiento de las arquitecturas propuestas, basadas en capas LSTM y Bi-LSTM, bajo diferentes condiciones de visibilidad y densidad poblacional revela diferencias notables en el rendimiento. En escenarios con alta visibilidad, como muestra la Tabla 4.1, los modelos Bi-LSTM alcanzan una precisión de validación ligeramente superior (0.97) en comparación con los modelos LSTM (0.96). En condiciones de alta densidad poblacional (Tabla 4.2), ambos tipos de arquitectura

lograron un rendimiento comparable, con una precisión de validación consistente de 0.96 en todos los modelos seleccionados. Finalmente, en situaciones de baja visibilidad (Tabla 4.11), se observa una disminución general en el rendimiento, donde tanto las arquitecturas LSTM como Bi-LSTM alcanzaron una precisión máxima de 0.87, lo que sugiere que la visibilidad limitada afecta negativamente la capacidad de los modelos para detectar comportamientos violentos. En conjunto, los resultados indican que las arquitecturas Bi-LSTM tienden a ofrecer un mejor desempeño en condiciones óptimas, mientras que en entornos más desafiantes, la precisión se estabiliza independientemente del tipo de arquitectura empleada.

4.2.2. Rendimiento en entrenamiento de VGG-19 con capas LSTM/Bi-LSTM en distintos escenarios

Esta Sección tiene el objetivo de evaluar cuál de las arquitecturas propuestas -basadas en capas LSTM o Bi-LSTM— ofrece un mejor rendimiento global en distintos escenarios, se ha llevado a cabo un análisis comparativo utilizando los mejores modelos de validación en cada uno de los tres entornos definidos: alta visibilidad (Hockey Fights), alta densidad poblacional (Violent Flow) y baja visibilidad (RWF-2000).

Para realizar una evaluación conjunta de los modelos, se han calculado dos métricas agregadas: la media ponderada de *accuracy* y la desviación típica de los valores de *accuracy* en los distintos datasets. A continuación, se describen ambas métricas.

La **media ponderada de *accuracy***, expuesta en la Ecuación 4.1, permite tener en cuenta el tamaño de cada conjunto de validación al calcular el rendimiento global del modelo. Se define como:

$$\text{Accuracy}_{\text{ponderada}} = \frac{n_1 \cdot \text{Acc}_1 + n_2 \cdot \text{Acc}_2 + n_3 \cdot \text{Acc}_3}{n_1 + n_2 + n_3} \tag{4.1}$$

donde:

- Acc_i es la *accuracy* de validación obtenida por el mejor modelo en el entorno i,
- n_i es el número de muestras (videos) del conjunto de validación en dicho entorno.

Esta métrica permite evaluar la *accuracy* global del modelo a través de escenarios heterogéneos.

La desviación típica, expuesta en la Ecuación 4.2, permite medir la **precisión o estabilidad** del modelo: una desviación baja implica que los resultados son más consistentes entre entornos. La fórmula utilizada es:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{Acc}_i - \mu)^2} \tag{4.2}$$

donde:

- μ es la media de los valores de *accuracy*,
- N es el número total de entornos considerados (en este caso, 3),
- Acc_i es la *accuracy* en el entorno i.

Con todo, se presentan en la Tabla 4.4 los valores obtenidos para cada arquitectura, utilizando los mejores modelos identificados en cada entorno:

Tabla 4.4: Comparación global de rendimiento entre modelos LSTM y Bi-LSTM en entornos cambiantes.

Modelo	**Accuracy ponderada**	**Desviación típica**
LSTM	0.90	0.05
Bi-LSTM	0.91	0.05

Como puede observarse, el modelo Bi-LSTM obtiene una mayor **exactitud global**, mientras que el modelo LSTM muestra una ligeramente mayor **precisión** en sus resultados. Esta diferencia, aunque pequeña, puede resultar relevante dependiendo del contexto de aplicación, por ejemplo, si se prioriza exactitud frente a consistencia.

4.3. Testeo de VGG-19 con capas RNN

Esta Sección presenta los resultados obtenidos durante el proceso de testeo de la arquitectura que combina VGG-19 preentrenada combinada con capas LSTM y con

capas Bi-LSTM para distintos escenarios (alta visibilidad, alta densidad y baja visibilidad). La Sección 4.3.1 expone los resultados del testeo de VGG-19 preentrenada con capas LSTM y capas Bi-LSTM, agrupado según los distintos escenarios. La Sección 4.3.2 compara el rendimiento de arquitecturas con capas LSTM y Bi-LSTM en diferentes contextos de testeo, agrupados por tipo de escena (alta visibilidad, alta densidad y baja visibilidad), y concluye con un análisis en entornos cambiantes mediante métricas de exactitud y precisión.

4.3.1. Testeo de VGG-19 con capas LSTM/Bi-LSTM en distintos escenarios

Esta Sección contiene los resultados del testeo de VGG-19 preentrenada con capas LSTM y Bi-LSTM, agrupados según los distintos escenarios. Para cada escenario se presentan tres tablas: la primera resume las configuraciones y resultados de exactitud obtenidos en el conjunto de prueba; la segunda desglosa los valores absolutos y porcentuales de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos; y la tercera reporta métricas de evaluación complementarias como precisión, *recall*, *F1 score*, especificidad y AUC para los modelos seleccionados.

I) Escenas en condiciones de con condiciones de alta visibilidad

En primer lugar, las Tablas 4.5, 4.6 y 4.7, exponen los resultados de las arquitecturas entrenadas en condiciones de alta densidad poblacional.

Tabla 4.5: Resultados del testeo para el dataset *Hockey Fights* con capas LSTM y Bi-LSTM.

Tipo	**Modelo**	***Test Accuracy***	***Learning Rate***	**LSTM 1**	**LSTM 2**	**FCL 1**	**FCL 2**
LSTM	1.1	0.94	0.0001	256	128	64	64
	1.2	0.96	0.01	512	512	128	64
	1.3	0.96	0.001	64	1024	512	64
Bi-LSTM	1.1	0.93	0.01	64	64	64	512
	1.2	0.95	0.01	512	64	1024	512
	1.3	0.97	0.01	64	256	64	256

Tabla 4.6: Comparación de resultados entre capas LSTM y Bi-LSTM para el dataset *Hockey Fights*.

Modelo	***Tipo de Capa***	***Nº Test Videos***	**VP**	**VN**	**FP**	**FN**	**VP (%)**	**VN (%)**	**FP (%)**	**FN (%)**
1.2	LSTM	200	96	96	4	4	48.00	48.00	2.00	2.00
1.3	Bi-LSTM	200	20	23	2	5	40.00	46.00	4.00	10.00

Tabla 4.7: Métricas de testeo para los modelos con capas LSTM y Bi-LSTM en el dataset *Hockey Fights*.

Tipo	**Modelo**	**N.º de *test* videos**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Specificity**	**AUC**
LSTM	1.2	200	0.96	0.96	0.96	0.96	0.96	0.97
Bi-LSTM	1.3	200	0.97	0.95	0.98	0.97	0.95	0.98

II) Escenas en condiciones de alta densidad poblacional

Se muestran en segundo lugar en las Tablas 4.8, 4.9 y 4.10 los resultados de las arquitecturas testeadas en condiciones de alta densidad poblacional.

Tabla 4.8: Resultados del testeo para el dataset *Violent Flow* con capas LSTM y Bi-LSTM.

Tipo	**Modelo**	***Test Accuracy***	***Learning Rate***	**LSTM 1**	**LSTM 2**	**FCL 1**	**FCL 2**
LSTM	2.1	0.86	0.01	64	512	64	128
	2.2	0.86	0.01	128	64	128	64
	2.3	0.86	0.0001	512	512	128	1024
Bi-LSTM	2.1	0.90	0.01	64	512	64	128
	2.2	0.82	0.01	128	64	128	64
	2.3	0.88	0.0001	512	512	128	1024

Tabla 4.9: Comparación de resultados entre capas LSTM y Bi-LSTM para el dataset *Violent Flow*.

Modelo	***Nº Test Videos***	***Nº Test Videos***	**VP**	**VN**	**FP**	**FN**	**VP (%)**	**VN (%)**	**FP (%)**	**FN (%)**
1	LSTM	50	20	23	2	5	40.00	46.00	4.00	10.00
1	Bi-LSTM	50	24	21	4	1	48.00	42.00	8.00	2.00

Tabla 4.10: Métricas de testeo para los modelos con capas LSTM y Bi-LSTM en el dataset *Violent Flow*.

Tipo	**Modelo**	**N.º de *test* videos**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Specificity**	**AUC**
LSTM	2.1	50	0.86	0.85	0.88	0.86	0.84	0.93
Bi-LSTM	2.1	50	0.90	0.86	0.97	0.91	0.84	0.96

III) Escenas en condiciones de baja visibilidad

Finalmente, las Tablas 4.11, 4.12 y 4.13, muestran en tercer lugar los resultados de las arquitecturas testeadas en condiciones de baja visibilidad.

Tabla 4.11: Resultados del testeo para el dataset *RWF-2000* con capas LSTM y Bi-LSTM.

Tipo	**Modelo**	***Validation Accuracy***	***Learning Rate***	**Capa 1**	**Capa 2**	**FCL 1**	**FCL 2**
LSTM	3.1	0.72	0.001	1024	1024	256	512
	3.2	0.71	0.0001	512	64	128	128
	3.3	0.71	0.001	128	512	128	256
Bi-LSTM	3.1	0.73	0.0001	1024	1024	256	512
	3.2	0.73	0.0001	512	64	128	128
	3.3	0.68	0.001	1024	1024	256	512

Tabla 4.12: Comparación de resultados entre capas LSTM y Bi-LSTM para el dataset *RWF-2000*.

Modelo	***Tipo de Capa***	***Nº Test Videos***	**VP**	**VN**	**FP**	**FN**	**VP (%)**	**VN (%)**	**FP (%)**	**FN (%)**
1	LSTM	400	154	135	65	46	38.50	33.75	16.25	11.50
1	Bi-LSTM	400	143	149	51	57	35.75	37.25	12.75	14.25

Tabla 4.13: Métricas de testeo para los modelos con capas LSTM y Bi-LSTM en el dataset *RWF-2000*.

Tipo	**Modelo**	**N.º de *test* videos**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Specificity**	**AUC**
LSTM	3.1	400	0.72	0.70	0.77	0.74	0.68	0.79
Bi-LSTM	3.4	400	0.73	0.74	0.72	0.73	0.75	0.79

En resumen, los experimentos realizados con la arquitectura VGG-19 combinada con capas LSTM y Bi-LSTM en distintos escenarios revelan que el desempeño de los modelos varía significativamente según las condiciones del entorno. Mientras que en escenarios con alta visibilidad ambos tipos de capas ofrecen resultados sobresalientes, destacándose ligeramente Bi-LSTM, en contextos de alta densidad poblacional o baja visibilidad se observaron caídas notables en la precisión general. Sin embargo, en estos casos, las capas Bi-LSTM tienden a ofrecer un mejor equilibrio entre sensibilidad y especificidad, lo que sugiere su mayor capacidad para capturar relaciones temporales complejas en situaciones más desafiantes. Estos hallazgos permiten establecer lineamientos sobre qué configuraciones resultan más adecuadas según el contexto de aplicación.

4.3.2. Rendimiento en testeo de VGG-19 con capas LSTM/Bi-LSTM en distintos escenarios

Esta Sección expone la evaluación del rendimiento global de los modelos LSTM y Bi-LSTM en la fase de testeo. Se han calculado la media ponderada de *accuracy* y la desviación típica a partir de los mejores resultados obtenidos para cada dataset. Las ecuaciones empleadas son las mismas que las de la Sección 4.2.2.

Tabla 4.14: Comparación global de testeo entre modelos LSTM y Bi-LSTM.

Modelo	**Accuracy ponderada**	**Desviación típica**
LSTM	0.80	0.06
Bi-LSTM	0.82	0.07

Los resultados presentados permiten concluir que, en términos globales, los modelos con capas Bi-LSTM superan ligeramente a los modelos con LSTM en la fase de testeo, tanto en precisión ponderada como en capacidad de generalización, reflejada en una desviación típica comparable. Aunque la diferencia de rendimiento no es amplia, el uso de Bi-LSTM se muestra como una alternativa robusta y más consistente ante la variabilidad entre datasets, consolidando su utilidad en tareas de clasificación de escenas violentas en contextos diversos.

4.4. Análisis y comparación de resultados

En esta Sección se presentan los resultados obtenidos por las arquitecturas propuestas en este Capítulo, además de ser comparadas con otros trabajos relacionados de la literatura reciente. La Sección 4.4.1 analiza si una cierta combinación de los hiperparámetros implementados obtiene sistemáticamente mejores resultados. La Sección 4.4.2 expone los resultados obtenidos por la arquitectura propuesta en distintos escenarios, comparando entre el uso de capas LSTM y Bi-LSTM y frente a otros trabajos con arquitecturas alternativas.

4.4.1. Hiperparámetros en arquitecturas basadas en capas LSTM y Bi-LSTM

En la Sección 4.3 se realiza el testeo de los modelos propuestos basados en la combinación de la red VGG-19 preentrenada en el dataset de *ImageNet* con capas LSTM y capas Bi-LSTM, en diferentes tipos de escena (alta visibilidad, alta densidad poblacional y baja visibilidad). Durante el proceso de entrenamiento, se emplea en la Sección 4.2 la herramienta *Keras-Tuner* para explorar 200 combinaciones distintas de hiperparámetros en cada arquitectura. Estos hiperparámetros incluyen el número de neuronas en las capas LSTM y Bi-LSTM, el número de neuronas en las capas densamente conectadas, y la tasa de aprendizaje. Los valores posibles para las neuronas fueron 64, 128, 256, 512 y 1024, mientras que las tasas de aprendizaje consideradas fueron 10^{-2}, 10^{-3} y 10^{-4}. Estas combinaciones representan aproximadamente un 10 % del total de 1875 combinaciones posibles, debido a las limitaciones computacionales para realizar una búsqueda exhaustiva.

Con el objetivo de identificar tendencias generales que puedan influir en el rendimiento del modelo, se agruparon los resultados obtenidos en función de tres criterios clave:

1. **Variación en el número de neuronas en las capas LSTM/Bi-LSTM**. Para cada combinación evaluada, se clasifica si el número de neuronas en las capas LSTM/Bi-LSTM aumenta, disminuye o se mantiene constante entre la primera y la segunda capa. Se calcula la mediana de la *validation accuracy* obtenida para

cada uno de estos tres grupos. El análisis busca evaluar si una variación en la capacidad de las capas recurrentes tiene impacto en el rendimiento del modelo:

- Modelos en los que el número de neuronas aumentó entre la primera y segunda capa.
- Modelos en los que el número de neuronas se mantuvo constante.
- Modelos en los que el número de neuronas disminuyó.

2. **Variación en el número de neuronas en las capas densamente conectadas**. De forma análoga al punto anterior, se agrupan los modelos en función de cómo varía el número de neuronas en las capas densamente conectadas (*Fully Connected Layers*). Este análisis permite estudiar si la capacidad de representación final (antes de la salida) contribuye de manera clara a una mejor predicción:

   - Modelos con aumento de neuronas entre capas densas.
   - Modelos con número constante de neuronas.
   - Modelos con disminución de neuronas entre capas.

3. **Tasa de aprendizaje utilizada**. En este caso, los modelos se agrupan directamente según el valor específico de la tasa de aprendizaje utilizada durante el entrenamiento. El objetivo es analizar si alguno de estos valores se asocia consistentemente con mejores resultados de validación:

   - Tasa de aprendizaje 10^{-2}
   - Tasa de aprendizaje 10^{-3}
   - Tasa de aprendizaje 10^{-4}

A pesar de los esfuerzos por organizar y analizar los resultados de forma estructurada, no se identifican patrones estadísticamente significativos que permitan concluir que alguna de estas configuraciones —ni el tipo de variación en las neuronas ni un valor concreto de tasa de aprendizaje— conduzca de manera consistente a un mejor rendimiento.

Por otro lado, al limitar el análisis a patrones de aumento, reducción o estabilidad en el número de neuronas, no se ha explorado el impacto de combinaciones exactas de valores. Esta elección se debe al número limitado de combinaciones evaluadas, que no permite obtener conclusiones sólidas sobre configuraciones específicas.

Sin embargo, esta falta de conclusiones claras sobre el efecto de los hiperparámetros puede entenderse como una manifestación de una de las principales limitaciones del aprendizaje profundo: su falta de explicabilidad. Aunque se identifican modelos con *accuracy* elevada, no es posible determinar qué configuraciones lo han provocado con certeza ni por qué. Esta problemática justifica la necesidad de un enfoque más explicable, que será desarrollado en el Capítulo 5, como alternativa para interpretar y confiar en los resultados obtenidos por estos sistemas.

4.4.2. Arquitecturas propuestas frente a otros trabajos seleccionados

Esta Sección tiene como objetivo analizar los resultados obtenidos por los modelos desarrollados en esta Tesis Doctoral entre sí y con otros trabajos seleccionados que implementen arquitecturas similares. Se incluyen en este análisis los resultados obtenidos en el Capítulo 3, con el fin de poder contrastar la mejora que supone el uso de capas LSTM o capas Bi-LSTM frente al uso exclusivo de VGG-19 preentrenada.

A continuación, se presentan los resultados obtenidos por los modelos propuestos en esta Tesis Doctoral (Tabla 4.15) y se comparan, en función del escenario, con modelos del estado del arte que emplean combinaciones de redes neuronales convolucionales (CNN) y redes recurrentes (RNN), principalmente LSTM. Esta comparación se organiza según los conjuntos de datos utilizados:

- En la Tabla 4.16 se muestran los resultados correspondientes al conjunto de datos *Hockey Fights*.
- En la Tabla 4.17, los resultados obtenidos con el conjunto *Violent Flow*.
- En la Tabla 4.18, los resultados asociados al conjunto *RWF-2000*.

En dichas tablas, las celdas en color verde indican los modelos propuestos en esta investigación. Las celdas en color rojo representan modelos que emplean como CNN base a VGG-16 o VGG-19. Finalmente, las celdas en blanco corresponden a trabajos seleccionados que no utilizan dichas arquitecturas. A partir de esta comparación, se presentan una serie de conclusiones relevantes que se detallan a continuación.

Al realizar la comparación de los resultados de predicción de VGG-19 en base a *frames* debe hacerse con cuidado, ya que se realiza imagen a imagen y no con respecto a un vídeo completo. Aun así, en el caso de los datasets de *Hockey fights* (escenario con alta visibilidad) y *Violent Flow* (escenario con alta densidad poblacional), se observa cómo los resultados son notables en comparación con aquellos combinados con capas LSTM y Bi-LSTM. Con respecto al modelo que combina VGG-19 junto a una lógica manual, si bien el número mínimo de *frames* detectados como violentos para considerar el vídeo en su conjunto como violento se establece a posteriori del entrenamiento, obtiene mejores resultados que analizando los *frames* individualmente. Este método obtiene mejores resultados en los datasets de *RWF-2000* (escenario con baja visibilidad) y *Violent Flow* que las combinaciones con capas LSTM y Bi-LSTM, incluso superándolas en un 6 % de *accuracy* en el caso del dataset *Violent Flow*.

Los hallazgos expuestos en el párrafo anterior plantean la posibilidad de que la violencia pueda detectarse utilizando únicamente información visual, sin necesidad de analizar explícitamente la relación temporal entre las imágenes. Aunque un acto violento es, en esencia, una acción que se desarrolla en una ventana de tiempo —diferenciándose así de la detección estática de objetos—, los resultados sugieren que las CNN tienen un papel notablemente superior al de las LSTM en la combinación de ambas para la detección de violencia. Esto lleva a cuestionarse si la mayor efectividad de las CNN se debe a que las características espaciales que extraen se emplean luego como *inputs* para las LSTM. Una línea de investigación futura derivada de esta Tesis Doctoral será explorar si la detección de violencia en vídeo podría mejorar al realizar en paralelo la extracción de características espaciales, mediante una CNN, y la extracción de características temporales, mediante una LSTM, alimentando ambas a un clasificador común. Esta estrategia podría contribuir a incrementar la precisión del sistema de detección.

Por otro lado, se confirma que las estructuras que utilizan Bi-LSTM obtienen mejores resultados que las que utilizan LSTM. En el caso del dataset de *Hockey fights* y *RWF-2000*, es solo del 1 % de *accuracy*, mientras que en el caso del dataset *Violent Flow* es del 4 %. Esta diferencia sugiere que la efectividad de las Bi-LSTM puede depender del tipo de escenario, siendo la mejora más notable en escenas con alta densidad poblacional (dataset *Violent Flow*), donde la estructura con capas Bi-LSTM capturaría mejor las secuencias temporales relevantes.

Al analizar los resultados obtenidos con el conjunto de datos *Hockey Fights*, se observa que todos los modelos propuestos en esta Tesis Doctoral superan la *accuracy* del modelo presentado por Mumtaz et al. (2022b), el cual alcanza un 91.29 % utilizando una arquitectura basada en VGG-19 preentrenada combinada con una Bi-LSTM. En comparación, los modelos propuestos que también emplean VGG-19 preentrenada alcanzan precisiones que oscilan entre el 94 % y el 97 %. Sin embargo, aunque algunos modelos propuestos superan estudios previos, ninguno logra resultados superiores a los basados en VGG-16 preentrenada, que alcanzan hasta un 99.1 %, como en los trabajos de Aarthy y Nithya (2022) y Mugunga et al. (2021b). Esto indica que, pese a la mayor profundidad de VGG-19, VGG-16 sigue siendo más eficaz en este contexto.

Escenario de alta visibilidad: Dataset *Hockey Fights*. Los mejores desempeños en este escenario se observan al combinar CNN preentrenadas con métodos secuenciales como LSTM o Bi-LSTM, con modelos como los de Vijeikis et al. (2022) y Halder y Chatterjee (2020) logrando un 99.5 % y 99.27 % de *accuracy*, respectivamente. Además, el uso de características manuales o análisis *frame* a *frame* no supera a los modelos de aprendizaje profundo, destacando la superioridad de arquitecturas que aprenden características de manera autónoma.

Escenario de alta densidad poblacional: Dataset *Violent Flow*). En cuanto a los resultados obtenidos con el dataset *Violent Flow*, el trabajo de Mumtaz et al. (2022b) alcanza una *accuracy* del 90.5 % utilizando VGG-19 preentrenada junto con Bi-LSTM, el mismo resultado que uno de los modelos desarrollados en esta Tesis Doctoral, que también combina VGG-19 y capas Bi-LSTM. No obstante, el modelo propuesto en esta Tesis Doctoral que combina VGG-19 con características manuales logra un mejor rendimiento, alcanzando un 96.0 % de *accuracy*. Este resultado supera al obtenido por otros trabajos basados en VGG-16 preentrenada, como el de Gupta y Ali (2022), que logra entre un 92.2 % y 95.1 % con la combinación de VGG-16 y LSTM/Bi-LSTM, y el de Asad et al. (2021), que obtiene un 97.1 %. Este resultado demuestra la eficacia de la combinación de VGG-19 con características manuales en la mejora del rendimiento en la detección de violencia.

Escenario de baja visibilidad: Dataset *RWF-2000*. En el caso del dataset *RWF-2000*, ninguno de los modelos propuestos en esta Tesis Doctoral alcanza resultados cercanos a los obtenidos por los tres trabajos que también utilizan este conjunto de datos.

En particular, el modelo de Mugunga et al. (2021b), que combina VGG-16 preentrenada con Bi-Conv-LSTM, logra un 92.4 % de *accuracy*, mientras que el modelo de Mumtaz et al. (2022b), basado en VGG-19 preentrenada con capas CNN adicionales y Bi-LSTM, obtiene un 90.5 %. Los modelos propuestos en esta tesis alcanzan *accuracies* más bajas, variando desde un 76.0 % con VGG-19 y características manuales, hasta un 64.0 % con el enfoque *frame* a *frame*. Esto demuestra que *RWF-2000* es el dataset con el que los modelos propuestos tienen mayores dificultades para detectar violencia en vídeo, probablemente debido a la complejidad y variedad de las escenas presentes en este conjunto de datos.

Resumen comparativo por escenarios:

- En el dataset *Hockey Fights*, los modelos más efectivos combinan CNN preentrenadas con métodos secuenciales, alcanzando hasta un 99.5 % de *accuracy*.
- En el dataset *Violent Flow*, la combinación de VGG-19 con características manuales supera a modelos previos, alcanzando un 96.0 % de *accuracy*.
- En el dataset *RWF-2000*, los modelos propuestos no logran superar los resultados de los trabajos previos, obteniendo *accuracies* más bajas, especialmente en escenarios de baja visibilidad.

Por último, es importante destacar que las arquitecturas presentadas en esta Tesis Doctoral también superan los resultados obtenidos por otros algoritmos del estado del arte, expuestos en el Capítulo 2 y recopilados en la Tabla 4.15. Por ejemplo, en el dataset *Hockey Fights*, las arquitecturas propuestas que integran capas LSTM y Bi-LSTM muestran un desempeño superior al modelo basado en Transformers presentado por Kumar et al. (2023). Asimismo, los resultados obtenidos superan a los de modelos como los propuestos por Wintarti et al. (2022), Lohithashva y Aradhya (2021) y Jaiswal y Mohod (2021), los cuales combinan técnicas manuales con enfoques de *Machine Learning*. Estos logros resaltan la efectividad de las arquitecturas propuestas en esta investigación en comparación con enfoques tradicionales o híbridos.

En base a los resultados obtenidos para los tres datasets seleccionados, se puede observar cómo al combinar con capas LSTM o capas Bi-LSTM, una red más profunda como VGG-19 genera características más complejas o redundantes, lo que dificulta la

TABLA 4.15: *Accuracy* de los modelos propuestos en esta Tesis Doctoral.

Dataset	**VGG-19 preentrenada**			
	frame* a *frame	**+ Característica Manual**	**+ LSTM**	**+ Bi-LSTM**
Hockey Fights	94 %	95 %	96 %	97 %
RWF-2000	64 %	76 %	72 %	73 %
Violent Flow	92 %	96 %	96 %	90 %

TABLA 4.16: *Accuracy* de los modelos propuestos en esta Tesis Doctoral y de otros trabajos seleccionados que utilizan el dataset *Hockey Fights*.

Referencia	**CNN**	**Combinación**	***Hockey Fights***
Vijeikis et al. (2022)	MobileNet V2	LSTM	99.5 %
Halder y Chatterjee (2020)	CNN propia	Bi-LSTM	99.3 %
Mugunga et al. (2021b)	VGG-16 PT (ImageNet)	Bi-Conv-LSTM	99.1 %
Aarthy y Nithya (2022)	VGG-16 PT (ImageNet)	LSTM	99.1 %
Jahlan y Elrefaei (2021)	MNAS CNN PT	Conv-LSTM	99.0 %
Traoré y Akhloufi (2020b)	Dos EfficienNet-B0 PT en ImageNet	Bi-LSTM	99 %
Gupta y Ali (2022)	VGG-16 PT en ImageNet	LSTM/Bi-LSTM	97.6 %/98.8 %
Asad et al. (2021)	Dos VGG-16 PT en ImageNet + Bloques Residuales Densos y Anchos (WDRB)	LSTM	98.8 %
Ullah et al. (2022)	Darknet + Flujo óptico residual	M-LSTM	98.0 %
Traoré y Akhloufi (2020a)	VGG-16 PT en INRA	Bi-GRU	98.0 %
Modelo propuesto	VGG-19 PT en ImageNet	Bi-LSTM	97.0 %
Islam et al. (2021a)	CNN propia	LSTM	97.0 %
Sharma et al. (2021)	Xception PT en ImageNet	LSTM	96.6 %
Modelo propuesto	VGG-19 PT en ImageNet	LSTM	96.0 %
Modelo propuesto	VGG19 PT en ImageNet	Característica manual	95.0 %
Talha et al. (2022)	CNN	LSTM	94.9 %
Modelo propuesto	VGG-19 PT en ImageNet (frame by frame)	N	94 %
Mumtaz et al. (2022b)	VGG-19 PT en ImageNet + capas CNN extra	Bi-LSTM	91.29 %

extracción de patrones temporales relevantes por parte del LSTM. En cambio, VGG-16, al ser más simple, produce representaciones más compactas y manejables, facilitando así el modelado secuencial, lo que demuestra los mejores resultados obtenidos con respecto a las arquitecturas propuestas basadas en VGG-19.

4.5. Discusión

Las agresiones físicas representan una preocupación notable en nuestra sociedad, afectando a individuos en todo el mundo e influyendo directamente en las víctimas

TABLA 4.17: *Accuracy* de los modelos propuestos en esta Tesis Doctoral y de otros trabajos seleccionados que utilizan el dataset *Violent Flow*.

Referencia	**CNN**	**LSTM**	***Violent Flow***
Gupta y Ali (2022)	VGG-16 PT en ImageNet	LSTM/Bi-LSTM	92.2 %/95.1 %
Jahlan y Elrefaei (2021)	MNAS CNN PT	Conv-LSTM	100.0 %
Halder y Chatterjee (2020)	CNN propia	Bi-LSTM	98.6 %
Mugunga et al. (2021b)	VGG-16 PT en ImageNet	Bi-Conv-LSTM	98.4 %
Ullah et al. (2022)	Darknet + Flujo Óptico Residual	M-LSTM	98.2 %
Asad et al. (2021)	Dos VGG-16 PT en ImageNet + Bloques Residuales Densos y Anchos (WDRB)	LSTM	97.1 %
Modelo propuesto	VGG-19 PT en ImageNet	Manual Feature	96.0 %
Traoré y Akhloufi (2020a)	VGG-16 PT en INRA	Bi-GRU	95.5 %
Traoré y Akhloufi (2020b)	Dos EfficienNet-B0 PT en ImageNet	Bi-LSTM	93.7 %
Modelo propuesto	VGG-19 PT en ImageNet (frame a frame)	N	92.0 %
Mumtaz et al. (2022b)	VGG-19 PT en ImageNet + capas CNN extra	Bi-LSTM	90.5 %
Modelo propuesto	VGG-19 PT en ImageNet	Bi-LSTM	90.0 %
Modelo propuesto	VGG-19 PT en ImageNet	LSTM	86.0 %
Talha et al. (2022)	CNN	LSTM	77.3 %

TABLA 4.18: *Accuracy* de los modelos propuestos en esta Tesis Doctoral y de otros trabajos del estado del arte utilizando el dataset *RWF-2000*.

Referencia	**CNN**	**LSTM**	***RWF-2000***
Mugunga et al. (2021b)	VGG-16 PT en ImageNet	Bi-Conv-LSTM	92.4 %
Mumtaz et al. (2022b)	PT VGG-19 (ImageNet) + extra CNN layers	Bi-LSTM	90.5 %
Vijeikis et al. (2022)	MobileNet V2	LSTM	82.0 %
Modelo propuesto	VGG-19 PT en ImageNet	Manual Feature	76.0 %
Modelo propuesto	VGG-19 PT en ImageNet	Bi-LSTM	73.0 %
Modelo propuesto	VGG-19 PT en ImageNet	LSTM	72.0 %
Modelo propuesto	VGG-19 PT en ImageNet (frame a frame)	N	64.0 %

y su bienestar psicológico (Long et al., 2021; Muarifah et al., 2022). La identificación de la violencia en vídeos en tiempo real sirve como última barrera en la protección de las víctimas, donde la inteligencia artificial ha demostrado excelentes resultados en esta tarea (Negre et al., 2024b). El objetivo principal de este capítulo ha sido mejorar la detección automática de violencia en vídeo mediante el desarrollo de modelos basados en inteligencia artificial que integraran información espacial y temporal. Para ello, se propusieron arquitecturas que combinaron la red convolucional preentrenada VGG-19 con capas LSTM y Bi-LSTM, con el fin de capturar tanto las características visuales como la dinámica temporal de las secuencias de vídeo.

Este enfoque se planteó como una extensión directa de los modelos desarrollados en el Capítulo 3, en el que se emplearon redes convolucionales junto con características manuales, pero sin considerar explícitamente la dimensión temporal. En este contexto, el trabajo realizado en este capítulo buscó incorporar el análisis secuencial entre fotogramas mediante redes recurrentes, con el propósito de superar las limitaciones observadas anteriormente y mejorar el rendimiento general de los modelos.

En particular, se evaluó el impacto del uso de capas LSTM y Bi-LSTM sobre el rendimiento en tres escenarios representados por distintos conjuntos de datos: escenas de alta visibilidad (*Hockey Fights*), alta densidad poblacional (*Violent Flow*) y baja visibilidad (*RWF-2000*). Los resultados mostraron que la inclusión de la información temporal permitió alcanzar mejoras significativas en algunos de estos contextos, especialmente cuando se combinaron redes CNN profundas con mecanismos secuenciales.

Este capítulo respondió directamente al segundo objetivo específico de la tesis: explorar arquitecturas híbridas que integraran redes convolucionales con redes neuronales recurrentes para capturar tanto características espaciales como temporales en vídeos violentos. Con este propósito, se implementaron dos modelos basados en la red VGG-19 preentrenada, utilizada para extraer características espaciales *frame* a *frame* a partir de vídeos de distintos datasets de violencia. Estas características se combinaron posteriormente con capas LSTM o Bi-LSTM, responsables de modelar la dinámica temporal presente en las secuencias de vídeo. Los hallazgos obtenidos permitieron identificar qué combinaciones de modelos resultaban más eficaces según el tipo de escenario visual, aportando evidencia clara sobre la relevancia del análisis temporal en la mejora del rendimiento de los sistemas de detección de violencia.

A partir de los modelos diseñados e implementados, se analizan los resultados obtenidos, lo que permite extraer una serie de conclusiones. A continuación, se presentan dichas **conclusiones, que recogen los principales hallazgos y aportes relevantes de esta Tesis Doctoral**.

- Los resultados obtenidos permiten analizar el impacto del uso de capas LSTM y Bi-LSTM en distintos contextos visuales. A continuación, se resumen los hallazgos más relevantes organizados por escenario:

- **Escenario de alta visibilidad: *Hockey Fights*.** En este escenario, la inclusión de capas LSTM y Bi-LSTM aporta mejoras claras respecto al uso exclusivo de VGG-19 preentrenada, que alcanza un 94 % de *accuracy*, y de VGG-19 combinada con características manuales (95 %). Al incorporar capas LSTM y Bi-LSTM, la *accuracy* asciende a 96 % y 97 %, respectivamente. Estos resultados sugieren que, en contextos con buena visibilidad y escenas bien delimitadas, el modelado temporal mediante redes recurrentes contribuye positivamente a la detección de violencia.
- **Escenario de alta densidad poblacional : *Violent Flow*.** En este caso, el modelo basado en VGG-19 con características manuales alcanza el mejor rendimiento (96 % de *accuracy*), seguido por el modelo con LSTM (96 %) y luego por Bi-LSTM (90 %). Esto indica que, aunque la inclusión de LSTM aporta un rendimiento comparable, la complejidad del entorno puede afectar negativamente al rendimiento de Bi-LSTM. Las características manuales, en cambio, ofrecen una ventaja competitiva en escenarios densos donde las relaciones espaciales adicionales pueden ser más relevantes que la secuencia temporal.
- **Escenario de baja visibilidad: *RWF-2000*.** Este escenario resulta el más desafiante. La arquitectura con VGG-19 preentrenada obtiene una *accuracy* del 64 %, mientras que su combinación con características manuales mejora hasta un 76 %. Al incorporar LSTM y Bi-LSTM, los resultados alcanzan un 72 % y 73 %, respectivamente. Aunque no se supera el rendimiento del modelo con características manuales, sí se observan mejoras respecto a la VGG-19 básica, lo que sugiere que el modelado temporal ayuda parcialmente, pero no compensa la pérdida de visibilidad en las escenas.
- **Comparativa LSTM vs. Bi-LSTM.** La comparación entre LSTM y Bi-LSTM muestra que la variante bidireccional tiende a ofrecer mejores resultados en entornos con alta visibilidad (*Hockey Fights*), pero no siempre aporta ventajas en escenarios complejos como *Violent Flow* o *RWF-2000*. En algunos casos, incluso se observa un leve descenso en el rendimiento. Esto sugiere que la selección del tipo de red recurrente debe adaptarse a las características del entorno visual y al tipo de secuencia de vídeo analizada.

En conjunto, estos resultados demuestran que la incorporación de capas LSTM y Bi-LSTM puede mejorar la detección de violencia en vídeos, especialmente en escenarios donde la visibilidad favorece el análisis secuencial. Sin embargo, su efectividad varía según el contexto, por lo que su uso debe evaluarse en función del tipo de escenario y de las características del conjunto de datos empleado.

- En cuanto a la ventaja específica de Bi-LSTM sobre LSTM, los resultados particulares muestran que, en el escenario de clasificación de acciones violentas del dataset *Hockey Fights*, la precisión mejora del 96 % al 97 % al usar Bi-LSTM. En el escenario de multitudes violentas del dataset *Violent Flow*, la mejora es más notable, pasando del 86 % al 90 %. Finalmente, en el escenario de videovigilancia del dataset *RWF-2000*, el rendimiento se incrementa ligeramente del 72 % al 73 %. Estas mejoras específicas se ven respaldadas por la comparación global en la fase de testeo, donde los modelos con capas Bi-LSTM alcanzan una precisión media ponderada del 82 %, superando en dos puntos porcentuales a los modelos con LSTM (80 %), manteniendo además una desviación típica comparable (0.07 frente a 0.06). Esto sugiere que, aunque la mejora general no es drástica, Bi-LSTM ofrece una ventaja consistente en distintos contextos, consolidándose como una alternativa más robusta y fiable para la clasificación de escenas violentas.

- En cuanto al análisis de hiperparámetros en arquitecturas que combinan la red VGG-19 preentrenada con capas LSTM y Bi-LSTM, si bien no se han identificado mejoras estadísticamente significativas en la *accuracy* obtenida tras evaluar 200 combinaciones distintas, el estudio permite comprender mejor la estabilidad del modelo ante variaciones en parámetros clave. En particular, no se observaron patrones claros de impacto en el rendimiento al ajustar el número de neuronas en las capas LSTM o densas, ni al modificar la tasa de aprendizaje, lo cual sugiere una relativa robustez del modelo frente a estos cambios. Esta estabilidad aporta valor desde la perspectiva de la explicabilidad, ya que permite identificar qué aspectos del modelo son menos sensibles al ajuste. Cabe señalar que solo se exploró aproximadamente un 10 % del espacio de combinaciones posibles debido a restricciones computacionales, por lo que se considera necesario un análisis más exhaustivo para descubrir configuraciones que puedan aportar mejoras adicionales en el desempeño.

- Extendiendo el análisis para comparar los resultados con los obtenidos por otras arquitecturas y en relación con la posible mejora de VGG-19 respecto a VGG-16 en arquitecturas combinadas con capas LSTM y Bi-LSTM, se evidencia que, **a pesar de que VGG-19 cuenta con un mayor número de capas convolucionales, su desempeño en términos de *accuracy* no supera al de VGG-16** en varios conjuntos de datos. Por ejemplo, en el dataset *Hockey Fights*, VGG-19 obtiene una *accuracy* del 97 % con un modelo Bi-LSTM, mientras que un trabajo que emplea VGG-16 alcanza una *accuracy* superior del 99.5 % (Vijeikis et al., 2022). En el caso del dataset *Violent Flow*, el modelo desarrollado en esta Tesis Doctoral que combina VGG-19 junto con características manuales un 96 % con características manuales, que es inferior al 98.4 % registrado por VGG-16 con Bi-Conv-LSTM (Mugunga et al., 2021b). Estos resultados indican que al combinar con capas LSTM y capas Bi-LSTM, una red más profunda como VGG-19 genera características más complejas o redundantes, lo que dificulta la extracción de patrones temporales relevantes por parte del LSTM. En cambio, VGG-16, al ser más simple, producir representaciones más compactas y manejables, facilitando así el modelado secuencial.

Capítulo 5

Arquitectura explicable basada en VGG-19 preentrenada con capas LSTM/Bi-LSTM

Capítulo 5

Arquitectura explicable basada en VGG-19 preentrenada con capas LSTM/Bi-LSTM

Índice

El aumento en el uso de la inteligencia artificial ha despertado una preocupación notable sobre la fiabilidad de los algoritmos, motivando la creación de informes para establecer definiciones y estándares claros, liderados por instituciones como la Comisión Europea (European Commission, 2019). La **identificación de actos violentos en tiempo real**, facilitada por los avances tecnológicos, utiliza algoritmos de inteligencia artificial para procesar y examinar imágenes y videos de seguridad (Afra & Alhajj, 2020; Ageed & Zeebaree, 2021; Collins et al., 2021). Sin embargo, el uso generalizado de IA plantea preocupaciones sobre su fiabilidad y ética (European Commission, 2019; ISO/IEC, 2020). En este contexto, la **XAI** se destaca como una herramienta crucial para garantizar la transparencia y fiabilidad de los sistemas de IA (Abhishek & Kamath, 2022; Speith, 2022). La XAI permite comprender cómo los algoritmos de IA toman decisiones, lo que es fundamental para aplicaciones críticas como la detección de violencia (European Commission, 2019; Speith, 2022). No obstante, aunque las arquitecturas y modelos propuestos para la detección de violencia en video han demostrado obtener buenos resultados en términos de precisión y robustez, uno de los aspectos clave que falta en

la mayoría de estos enfoques es la explicabilidad del modelo (Chamola et al., 2023; Kaur et al., 2022; López-Blanco et al., 2024). La capacidad de comprender por qué un determinado acto se clasifica como violento es fundamental para tomar decisiones informadas y mejorar la confianza en los sistemas automatizados. En los modelos presentados en los capítulos 3 y 4, aunque se ha logrado un rendimiento destacado, el análisis de los hiperparámetros y la falta de claridad sobre las decisiones del modelo muestran que la explicabilidad sigue siendo un reto importante. A pesar de evaluar 200 combinaciones de hiperparámetros, no se observó una mejora significativa en la *accuracy*, lo que subraya la necesidad de modelos que no solo sean precisos, sino también transparentes y capaces de justificar sus predicciones en tareas sensibles como la detección de violencia.

Con el fin de mejorar la explicabilidad de los modelos presentados en los capítulos 3 y 4, en este Capítulo 5 se proponen los siguientes objetivos, alineados con los planteados en el Capítulo 1, basados en la arquitectura combinada de VGG-19 preentrenada con capas LSTM/Bi-LSTM:

- Investigar técnicas de inteligencia artificial explicable (XAI) para mejorar la explicabilidad en la extracción de fotogramas característicos mediante la diferencia de fotogramas. Estas técnicas estarán basadas en la arquitectura YOLOv8 para la detección de objetos y la selección de fotogramas relevantes. *En línea con los objetivos* ***OB4*** *y* ***OB6*** *planteados en esta Tesis Doctoral.*

- Emplear Grad-CAM para extraer información visual sobre la atención espacial de VGG-19 en los fotogramas, con el fin de visualizar las áreas específicas a las que el modelo recurre para detectar actos de violencia. Esta técnica permitirá comprender mejor el funcionamiento del modelo y su enfoque en las características relevantes de las imágenes. *Alineado con los objetivos* ***OB4*** *y* ***OB6*** *de esta Tesis Doctoral.*

- Diseñar un algoritmo que cuantifique la relevancia temporal de los fotogramas en una escena violenta, evaluando su importancia dentro del contexto de la escena para mejorar la identificación de momentos clave. *De acuerdo con los objetivos* ***OB4*** *y* ***OB6*** *establecidos en esta Tesis Doctoral.*

El presente Capítulo se divide en cuatro Secciones. La Sección 5.1 expone la arquitectura explicable de detección de violencia en vídeo creada en esta Tesis Doctoral. La Sección 5.3 presenta los resultados obtenidos durante la aplicación de las técnicas de XAI, centradas

en la parte del proceso de selección de *frames* característicos. La Sección 5.4 trata los resultados obtenidos durante la aplicación de las técnicas de XAI, centradas en la parte detección de violencia. La Sección 5.5 analiza las conclusiones de este Capítulo.

5.1. Arquitectura VGG-19-LSTM/Bi-LSTM explicable

Esta Sección contiene la propuesta de una arquitectura explicable a partir de la arquitectura propuesta en el Capítulo 4 que combina el uso de VGG-19 junto con capas LSTM-BI-LSTM. La Sección 5.1.1 expone las técnicas de explicabilidad empleadas en la arquitectura explicable propuesta en este Capítulo. La Sección 5.1.2 presenta la arquitectura explicable creada en esta Tesis Doctoral.

5.1.1. Técnicas de explicabilidad empleadas en esta Tesis Doctoral

Esta Sección presenta las técnicas de explicabilidad empleadas en esta Tesis Doctoral en la arquitectura propuesta expuesta en la Sección 5.1.2

En primer lugar, se expone el algoritmo **YOLO (You Only Look Once)** es un algoritmo de detección de objetos diseñado para operar en tiempo real. Su principal ventaja radica en que procesa la imagen completa en una única pasada por la red, a diferencia de los métodos basados en propuestas de regiones, que requieren varias etapas. Para la detección de personas, YOLO divide la imagen en una cuadrícula y, para cada celda, predice múltiples bounding boxes junto con la probabilidad de que cada una contenga un objeto de una clase determinada. Cada bounding box incluye coordenadas espaciales, dimensiones y un valor de confianza que refleja la probabilidad de presencia de una persona en esa región. Esta estructura permite detectar simultáneamente múltiples individuos en escenas complejas y dinámicas con alta eficiencia Diwan et al., 2023. Además, YOLO incorpora técnicas como Non-Maximum Suppression (NMS) para eliminar detecciones redundantes y mejorar la precisión. Gracias a su equilibrio entre velocidad y exactitud, YOLO es ampliamente utilizado en aplicaciones como videovigilancia, robótica y análisis automático de vídeo, donde la detección rápida y fiable de personas es fundamental Zhang et al., 2022. En el contexto de esta tesis y aplicado a la extracción de fotogramas característicos, YOLO ofrece una solución óptima para la detección de personas gracias a su capacidad de procesar cada imagen en una única pasada y generar bounding boxes precisos en tiempo real. Esta eficiencia

computacional y su alta fiabilidad en escenarios dinámicos facilitan la integración de sistemas de detección de violencia basados en vídeo.

En segundo lugar, la técnica de **diferencia de *frames*** consiste en calcular la sustracción pixel a pixel entre fotogramas consecutivos para resaltar regiones con movimiento significativo. En el contexto de esta tesis, aplicar dicha diferencia permite identificar automáticamente los instantes en los que se producen cambios abruptos o patrones de desplazamiento característicos de escenas violentas. Al generar un mapa de movimiento, podemos seleccionar aquellos fotogramas con mayor actividad dinámica como candidatos a *frames* característicos, reduciendo el volumen de datos y enfocando el análisis en los momentos más informativos. Esto no solo mejora la eficiencia del sistema—al procesar únicamente los fotogramas con mayor probabilidad de contener eventos relevantes—sino que, combinado con las métricas de importancia y los mapas de calor de Grad-CAM, ofrece una visión complementaria que integra señal estática y dinámica para una detección de violencia más robusta.

Pasando al algoritmo de **Grad-CAM (Gradient-weighted Class Activation Mapping)** es una técnica de visualización que permite interpretar las decisiones de redes neuronales convolucionales (CNN) mediante la generación de mapas de calor sobre imágenes de entrada Sahatova y Balabaeva, 2022. Este método calcula los gradientes de la predicción de una clase específica respecto a las activaciones de una capa convolucional intermedia, lo que permite ponderar las características espaciales más relevantes para dicha clase. El resultado es un mapa de calor que destaca las regiones de la imagen que más influyen en la predicción del modelo Abhishek y Kamath, 2022. En el contexto de esta Tesis y aplicado a la explicabilidad de las características espaciales extraídas por parte de la CNN, Grad-CAM se utiliza para generar mapas de calor sobre las imágenes procesadas por la CNN, con el fin de identificar visualmente las zonas que el modelo considera importantes para detectar violencia. Esta visualización facilita la comprensión del comportamiento interno de la red y aporta un nivel de explicabilidad que es fundamental para validar y mejorar los sistemas de detección basados en aprendizaje profundo. De esta forma, Grad-CAM contribuye a aumentar la confianza en las predicciones del modelo, al mostrar explícitamente las áreas de interés dentro de las imágenes.

Por último, se pasa a exponer el algoritmo creado en esta Tesis Doctoral enfocado en proporcionar explicabilidad sobre **cuáles de los *frames* del vídeo son más**

relevantes para la detección de violencia. Hasta donde tiene conocimiento el autor de esta Tesis Doctoral, no existen algoritmos que proporcionen explicabilidad sobre la relación entre eventos, en este caso entre *frames*, excepto TimeSHAP. En este caso, se proporciona una solución diferente para entender qué *frames* son más relevantes para la detección de violencia.

Con este propósito, se ejecuta un bucle que elimina un número determinado experimentalmente de *frames* del vídeo en cada iteración, comenzando desde el principio y terminando al final del vídeo, y agrega el último *frame* del vídeo tantos frames como han sido eliminados. Esta operación implica hacer predicciones sin esos *frames* para cada combinación de *frames* eliminados, por lo que si el porcentaje de *accuracy* de esa combinación es menor que el original, esos *frames* son relevantes para la predicción. Así, cuanto menor sea la *accuracy* predicha, más relevantes son dichos *frames*. A continuación se expone un ejemplo, eliminando recursivamente un *frame*.

- **Paso 1:** Se elimina el *frame F1*.
  - Input al modelo: *F2*, *F3*, *F4*, ..., *F10* + *F10* (*F10* repetido para mantener la longitud del vídeo).
  - Nueva *accuracy*: 85 %.
  - Como la precisión bajó, concluimos que *F1* es un *frame* relevante.
- **Paso 2:** Se elimina el *frame F2*.
  - Input al modelo: *F1*, *F3*, *F4*, ..., *F10* + *F10*.
  - Nueva *accuracy*: 88 %.
  - La precisión bajó menos que con el *frame F1*, por lo tanto, *F2* es un *frame* menos relevante.
- **Paso 3:** Se elimina el *frame F3*.
  - Input al modelo: *F1*, *F2*, *F4*, ..., *F10* + *F10*.
  - Nueva *accuracy*: 90 %.
  - Como no hubo descenso en la *accuracy*, el *frame F3* no es relevante.
- **(...)**

Para calcular la relevancia de estos *frames* en la detección de violencia para un vídeo dado, en esta Tesis Doctoral se crea la siguiente métrica denominada importancia, donde tanto la accuracy del vídeo original como la del vídeo modificado son valores que varían entre 0 y 1. Por lo tanto, la importancia resultante es adimensional, lo que permite interpretarla como una relación entre estas dos precisiones.

Sea un vídeo compuesto por $n_f \in \mathbb{N}$ *frames*. Consideramos la versión i-ésima del vídeo, denotada como V_i, obtenida al eliminar un subconjunto de $n_i \in (0, n_f)$ *frames* del vídeo original. Definimos la **importancia** de dicho conjunto de *frames*, denotada como $I_i \in \mathbb{R}^+$, como la razón entre la precisión (*accuracy*) del modelo al evaluar el vídeo original y la precisión obtenida al evaluar V_i:

$$I_i = \frac{\text{acc}_{n_f}}{\text{acc}_{n_i}} \tag{5.1}$$

donde:

- $\text{acc}_{n_f} \in (0, 1)$ es la *accuracy* del vídeo original (con los n_f *frames*),
- $\text{acc}_{n_i} \in (0, 1)$ es la *accuracy* del modelo sobre la versión modificada V_i, con $n_f - n_i$ *frames*.

Esta métrica es adimensional y permite cuantificar la relevancia del conjunto eliminado: cuanto mayor sea I_i, mayor será la importancia de los *frames* eliminados para la tarea de detección de violencia. La **Interpretación de la métrica** I_i se expone a continuación:

- $I_i > 1$: Indica que la eliminación de esos *frames* reduce la precisión del modelo, sugiriendo que son relevantes para la detección de violencia.
- $I_i = 1$: Indica que la eliminación de los *frames* no afecta la precisión del modelo, sugiriendo que esos *frames* no son importantes.
- $I_i < 1$: Indica que el modelo es más preciso sin esos *frames*, sugiriendo que podrían estar introduciendo ruido o información irrelevante.

A continuación, se presenta la Figura 5.1, que ilustra la métrica de importancia I_i calculada para cada *frame* del vídeo. Esta visualización tiene como objetivo facilitar la comprensión de qué *frames* son más relevantes para la detección de violencia por parte del modelo. En la gráfica, el eje horizontal representa el número de *frame* dentro

del vídeo, mientras que el eje vertical muestra el valor de importancia I_i. Los diferentes colores de los puntos permiten identificar rápidamente la relevancia de cada *frame*:

- **Rojo (valores de importancia superiores a 1):** *frames* cuya eliminación disminuye la *accuracy* del modelo, por lo que son altamente relevantes.
- **Azul (valores de importancia iguales a 1):** *frames* cuya eliminación no afecta la *accuracy*, considerándose neutros.
- **Verde (valores de importancia inferiores a 1):** *frames* cuya eliminación mejora la *accuracy*, sugiriendo que podrían ser irrelevantes o introducir ruido.

La Figura 5.1 muestra un pico claro de importancia en el centro del vídeo, destacando un conjunto de *frames* críticos para la detección de violencia. Esta representación gráfica contribuye a interpretar temporalmente la importancia de los *frames* y a explicar el comportamiento del modelo de forma visual y sencilla.

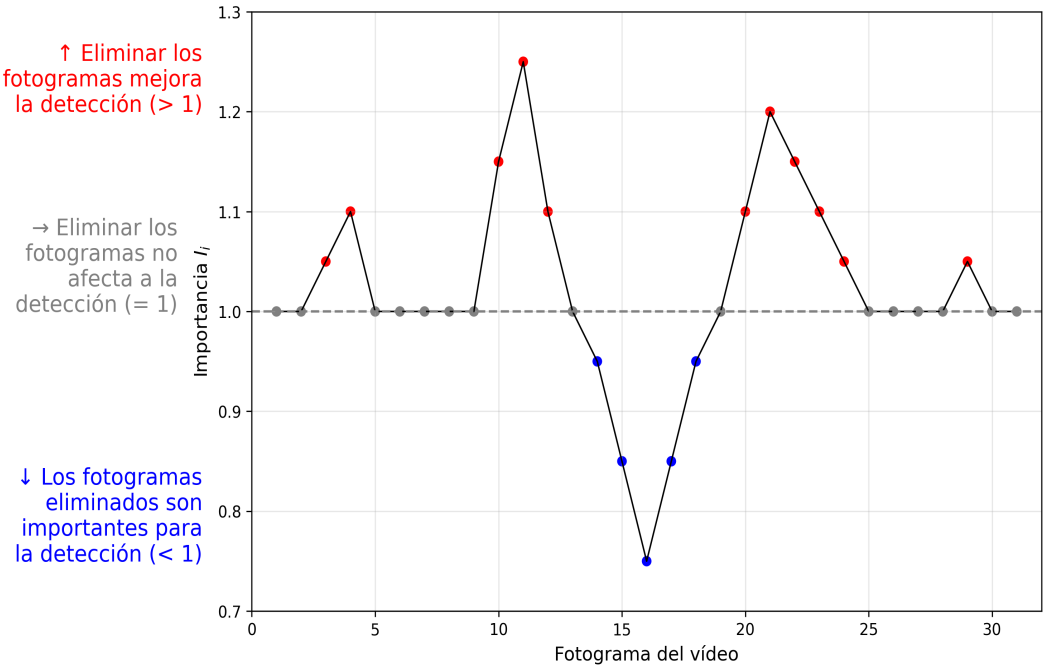

FIGURA 5.1: Ejemplo artificial de la relevancia de cada frame en la detección de violencia.

En el contexto de esta Tesis y aplicado a la explicabilidad de las características temporales extraídas por parte de las capas LSTM/Bi-LSTM, este método es relevante ya que permite comprender qué partes del vídeo son temporalmente significativas para la detección de violencia. Además, la aplicación iterativa del modelo no tiene un coste

computacional elevado, ya que solo se aplica a las capas LSTM/Bi-LSTM y capas densas, a partir de las características espaciales extraídas por VGG-19, lo que hace que su cálculo sea rápido, del orden de dos segundos.

5.1.2. Arquitectura explicable propuesta en esta Tesis Doctoral

Esta Sección presenta la arquitectura del modelo de detección de violencia en vídeo mostrado en la Figura 4.2 expuesta en el Capítulo 4, junto con la aplicación de técnicas de inteligencia artificial explicable expuestas en la Sección 5.1.1. Esta arquitectura se muestra en la Figura 5.2, donde se implementa una vez que el modelo de detección de violencia está entrenado. Cabe destacar que los ejemplos presentados en este capítulo se basan en el conjunto de datos de vídeos con alta visibilidad (*Hockey Fights dataset*). Esta elección responde exclusivamente a que sus primeros planos y condiciones de iluminación facilitan una mejor comprensión visual de los resultados obtenidos.

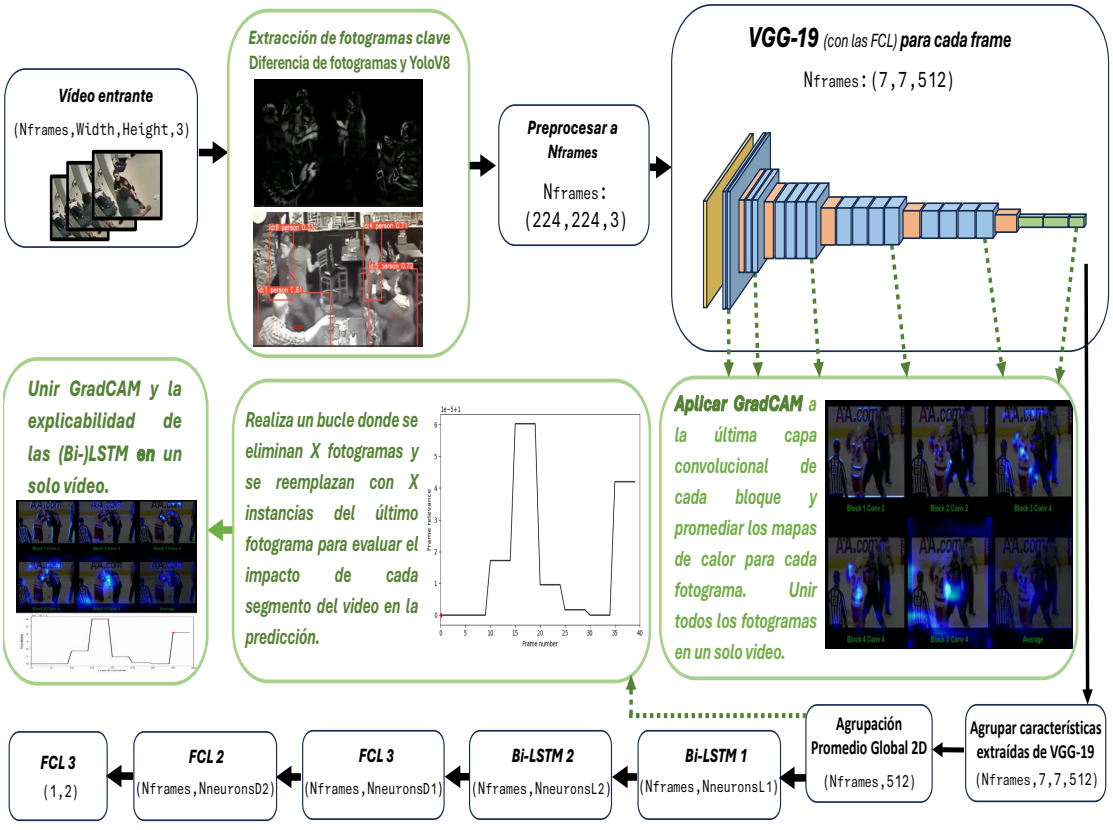

FIGURA 5.2: Arquitectura de detección de violencia explicable propuesta en esta Tesis Doctoral.

- Como primer elemento que proporciona explicabilidad, se aplica un proceso de **extracción de *frames* clave basado en el uso de diferencia de *frames* y YoloV8**. Se expone en el Capítulo 2, Sección 2.3, cómo el análisis de vídeo en tiempo real es costoso computacionalmente, por lo que es importante aplicar modelos livianos que seleccionen *frames* con probabilidad de contener violencia. Se aplica YoloV8 como algoritmo de detección de personas rápido y optimizado. Por otro lado, la diferencia de *frames* resalta las partes del vídeo donde se detecta más movimiento.

  Ambos algoritmos se emplean por separado en el estado del arte y se propone la combinación de ambos para minimizar el número de *frames* clave analizados por el algoritmo de detección de violencia. Dependiendo del tipo de escena que la cámara de vídeo esté analizando, se debe establecer un número mínimo de personas a detectar por YoloV8 y un valor mínimo de diferencia de *frames* para pasar estos *frames* al algoritmo de detección de violencia cuando se cumplan ambas condiciones. Este primer elemento implica comprender por qué la arquitectura considera relevantes ciertos *frames* como candidatos para la detección de violencia.

- El segundo elemento de explicabilidad aplicado en la estructura propuesta es **GradCAM**, empleado sobre VGG-19. Se configura GradCAM para obtener los gradientes de la última capa convolucional de cada bloque, lo que nos permite entender en qué partes está poniendo el foco VGG-19. En consecuencia, se obtienen un total de cinco mapas de calor, que se superponen en los *frames* originales. Además, se calcula el promedio de los 5 mapas de calor, aunque inicialmente se presume que la última capa convolucional contiene el conocimiento más relevante, maximizando así las áreas más prominentes de cada bloque.

  En última instancia, se obtiene un vídeo con seis sub-vídeos divididos en cuadrículas para cada mapa de calor superpuesto con el vídeo original. Aplicar GradCAM de esta manera significa una comprensión de cómo VGG-19 procesa los *frames* de vídeo y dónde se centra para localizar la violencia.

- El tercer elemento de explicabilidad es el algoritmo creado en esta Tesis Doctoral para la explicabilidad de la importancia temporal de ciertos fotogramas del vídeo en la detección de violencia. Se escoge eliminar recursivamente el valor de cinco *frames* consecutivos por dos razones: por un lado, se ha observado empíricamente que la eliminación de un solo *frame* no produce variaciones significativas en la *accuracy*;

por otro, cinco *frames* corresponden aproximadamente a un sexto de segundo, lo cual es suficiente para capturar eventos de corta duración como actos violentos. Esta decisión permite detectar cambios relevantes sin perder la generalidad de la arquitectura propuesta.

- Finalmente, los resultados de la detección por parte de Yolo, la diferencia de *frames* y, por otro lado, los seis mapas de calor generados por GradCAM y su unificación en un solo vídeo junto con la serie temporal son almacenados. Este proceso de explicabilidad proporciona información valiosa al usuario desde el principio hasta el final del proceso, especialmente considerando que las CNN y las LSTM son algoritmos completamente opacos.

5.2. Implementación de la arquitectura explicable

Esta Sección expone la implementación de la arquitectura expuesta en la Sección 5.1. Al igual que en los Capítulos 4 y 5, todas las computaciones de esta Tesis Doctoral se realizan en un servidor con un procesador Intel(R) Core(TM) i9-10940X con 188 gigabytes de RAM y una GPU NVIDIA GeForce RTX 3090 con 24 gigabytes de memoria.

Para la detección de escenas potencialmente violentas, se implementa un procedimiento basado en la detección de personas mediante la red **YOLOv8** y el análisis de diferencias entre fotogramas consecutivos. En primer lugar, se emplea un modelo preentrenado (`yolov8n.pt`) para detectar personas en cada fotograma del video de entrada. Se considera una escena como candidata a violencia únicamente si se detectan al menos dos personas en el mismo fotograma, lo cual permite restringir el análisis a interacciones humanas relevantes. Cada fotograma se procesa secuencialmente mediante el método de `tracking`, permitiendo además la asignación de identificadores únicos y el seguimiento de trayectorias. Paralelamente, se calcula la **diferencia de intensidad entre fotogramas consecutivos** con el objetivo de cuantificar el movimiento o cambio brusco en la escena. Este cálculo se realiza convirtiendo cada fotograma a escala de grises y aplicando una diferencia absoluta entre imágenes sucesivas. Se extraen estadísticas básicas como la media, mediana, desviación estándar, mínimo y máximo del mapa de diferencia, que sirven como medidas indirectas de la actividad o agitación visual. Finalmente, los mapas de diferencia se almacenan como video para su análisis

posterior, permitiendo una visualización clara de los cambios dinámicos en el entorno. Esta combinación de criterios permite la seleccion de fotogramas potencialmente violentos.

Pasando ahora a la implementación de **GradCAM**, como ya se ha expuesto, esta es una técnica de visualización que permite identificar qué regiones de una imagen influyen más en la decisión de una red neuronal convolucional. La implementación desarrollada en esta Tesis Doctoral, basada en la red **VGG-19**, se expone a continuación, además de mostrarse un ejemplo resultante de un fotograma del dataset *Hockey Fights* en la Figura 5.6:

- Se generan visualizaciones sobre la última capa de cada uno de los cinco bloques convolucionales de VGG-19, seleccionando capas profundas del modelo previamente entrenado.
- El video de entrada se divide en un número fijo de fotogramas (*frames*), seleccionados equidistantemente. Cada fotograma se redimensiona a un tamaño de 224×224 píxeles, y se normaliza para tener valores entre 0 y 1, dividiendo cada canal RGB por 255.
- Para cada fotograma, se obtiene la predicción de clase a partir del modelo, y se calculan los gradientes de dicha clase con respecto a la salida de la capa convolucional seleccionada.
- Se realiza un promedio espacial de los gradientes sobre cada canal, lo que permite estimar la importancia relativa de cada mapa de activación.
- Las activaciones de la capa convolucional se combinan ponderadamente con estos valores medios de los gradientes, generando así un mapa de activación (*heatmap*) que destaca las regiones más influyentes.
- Este mapa se redimensiona al tamaño original del fotograma y se superpone visualmente, ya sea con un esquema de color (por ejemplo, *hot colormap*) o ajustando la opacidad del canal alfa.
- El proceso se repite para cada una de las capas seleccionadas, guardando un video por capa para su análisis individual.

- Finalmente, se calcula el promedio de todos los mapas de calor generados, obteniendo una visualización agregada. El resultado se organiza en una cuadrícula de videos unificados, facilitando la comparación entre capas.

FIGURA 5.3: Mapas de calor generados mediante Grad-CAM aplicados sobre la última capa convolucional de cada bloque de la red VGG-19, junto con el mapa promedio resultante. Se utiliza como entrada un fotograma representativo de un vídeo del dataset *Hockey Fights*.

Por último, con respecto a la implementación del **método de explicabilidad desarrollado en esta Tesis Doctoral para cuantificar la importancia temporal de los *frames*** en la detección de violencia (ecuación 5.1), se presentan en la Tabla 5.1 un experimento diseñado para analizar cómo afecta la eliminación de distintos bloques de *frames* (1, 2, 5 y 10) en la salida del modelo. Para mantener constante la longitud de entrada, los *frames* eliminados fueron sustituidos por repeticiones del último *frame* existente. Para cada caso, se calcularon varias métricas agregadas sobre el dataset con escenas de alta visibilidad *Hockey Fights*:

- La **mediana de la media de importancias por vídeo**, como indicador robusto de cambio global.

- La **desviación estándar entre las medias de importancia de todos los vídeos**, para reflejar la variabilidad inter-vídeo.
- La **media y la mediana de las desviaciones estándar dentro de cada vídeo**, como medida de inestabilidad local (*frame* a *frame*).
- El **tiempo medio de computación** de la importancia por vídeo.

Número de frames eliminados de forma secuencial	**Mediana de la media de importancia por vídeo**	**Desviación estándar entre las medias de importancia**	**Media de la desviación estándar por vídeo**	**Mediana de la desviación estándar por vídeo**	**Tiempo medio de cómputo por vídeo (s)**
1	1.0000	0.0059	0.0060	0.0001	1.94
2	1.0010	0.0119	0.0118	0.0001	0.95
5	1.0001	0.3169	0.1296	0.0002	0.46
10	1.0001	0.6644	0.1803	0.0003	0.28

TABLA 5.1: Resultados detallados en función del número de *frames* eliminados secuencialmente.

Se observa que la **mediana de las medias de importancia por vídeo permanece extremadamente próxima a 1**, lo que implica que, en términos globales, la predicción del modelo no se ve sustancialmente afectada por la eliminación de hasta 10 *frames*. Sin embargo, la **variabilidad entre vídeos** (columna 3) se incrementa considerablemente a partir de 5 *frames* eliminados, lo que sugiere que ciertos vídeos contienen secuencias críticas cuya ausencia altera la salida más que en otros.

Adicionalmente, la **desviación estándar interna a cada vídeo** (columnas 4 y 5) también muestra un crecimiento moderado, lo cual indica que, además de diferenciarse entre vídeos, también hay fluctuaciones mayores dentro de la misma secuencia cuando se alteran múltiples bloques temporales.

Por tanto, se concluye que la eliminación de bloques de **5 *frames*** representa un compromiso óptimo entre:

- Detectar la sensibilidad del modelo a regiones temporales relevantes.
- Minimizar la distorsión general de la predicción.
- Reducir el coste computacional (menos de medio segundo por vídeo).
- Evitar excesiva sobresimplificación como ocurre al eliminar 10 *frames*, lo cual reduce aún más el tiempo, pero genera alta variabilidad.

5.3. Resultados explicables en la extracción de fotogramas característicos

En esta Sección se presentan los resultados obtenidos durante la aplicación de las técnicas de XAI expuestas en la Figura 5.2 centradas en la parte del proceso de selección de *frames* característicos. Por un lado, la inteligencia artificial explicable es un campo que aún se encuentra en desarrollo (Mahmoudi et al., 2023). Por otro lado, cuantificar la relevancia de un área específica de un *frame* en la detección de violencia, así como determinar la importancia de un *frame* en un vídeo violento, constituye una tarea compleja.

En primer lugar, la aplicación de la **diferencia de *frames*** aporta explicabilidad al resaltar zonas del vídeo donde se producen variaciones bruscas de píxeles, asociadas a movimientos significativos. Esto permite vincular visualmente la predicción del modelo con momentos concretos en los que ocurren acciones relevantes, como una pelea o un empujón. El valor se calcula de forma experimental como la media de las diferencias absolutas de intensidad entre píxeles en dos *frames* consecutivos. Este valor depende de la posición de la cámara: por ejemplo, cuando el sujeto aparece en primer plano y ocupa buena parte del encuadre, una acción violenta puede generar un valor medio de diferencia de *frames* superior a 40 unidades. En cambio, si la cámara está más alejada, la misma acción produce valores considerablemente menores, debido a la menor variación de píxeles visibles. Además, condiciones de iluminación muy cambiantes entre *frames* también pueden alterar significativamente estos valores, generando falsos positivos si no se controlan adecuadamente. Se representa en la Figura 5.4

Figura 5.4: *Frame* del vídeo *fi50_xvid* del dataset *Hockey Fights* (izquierda) al que se le aplica la diferencia de *frames*

Con respecto a la **detección del número de personas en la escena por parte de YoloV8**, se muestra un ejemplo en la Figura 5.5. En esta se observa cómo la **oclusión entre personas** y con otros objetos puede impedir el conteo correcto en los *frames* donde dicha oclusión ocurre. Sin embargo, dado que YoloV8 se implementa con la opción de seguimiento de objetos, el modelo mantiene la identidad de una persona ya detectada incluso si queda parcialmente oculta, hasta que desaparece completamente de la escena. Se puede observar cómo, aunque el segundo jugador de hockey está oculto por el jugador con la camiseta número 24, YoloV8 lo sigue detectando, lo incluye en el recuento de personas en la escena e identifica correctamente al segundo árbitro como la cuarta persona implicada en la acción.

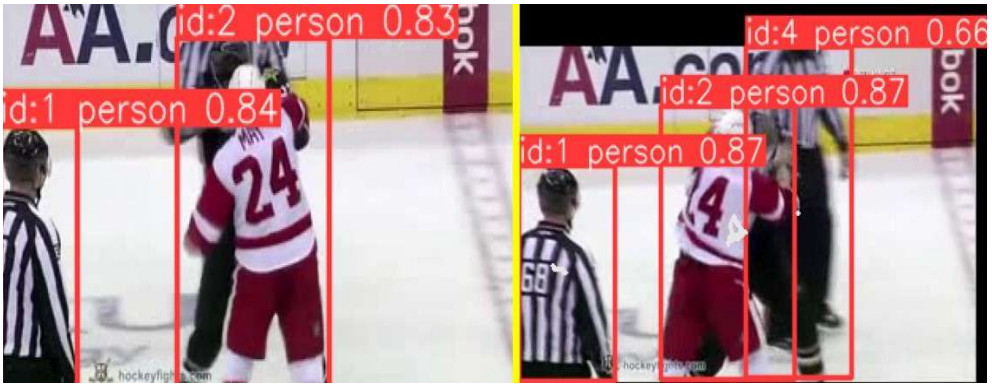

FIGURA 5.5: Detección de personas con oclusión entre ellas en uno de los vídeos del dataset *Hockey Fights*.

Con todo, la aplicación conjunta de la extracción de *frames* clave utilizando la combinación de **YoloV8** y la diferencia de *frames* ha dado resultados positivos en varios aspectos. Por un lado, el tiempo de computación de la diferencia de *frames* se encuentra, de media, en el orden de 0.01 *ms*, mientras que el de YoloV8 se encuentra en un promedio de 45 *ms* al realizar la detección de personas, lo que cumple su propósito de aplicación en tiempo real. La condición que debe cumplirse es que YoloV8 reconozca un número mínimo de personas en la escena, mientras que la diferencia de *frames* debe alcanzar un valor mínimo específico para que el contenido se considere potencialmente violento y pueda ser procesado adecuadamente por el algoritmo de detección de violencia a partir de ese momento.

5.4. Resultados explicables en la detección de violencia

Esta Sección expone la aplicación de técnicas de inteligencia artificial explicable en el algoritmo de detección de violencia. En este caso, se aplica GradCAM a la red VGG-19 y un algoritmo de desarrollo propio a las capas LSTM/Bi-LSTM. Comenzando con la aplicación de **GradCAM** en la predicción de VGG-19, esta resulta ser muy útil para comprender cómo se enfoca la **atención de VGG19 en la última capa de cada bloque convolucional**. Aunque GradCAM se aplica *frame* a *frame*, produciendo un vídeo de mapas de calor, se adjunta un ejemplo muy representativo de lo observado con sus resultados en la Figura 5.6.

FIGURA 5.6: Resultado de GradCAM para la última capa convolucional de cada bloque de VGG-19 y el promedio de mapas de calor, para el *frame* de un vídeo del dataset *Hockey fights*. Las zonas resaltadas son aquellas sobre las que la CNN presta más atención.

Los bloques convolucionales, que consisten en una sucesión de capas convolucionales agrupadas funcionalmente, operan de forma jerárquica extrayendo características visuales de manera progresiva. En las primeras capas, se detectan patrones simples como bordes, contornos y texturas básicas. A medida que la información fluye hacia capas más profundas, se capturan estructuras más complejas, como formas corporales o elementos

específicos del entorno. Finalmente, los bloques más profundos tienden a enfocarse en regiones semánticamente relevantes para la tarea, como interacciones entre objetos o sujetos. Esta evolución en la representación permite que la red aprenda desde patrones generales hasta detalles clave, como se analiza a continuación Omarov et al., 2022.

La Figura 5.6 muestra los mapas de calor generados por Grad-CAM para la última capa convolucional de cada bloque de la red VGG-19, la cual ha sido seleccionada como base de la arquitectura propuesta en este capítulo (ver Figura 5.2). Cada uno de estos bloques resalta las regiones espaciales de la imagen en las que la CNN focaliza su atención durante el proceso de inferencia. Esto permite interpretar de forma visual qué partes del fotograma influyen en la toma de decisiones del modelo, facilitando así un análisis más comprensible y explicable del comportamiento de la red. A continuación, se detalla cómo varía esta atención a lo largo de los diferentes bloques de la red:

- **Bloques 1 y 2:** Se enfocan principalmente en los contornos y patrones básicos, como los cuerpos, la ropa y las señales visibles en la escena.
- **Bloques 3 y 4:** Resaltan la posición de los brazos y del cuerpo de los jugadores. Sin embargo, también muestran atención no deseada hacia elementos irrelevantes, como señales y el árbitro ubicado a la izquierda del fotograma.
- **Bloque 5 (último):** Centra su atención en el área de interacción entre los dos jugadores de hockey, que es donde se localiza el acto de violencia. A pesar de ello, sigue resaltando parcialmente y de forma incorrecta al árbitro en el lado izquierdo del *frame*.
- **Promedio de mapas de calor (último recuadro):** Refuerza la zona de interacción identificada por el bloque 5 y resalta además los cuerpos de los jugadores. No obstante, también incluye algunas áreas destacadas que no están relacionadas con la violencia.

Una de las utilidades interesantes consideradas para la aplicación de GradCAM es **poder verificar que el algoritmo se centra correctamente en las áreas donde ocurre la violencia** al predecir un acto violento, y observar en qué áreas no se enfoca al hacer predicciones incorrectas. Esto puede facilitar la comprensión de las debilidades del modelo y equilibrar el conjunto de datos de entrenamiento en consecuencia.

Con respecto al **método de explicabilidad desarrollado en esta Tesis Doctoral para cuantificar la importancia temporal de los *frames*** en la detección de violencia (ecuación 5.1), los resultados obtenidos a través de la eliminación secuencial de bloques de *frames* muestran que, aunque la mediana de la media de importancia por vídeo permanece prácticamente inalterada incluso al eliminar hasta 10 *frames*, existe una marcada variabilidad inter-vídeo a partir de la supresión de 5 *frames*. Este comportamiento sugiere que ciertos vídeos contienen secuencias temporalmente críticas cuya ausencia afecta de forma más pronunciada la respuesta del modelo. Asimismo, se observa un incremento progresivo en la desviación estándar dentro de cada vídeo, lo cual indica una mayor inestabilidad local en la importancia asignada *frame* a *frame*. De este modo, la eliminación de bloques de 5 *frames* se identifica como un punto de equilibrio entre detectar regiones temporales relevantes, mantener la estabilidad global de las predicciones, y optimizar el tiempo de cómputo (menos de 0.5 segundos por vídeo). A modo de ejemplo ilustrativo, la Figura 5.7 muestra cómo el método detecta una menor relevancia en *frames* con plano perpendicular a la agresión, donde los agresores son visualmente más distinguibles. En conjunto, los experimentos validan la eficacia del método propuesto como herramienta de análisis interpretativo para modelos LSTM/Bi-LSTM aplicados a clasificación de vídeo, al permitir identificar con precisión la relevancia temporal de los *frames* sin incurrir en altos costes computacionales.

5.5. Discusión

En este Capítulo se ha propuesto una arquitectura de detección de violencia en vídeo que incorpora técnicas de inteligencia artificial explicable (XAI) sobre la combinación de VGG-19 y capas LSTM/Bi-LSTM. A partir de las distintas secciones del capítulo, podemos extraer tres conclusiones principales:

1. **Cuantificación de la importancia temporal de los *frames* (Sección 5.1).** La métrica adimensional creada en la Ecuación 5.1, Sección 5.1.1, llamada Importancia (I_i), permite identificar con claridad qué segmentos del vídeo aportan mayor o menor valor a la detección de violencia.

2. **Selección de *frames* clave y eficiencia computacional (Sección 5.2).** La combinación de YoloV8 (45 ms por detección de personas) y diferencia de

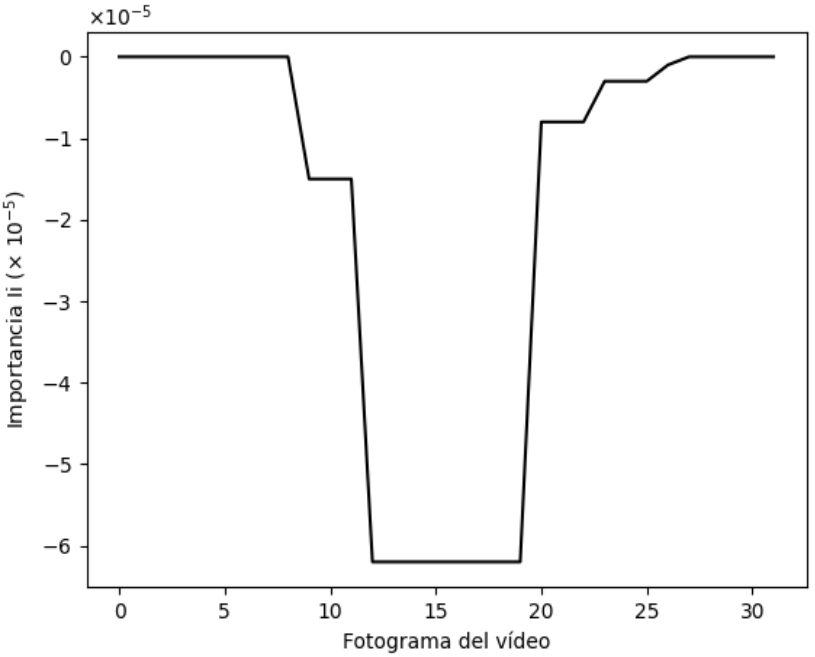

FIGURA 5.7: Resultado del vídeo «fi5_xvvid» perteneciente al dataset *Hockey Fights* mediante la técnica desarrollada para cuantificar la relevancia de *frames* en la detección de violencia.

frames (0,01 ms) mejora la transparencia del preprocesado, pues justifica por qué ciertos *frames* entran al detector de violencia. Además, mantener un umbral de conteo mínimo de personas y un valor de movimiento garantiza:

- Reducción del número de *frames* analizados sin perder precisión.
- Rendimiento en tiempo real adecuado para aplicaciones de videovigilancia.
- Resiliencia frente a oclusiones parciales gracias al seguimiento de objetos de YoloV8.

3. **GradCAM: explicabilidad espacial (Sección 5.3).**

   La aplicación de Grad-CAM sobre las cinco últimas capas convolucionales de VGG-19 ha resultado muy útil para interpretar en que área de la imagen se centra la red al predecir violencia. Los mapas de calor resaltan desde contornos básicos en los primeros bloques hasta el área de interacción violenta en los bloques más profundos, permitiendo:

   - Verificar que la atención se concentra en la zona correcta cuando la predicción es exitosa.
   - Detectar elementos distractores (árbitro, señales) en los bloques intermedios.

- Guiar posibles ajustes de conjunto de datos o arquitecturas posteriores.

En definitiva, la arquitectura propuesta ha demostrado ser adecuada para mejorar la explicabilidad de un sistema de detección de violencia en vídeo. Integrar métricas temporales de relevancia, mapas de atención espaciales y criterios automatizados de selección de *frames* no solo facilita la comprensión de las decisiones del modelo, sino que también optimiza su eficiencia y fiabilidad.

Capítulo 6

Conclusiones y Trabajo Futuro

Capítulo 6

Conclusiones y Trabajo Futuro

Índice

El creciente uso de la inteligencia artificial ha generado una preocupación significativa sobre el desarrollo de algoritmos fiables, lo que ha llevado a la creación de numerosos informes detallados para definir y estandarizar estos términos por parte de importantes organizaciones como la Comisión Europea (European Commission, 2019).

En este ámbito, existen múltiples soluciones propuestas para afrontar la violencia en las sociedades (Vomfell et al., 2018). La detección de violencia en tiempo real es la última barrera para proteger a las víctimas. Entre las tecnologías que permiten esta detección, la inteligencia artificial proporciona una solución directa y eficaz que puede preservar vidas y disminuir la dependencia de la vigilancia humana continua (Omarov et al., 2022). En este contexto, en la literatura se han propuesto múltiples arquitecturas y modelos, donde los algoritmos basados en la combinación de *Convolutional Neural Networks* (CNN) junto con capas *Long Short Term Memory* (LSTM) es la arquitectura que mejores resultados ha obtenido y, por ello, la más utilizada (Islam et al., 2021a; Negre et al., 2024b; Vijeikis et al., 2022).

En este sentido, el objetivo de esta Tesis Doctoral ha sido diseñar e implementar una arquitectura basada en la combinación de CNN junto con el uso de características manuales y capas RNN, que mejore los resultados obtenidos en otras investigaciones y, dote de mayor explicabilidad a estos resultados.

En primer lugar, se ha propuesto una arquitectura basada exclusivamente en redes convolucionales (CNN) junto con características manuales, evaluando los modelos VGG-16 y VGG-19 preentrenados para la detección de violencia en vídeo a nivel de frame. Los resultados muestran que VGG-19 supera de un 2 % a un 10 % en accuracy a VGG-16 en los tres escenarios estudiados: alcanza un 94 % frente al 92 % de VGG-16

en escenarios de alta visibilidad (dataset *Hockey Fights*), un 96 % frente al 92 % en escenarios de alta densidad poblacional (dataset *Violent Flow*, y un un 76 % frente al 64 % en escenarios de baja visibilidad (dataset *RWF-2000*). Aunque VGG-19 reduce ligeramente el tiempo medio de inferencia (1,56 ms vs. 1,64 ms), su mayor precisión justifica su uso. Estos hallazgos confirman que, incluso sin modelado temporal, VGG-19 aporta ventajas significativas sobre VGG-16, asentando la elección de VGG-19 para los siguientes capítulos.

En segundo lugar, se ha ampliado la arquitectura integrando redes recurrentes (LSTM y Bi-LSTM) sobre la extracción de características de VGG-19 para capturar la dinámica temporal en secuencias de vídeo. Los modelos con LSTM y Bi-LSTM elevan la accuracy hasta un 96 % y 97 % en escenarios de alta visibilidad, frente al 94 % / 95 % de las versiones sin recurrentes; en escenarios de alta densidad poblacional logran 96 % (LSTM) y 90 % (Bi-LSTM), comparado con 96 % sin RNN; y en escenarios con baja visibilidad alcanzan 72 % y 73 % frente al 76 % con características manuales. La media ponderada de Bi-LSTM (82 %) supera en dos puntos a LSTM (80 %), demostrando que el análisis temporal aporta valor en contextos de buena visibilidad, aunque su eficacia depende del escenario.

En tercer lugar, se ha completado la arquitectura incorporando mecanismos de explicabilidad que actúan en paralelo sobre las dimensiones temporal, espacial y de selección de *frames*, con el objetivo de mejorar la transparencia del modelo sin afectar su rendimiento. La métrica de importancia temporal identifica segmentos críticos con una relación I_i que varía entre 10^{-3} y 10^{-6}, revelando robustez ante la supresión de información; la estrategia de selección de *frames* combina conteo de personas con diferencia de movimiento para reducir la carga computacional manteniendo precisión y operatividad en tiempo real; y la aplicación de Grad-CAM en cinco capas de VGG-19 muestra atención progresiva desde contornos simples hasta focos de interacción violenta. La integración conjunta de estas herramientas incrementa la confianza en las predicciones y ofrece trazabilidad operativa, lo que resulta clave en contextos de videovigilancia crítica.

En base a ello, otro de los objetivos de esta Tesis Doctoral es implementar técnicas de inteligencia artificial explicable sobre las arquitecturas propuestas para hacer más fiable todo el proceso de detección de violencia. Así, se han implementando técnicas intrínsicamente explicables como YoloV8 (Diwan et al., 2023) para la extracción de

frames característicos, se ha utilizado GradCAM (Chamola et al., 2023; Mahmoudi et al., 2023) como algoritmo de explicabilidad enfocado en CNN para la generación de mapas de saliencia y se ha creado un algoritmo que cuantifica la relevancia de las partes de un vídeo en la detección de violencia.

Este Capítulo sirve, por tanto, para concluir esta Tesis Doctoral y presentar posibles líneas de trabajo futuro. La Sección 6.1 enumera las conclusiones obtenidas durante el trabajo de investigación llevado a cabo en esta Tesis Doctoral, así como su alineamiento con los objetivos planteados al inicio de la misma. La Sección 6.2 propone diferentes líneas de investigación futura posibles que surgen a raíz de los resultados y conclusiones obtenidos por esta Tesis Doctoral.

6.1. Conclusiones

Se han sugerido en la literatura diversas opciones para abordar los casos de violencia en las comunidades (Killgore et al., 2021; Vomfell et al., 2018). En lo que respecta a la detección en tiempo real de actos violentos representa una medida crucial para salvaguardar a las víctimas. En este contexto, la inteligencia artificial ha destacado por su eficacia, generando un ahorro significativo en términos de recursos humanos para la vigilancia y permitiendo la capacidad de analizar múltiples vídeos de manera simultánea (Jing et al., 2021).

En esta Tesis Doctoral se ha desarrollado un estudio sistemático de la literatura reciente enfocada en trabajos que tratasen la detección de violencia en vídeo mediante el uso de inteligencia artificial, centrado en el uso de inteligencia artificial fiable (Negre et al., 2024b), recopilando un total de 63 artículos entre junio de 2021 y junio de 2023. En este estudio desarrollado en el Capítulo 2, se han recopilado y categorizado un total de 28 conjuntos de datos de detección de violencia, lo que pone de manifiesto la creación de nuevos conjuntos en los últimos años (Negre et al., 2024a).

Además, se han compilado 13 parámetros de evaluación utilizados en la detección de violencia, destacando la *accuracy* como el más utilizado. También se han analizado y categorizado un total de 21 métodos clave de extracción de características en los artículos seleccionados, los cuales son esenciales para un procesado en tiempo real más ligera y económico del proceso de detección de violencia en vídeo.

En el análisis de los distintos *inputs* de los algoritmos de detección de violencia, el más

utilizado es el vídeo en color, y la combinación de algoritmos que mejores resultados ha obtenido, y por ello, la más utilizada es la de *CNN* y *LSTM* (Negre et al., 2024a).
Sin embargo, entre los artículos seleccionados en el mapeo sistemático, ninguno hace uso explícito de técnicas de inteligencia artificial fiable. Aunque un total de 15 artículos mencionan la robustez de su modelo, este concepto está centrado en la cantidad de falsos positivos y falsos negativos predichos y no tanto en la robustez como pilar de la fiabilidad. Por otro lado, varios artículos emplean técnicas o algoritmos en el proceso de detección de violencia en vídeo que pueden facilitar la explicabilidad del modelo, como el uso de algoritmos de detección de objetos o algoritmos *Skeleton-based* (Hung et al., 2021; Naik & Gopalakrishna, 2021; Narynov et al., 2021; Zhou, 2022).
En base a los resultados obtenidos en este estudio sistemático, se han desarrollado en esta Tesis Doctoral diversas arquitecturas. En el Capítulo 3 se propone una arquitectura basada en la combinación de CNN junto con el uso de características manuales, seleccionando las redes VGG-16 y VGG-19 preentrenadas en *ImageNet* como CNN candidatas. En el Capítulo 4 se propone una arquitectura basada en la combinación de VGG-19 preentrenada en *ImageNet* junto con el uso de capas LSTM y capas Bi-LSTM.

- **Conclusión 1. Capítulo 2**. Se ha realizado un estudio sistemático de la literatura reciente sobre detección de violencia en vídeo mediante inteligencia artificial fiable, recopilando 63 artículos durante los dos últimos años. Se han categorizado 28 conjuntos de datos, reflejando la creación continua de nuevos datasets, y se han identificado 13 parámetros de evaluación, destacando la precisión (*accuracy*) como el más utilizado. Asimismo, se analizaron 21 métodos clave de extracción de características, esenciales para un procesamiento en tiempo real eficiente. Los algoritmos que combinan redes convolucionales (CNN) con capas LSTM son los más frecuentes y muestran un mejor desempeño en comparación con otras arquitecturas. Por último, aunque ningún estudio aplica explícitamente técnicas completas de inteligencia artificial fiable, varios integran métodos que favorecen la explicabilidad. Estos resultados apoyan la elección de combinar CNN y LSTM como base para las arquitecturas desarrolladas en esta Tesis.

- **Conclusión 2. Capítulo 3**. Se ha diseñado e implementado una arquitectura basada en redes convolucionales (CNN) complementadas con características manuales, evaluando las redes VGG-16 y VGG-19 para la detección de violencia en vídeo a nivel de fotograma. Los resultados evidencian que VGG-19 supera en

precisión (*accuracy*) a VGG-16 en un rango del 2 % al 10 % en los tres escenarios analizados: alcanza un 94 % frente al 92 % en escenarios de alta visibilidad (dataset *Hockey Fights*), un 96 % frente al 92 % en escenarios con alta densidad poblacional (dataset *Violent Flow*) y un 76 % frente al 64 % en condiciones de baja visibilidad (dataset *RWF-2000*). Además, aunque VGG-19 presenta un tiempo medio de inferencia ligeramente inferior (1,56 ms frente a 1,64 ms), su mayor precisión justifica plenamente su selección. Con todo, estos resultados confirman que, incluso sin incorporar modelado temporal, VGG-19 ofrece ventajas notables respecto a VGG-16, consolidando así su elección para los capítulos siguientes.

- **Conclusión 3. Capítulo 4**. Se ha diseñado e implementado una arquitectura basada en la CNN preentrenada VGG-19 combinada con capas LSTM y Bi-LSTM. En base a ello, se ha comprobado que el uso de capas Bi-LSTM supera la precisión (*accuracy*) obtenida por las capas LSTM, aunque esta mejora no ha superado el 4 % en los experimentos realizados (Sección 4.4.2). Además, se evidenció la falta de correlación entre los valores de hiperparámetros —como la tasa de aprendizaje y el número de neuronas en las capas LSTM, Bi-LSTM o densas— y la mejora en la precisión obtenida (Sección 4.4.2). Finalmente, se ha observado que, si bien VGG-19 obtiene mejores resultados que VGG-16 en las arquitecturas propuestas en el Capítulo 3, esto no se mantiene cuando se combinan con capas LSTM o Bi-LSTM, ya que las características espaciales extraídas por VGG-16 generan mejores resultados al integrarse con estas capas.

- **Conclusión 4. Capítulo 5**. Se han aplicado técnicas de inteligencia artificial explicable sobre las arquitecturas propuestas, lo que ha demostrado aportar más comprensión sobre el proceso de detección de violencia. Así, se ha empleado para la extracción de *frames* característicos intrínsicamemente explicables YoloV8 y la diferencia de *frames*, si bien se ha expuesto que el mínimo número de personas que YoloV8 debe detectar y el valor mínimo de la diferencia entre *frames* para considerar la escena como potencialmente violencia se debe ajustar según las características de cada zona grabada por la cámara de vídeo. Además, el uso de GradCAM (Chamola et al., 2023) aplicado a cada fotograma del vídeo ha permitido observar cómo cada capa convolucional procesa y entiende el vídeo, así como observar si se enfoca correctamente en la zona del vídeo donde ocurre la violencia. Por último, el algoritmo desarrollado para cuantificar la importancia en

la detección de violencia de los *frames* de un vídeo, ha demostrado obtener valores superiores de relevancia en torno a los momentos violentos, permitiendo identificar la relevancia temporal de los *frames* del vídeo en la inferencia.

De este modo, las conclusiones permiten afirmar que se ha cumplido tanto con el objetivo principal, lo que permite comprobar la hipótesis planteada en el Capítulo 1, como con los objetivos específicos que han sido definidos en el mismo:

(**OB1**) *Analizar las técnicas y algoritmos empleados para la detección de violencia en vídeo mediante inteligencia artificial fiable, así como los retos actuales que enfrenta este campo.* En el Capítulo 2 se ha realizado un estudio sistemático de la literatura sobre la detección de violencia en vídeo mediante el uso de inteligencia artificial fiable, analizando en detalle todos los pasos involucrados en este proceso, así como los retos actuales que enfrenta este campo. Se han categorizado aspectos clave en los trabajos seleccionados, incluyendo 28 conjuntos de datos de agresiones físicas, 21 retos actuales en la detección mediante inteligencia artificial, y 21 métodos para la extracción de *frames* característicos, sin embargo, ninguno de los trabajos aborda el uso de inteligencia artificial fiable en el proceso de detección de violencia. Por último, entre las combinaciones de algoritmos utilizadas, la combinación de redes convolucionales (CNN) junto con capas de memoria a largo plazo (LSTM) es la más empleada y la que obtiene mejores resultados. Por ello, esta arquitectura es la más comúnmente utilizada en el campo de la detección de violencia en vídeo.

(**OB2**) *Diseñar arquitecturas de detección de violencia en vídeo mediante inteligencia artificial fiable, combinando CNN con características manuales y capas RNN, que mejore los métodos actuales y permita la detección de violencia en escenarios con condiciones cambiantes (escenario de alta visibilidad, escenario de alta densidad poblacional y escenario de baja visibilidad).* A este respecto, se ha detallado en el Capítulo 3 y en el Capítulo 4, el diseño de tres arquitecturas basadas en la combinación de CNN junto con el uso de características manuales, capas LSTM y capas Bi-LSTM. Se ha utilizado VGG-19 frente a VGG-16, al obtener mejores resultados en la detección igual de fotogramas violentos, así como en la arquitectura que combina CNN con características manuales.

(**OB3**) *Investigar el marco teórico y diseñar una arquitectura explicable para la detección de violencia en vídeo, integrando técnicas de XAI como GradCAM y desarrollando métodos específicos que evalúen y cuantifiquen la relevancia de los frames individuales dentro del vídeo.* En este sentido, en el Capítulo 5, se ha detallado el diseño e implementación de la arquitectura desarrollada para aplicar sobre éstas las técnicas de inteligencia artificial explicable. La arquitectura diseñada en el Capítulo 5 incluye el uso de: (I) un algoritmo propio diseñado para la cuantificación de la importancia de los fotogramas en la detección de violencia en vídeo, obteniendo resultados que han demostrado la efectividad del modelo. (II) YoloV8 y la diferencia de *frame* para la extracción de fotogramas relevantes (III) la aplicación de Grad-CAM sobre VGG-19 ha permitido interpretar en qué área de la imagen se centraba la red al predecir violencia, mostrando mapas de calor que iban desde contornos básicos hasta la interacción violenta.

(**OB4**) *Implementar y demostrar la idoneidad de las arquitecturas de detección de violencia comparándolas entre sí y con otras arquitecturas propuestas en el estado del arte.* Se han seleccionado tres datasets que representan diferentes condiciones de escenas violentas: 1) *Hockey Fights*, con alta visibilidad y violencia en entornos deportivos; 2) *Violent Flow*, con alta densidad poblacional y violencia en flujos dinámicos; y 3) *RWF-2000*, con baja visibilidad y situaciones más complejas. En el Capítulo 3 se diseñó una arquitectura basada en CNN con VGG-16 y VGG-19, donde esta última superó en precisión a VGG-16 entre un 2 % y 10 % según el dataset, alcanzando hasta un 96 % de *accuracy*, con un tiempo de inferencia ligeramente inferior. En el Capítulo 4 se combinaron las CNN con capas LSTM y Bi-LSTM, confirmando que Bi-LSTM mejora la precisión hasta un 4 %, Finalmente, en el Capítulo 5 se ha diseñado una arquitectura explicable para detectar violencia en vídeo, usando YoloV8, diferencia de *frames* y GradCAM. Además, se desarrolló un algoritmo para cuantificar la importancia de los *frames*, demostrando su efectividad.

6.2. Líneas futuras de investigación

Como parte final de este trabajo, se han definido cuatro líneas futuras, cada una relacionada directamente con uno de los objetivos principales planteados. Estas

propuestas buscan dar continuidad al desarrollo de sistemas fiables de detección de violencia en vídeo, ampliando tanto su capacidad técnica como su aplicabilidad en entornos reales. Las líneas futuras se detallan a continuación:

(**LI1**) *OB1 - Analizar las técnicas y algoritmos empleados para la detección de violencia en vídeo mediante el uso de inteligencia artificial fiable, así como los retos actuales a los que se enfrenta.* Si bien la combinación de redes CNN junto con RNN ha sido el objeto de estudio de esta Tesis Doctoral, y considerando los resultados prometedores obtenidos tanto en el estado del arte como en los experimentos desarrollados en este trabajo, se propone como línea futura la investigación del uso de arquitecturas basadas en Transformers, pudiendo ofrecer mejoras significativas en la capacidad de modelado de dependencias temporales y espaciales complejas.

(**LI2**) *OB2 - Diseñar arquitecturas de detección de violencia en vídeo mediante inteligencia artificial fiable, combinando CNN con características manuales y capas RNN, que mejore los métodos actuales y permita la detección de violencia en escenarios con condiciones cambiantes.* En esta Tesis Doctoral se ha analizado cómo la variación del número de neuronas en capas LSTM, Bi-LSTM y densamente conectadas no mejora significativamente la accuracy obtenida. Se propone como futura línea de investigación estudiar cómo el número de capas LSTM/Bi-LSTM y el número de capas densas puede influir positivamente en la detección de violencia en vídeo.

(**LI3**) *OB3 - Investigar el marco teórico y diseñar una arquitectura explicable para la detección de violencia en vídeo, integrando técnicas de XAI como GradCAM y desarrollando métodos específicos que evalúen y cuantifiquen la relevancia de los frames individuales dentro del vídeo.* Se ha expuesto en esta Tesis Doctoral cómo la arquitectura presentada que utiliza algoritmos de inteligencia artificial explicable aporta comprensión al proceso de detección de violencia. Por ello, se propone como línea de investigación futura la realización de un análisis exhaustivo a partir de las predicciones explicables que realiza la arquitectura desarrollada en la detección de vídeos potencialmente violentos. Éste análisis consiste en comprender si existe un patrón en el que la detección de violencia se realiza de forma errónea bajo un cierto tipo de escenas

(**LI4**) *OB4 - Implementar y demostrar la idoneidad de las arquitecturas de detección de violencia comparándolos entre sí y con otras arquitecturas propuestas en el estado del arte.* A partir de las conclusiones de esta Tesis Doctoral, se evidencia la necesidad de avanzar en la creación y adopción de métricas más precisas y objetivas para evaluar la explicabilidad de los modelos en el contexto de detección de violencia. La cuantificación de la explicabilidad es crucial para asegurar que las decisiones tomadas por los modelos de inteligencia artificial sean comprensibles y confiables tanto para la evaluación del modelo como para los usuarios finales.

Chapter 6

Conclusions and Future Work

Chapter 6

Conclusions and Future Work

The increasing use of artificial intelligence has raised significant concerns about the development of reliable algorithms, leading to the creation of numerous detailed reports aimed at defining and standardizing these terms by major organizations such as the European Commission (European Commission, 2019).

In this field, multiple proposed solutions exist to address violence in societies (Vomfell et al., 2018). Real-time violence detection is the last line of defense to protect victims. Among the technologies enabling such detection, artificial intelligence offers a direct and effective solution that can save lives and reduce reliance on continuous human surveillance (Omarov et al., 2022). In this context, multiple architectures and models have been proposed in the literature, with algorithms combining *Convolutional Neural Networks* (CNN) and *Long Short Term Memory* (LSTM) layers showing the best performance and thus being the most commonly used (Islam et al., 2021a; Negre et al., 2024b; Vijeikis et al., 2022).

In this regard, the objective of this Doctoral Thesis has been to design and implement an architecture based on the combination of CNN with handcrafted features and RNN layers, aiming to improve results obtained in previous studies and to enhance the explainability of these results.

First, an architecture based solely on Convolutional Neural Networks (CNN) and handcrafted features has been proposed, evaluating the pre-trained VGG-16 and VGG-19 models for frame-level violence detection in video. The results show that VGG-19 outperforms VGG-16 by 2% to 10% in accuracy across the three analyzed scenarios: it achieves 94% versus 92% in high visibility scenarios (dataset *Hockey Fights*), 96% versus 92% in high population density scenarios (dataset *Violent Flow*), and 76% versus 64% in low visibility scenarios (dataset *RWF-2000*). Although VGG-19 slightly reduces

the average inference time (1.56 ms vs. 1.64 ms), its higher accuracy justifies its use. These findings confirm that even without temporal modeling, VGG-19 offers significant advantages over VGG-16, justifying its selection for the subsequent chapters.

Second, the architecture was expanded by integrating recurrent networks (LSTM and Bi-LSTM) over the feature extraction from VGG-19 to capture temporal dynamics in video sequences. The LSTM and Bi-LSTM models increase accuracy up to 96% and 97% in high visibility scenarios, compared to 94% / 95% for non-recurrent versions; in high population density scenarios, they reach 96% (LSTM) and 90% (Bi-LSTM), compared to 96% without RNN; and in low visibility scenarios they achieve 72% and 73% versus 76% with handcrafted features. The weighted average of Bi-LSTM (82%) surpasses LSTM (80%) by two points, demonstrating that temporal analysis adds value in good visibility contexts, although its effectiveness depends on the scenario.

Third, the architecture was completed by incorporating explainability mechanisms acting in parallel on the temporal, spatial, and frame selection dimensions, aiming to improve model transparency without affecting its performance. The temporal importance metric identifies critical segments with a relationship I_i ranging from 10^{-3} to 10^{-6}, showing robustness against information suppression; the frame selection strategy combines people counting with motion difference to reduce computational load while maintaining accuracy and real-time operability; and the application of Grad-CAM on five layers of VGG-19 shows progressive attention from simple contours to violent interaction hotspots. The joint integration of these tools increases trust in predictions and offers operational traceability, which is key in critical video surveillance contexts.

Based on this, another objective of this Doctoral Thesis is to implement explainable artificial intelligence techniques on the proposed architectures to make the entire violence detection process more reliable. Accordingly, intrinsically explainable techniques such as YoloV8 (Diwan et al., 2023) have been implemented for characteristic frame extraction, GradCAM (Chamola et al., 2023; Mahmoudi et al., 2023) has been used as an explainability algorithm focused on CNN for saliency map generation, and an algorithm has been developed to quantify the relevance of video segments in violence detection.

This Chapter therefore concludes this Doctoral Thesis and presents possible future research directions. Section 6.1 lists the conclusions drawn from the research carried out in this Doctoral Thesis, as well as their alignment with the objectives stated at the

beginning. Section 6.2 proposes different possible future research lines that arise from the results and conclusions of this Doctoral Thesis.

6.1. Conclusions

Various options have been suggested in the literature to address violence in communities (Killgore et al., 2021; Vomfell et al., 2018). Real-time detection of violent acts represents a crucial measure to safeguard victims. In this context, artificial intelligence has stood out for its effectiveness, significantly saving human surveillance resources and enabling the simultaneous analysis of multiple videos (Jing et al., 2021).

This Doctoral Thesis has conducted a systematic study of recent literature focused on video violence detection using artificial intelligence, with a focus on trustworthy AI (Negre et al., 2024b), compiling a total of 63 articles between June 2021 and June 2023. This study, developed in Chapter 2, gathered and categorized 28 violence detection datasets, highlighting the creation of new datasets in recent years (Negre et al., 2024a).

Additionally, 13 evaluation parameters used in violence detection have been compiled, with *accuracy* being the most common. A total of 21 key feature extraction methods were analyzed and categorized, which are essential for more lightweight and cost-effective real-time processing of video violence detection.
In the analysis of different algorithm inputs, color video is the most used, and the combination of algorithms with the best results—and thus most used—is CNN and LSTM (Negre et al., 2024a).

However, among the articles selected in the systematic mapping, none explicitly use trustworthy AI techniques. Although a total of 15 articles mention model robustness, this concept is focused on the number of false positives and false negatives, rather than robustness as a pillar of trustworthiness. On the other hand, several articles use techniques or algorithms in the violence detection process that may facilitate model explainability, such as object detection algorithms or skeleton-based algorithms (Hung et al., 2021; Naik & Gopalakrishna, 2021; Narynov et al., 2021; Zhou, 2022).

Based on the results obtained in this systematic study, several architectures have been developed in this Doctoral Thesis. Chapter 3 proposes an architecture based on the

combination of CNN with handcrafted features, selecting VGG-16 and VGG-19 networks pretrained on *ImageNet* as candidate CNNs. Chapter 4 proposes an architecture based on the combination of pretrained VGG-19 with LSTM and Bi-LSTM layers.

- **Conclusion 1. Chapter 2**. A systematic review of recent literature on video violence detection using trustworthy artificial intelligence has been carried out, collecting 63 articles over the past two years. 28 datasets were categorized, reflecting the continuous creation of new datasets, and 13 evaluation parameters were identified, with *accuracy* being the most used. Additionally, 21 key feature extraction methods were analyzed, which are essential for efficient real-time processing. Algorithms combining convolutional networks (CNN) with LSTM layers are the most frequent and show better performance compared to other architectures. Finally, although no study explicitly applies complete trustworthy AI techniques, several integrate methods that support explainability. These results support the choice of combining CNN and LSTM as the basis for the architectures developed in this Thesis.

- **Conclusion 2. Chapter 3**. An architecture based on convolutional neural networks (CNN) complemented with handcrafted features has been designed and implemented, evaluating VGG-16 and VGG-19 networks for frame-level violence detection. The results show that VGG-19 surpasses VGG-16 in accuracy by 2% to 10% in the three analyzed scenarios: 94% vs. 92% in high visibility scenarios (*Hockey Fights*), 96% vs. 92% in high population density scenarios (*Violent Flow*), and 76% vs. 64% in low visibility conditions (*RWF-2000*). Also, although VGG-19 has a slightly lower average inference time (1.56 ms vs. 1.64 ms), its greater precision fully justifies its selection. Overall, these results confirm that even without temporal modeling, VGG-19 offers notable advantages over VGG-16, thus solidifying its selection for the following chapters.

- **Conclusion 3. Chapter 4**. An architecture based on the pretrained CNN VGG-19 combined with LSTM and Bi-LSTM layers has been designed and implemented. It has been shown that using Bi-LSTM layers outperforms the accuracy achieved by LSTM layers, although this improvement did not exceed 4% in the conducted experiments (Section 4.4.2). Furthermore, a lack of correlation was observed between hyperparameter values—such as learning rate and number

of neurons in LSTM, Bi-LSTM, or dense layers—and the improvement in achieved accuracy (Section 4.4.2). Finally, it was observed that while VGG-19 performed better than VGG-16 in the architectures proposed in Chapter 3, this did not hold when combined with LSTM or Bi-LSTM layers, as the spatial features extracted by VGG-16 yielded better results when integrated with these layers.

- **Conclusion 4. Chapter 5**. Explainable artificial intelligence techniques have been applied to the proposed architectures, which proved to bring greater understanding to the violence detection process. For characteristic frame extraction, intrinsically explainable YoloV8 and frame difference were used, although it was shown that the minimum number of people YoloV8 must detect and the minimum frame difference value to consider a scene as potentially violent must be adjusted based on the characteristics of each area captured by the video camera. Furthermore, the use of GradCAM (Chamola et al., 2023) applied to each video frame allowed the observation of how each convolutional layer processes and understands the video, and whether it correctly focuses on the video area where violence occurs. Lastly, the developed algorithm to quantify the importance of video frames in violence detection demonstrated higher relevance values around violent moments, enabling the identification of the temporal importance of the video frames during inference.

Thus, the conclusions allow us to state that both the main objective and the specific objectives defined in Chapter 1 have been achieved:

(**OB1**) *Analyze the techniques and algorithms used for video violence detection using trustworthy artificial intelligence, as well as the current challenges in this field.* Chapter 2 presents a systematic review of the literature on video violence detection using trustworthy AI, analyzing in detail all the steps involved in this process and the current challenges. Key aspects in the selected works were categorized, including 28 datasets of physical aggression, 21 current challenges in AI-based detection, and 21 methods for characteristic frame extraction. However, none of the works address the use of trustworthy AI in the violence detection process. Finally, among the algorithm combinations used, the combination of convolutional neural networks (CNN) with long short-term

memory (LSTM) layers is the most employed and the one with the best performance. Therefore, this architecture is the most commonly used in the field of video violence detection.

(**OB2**) *Design violence detection architectures using trustworthy artificial intelligence, combining CNN with handcrafted features and RNN layers, to improve current methods and enable violence detection in changing conditions (high visibility, high population density, and low visibility scenarios).* In this regard, Chapters 3 and 4 detail the design of three architectures based on the combination of CNN with handcrafted features, LSTM layers, and Bi-LSTM layers. VGG-19 was used instead of VGG-16, as it obtained better results in detecting violent frames, as well as in the architecture combining CNN with handcrafted features.

(**OB3**) *Research the theoretical framework and design an explainable architecture for violence detection in video, integrating XAI techniques such as GradCAM and developing specific methods to assess and quantify the relevance of individual frames within the video.* In this sense, Chapter 5 details the design and implementation of the architecture developed to apply explainable AI techniques. The architecture designed in Chapter 5 includes: (IV) a custom algorithm to quantify the importance of frames in video violence detection, obtaining results that demonstrate the model's effectiveness; (V) YoloV8 and frame difference for relevant frame extraction; (VI) the application of Grad-CAM over VGG-19, which allowed interpretation of the image areas the network focused on when predicting violence, showing heatmaps ranging from basic contours to violent interaction.

(**OB4**) *Implement and demonstrate the suitability of violence detection architectures by comparing them with each other and with other state-of-the-art architectures.* Three datasets representing different violent scene conditions were selected: 1) *Hockey Fights*, with high visibility and violence in sports environments; 2) *Violent Flow*, with high population density and violence in dynamic flows; and 3) *RWF-2000*, with low visibility and more complex situations. In Chapter 3, an architecture based on CNN with VGG-16 and VGG-19 was designed, where the latter outperformed VGG-16 in accuracy by 2% to 10% depending on the dataset, reaching up to 96% *accuracy*, with a slightly lower inference time. In

Chapter 4, CNNs were combined with LSTM and Bi-LSTM layers, confirming that Bi-LSTM improves accuracy by up to 4%. Finally, in Chapter 5, an explainable architecture was designed to detect violence in video, using YoloV8, frame difference, and GradCAM. Additionally, an algorithm was developed to quantify frame importance, demonstrating its effectiveness.

6.2. Future Research Directions

As the final part of this work, four future research directions have been defined, each directly related to one of the main objectives set. These proposals aim to continue the development of reliable video violence detection systems, expanding both their technical capacity and applicability in real-world environments. The future directions are detailed below:

(**FR1**) *OB1 - Analyze the techniques and algorithms used for video violence detection through the use of trustworthy artificial intelligence, as well as the current challenges faced.* While the combination of CNN networks along with RNNs has been the focus of this Doctoral Thesis, and considering the promising results obtained both in the state of the art and in the experiments developed in this work, a future line of research is proposed to investigate the use of Transformer-based architectures, which could offer significant improvements in modeling complex temporal and spatial dependencies.

(**FR2**) *OB2 - Design video violence detection architectures using trustworthy artificial intelligence, combining CNN with handcrafted features and RNN layers, to improve current methods and allow detection in scenarios with changing conditions.* This Doctoral Thesis analyzed how varying the number of neurons in LSTM, Bi-LSTM, and densely connected layers does not significantly improve the accuracy obtained. A future research line is proposed to study how the number of LSTM/Bi-LSTM layers and the number of dense layers could positively influence video violence detection.

(**FR3**) *OB3 - Investigate the theoretical framework and design an explainable architecture for video violence detection, integrating XAI techniques such as*

GradCAM and developing specific methods that evaluate and quantify the relevance of individual frames within the video. This Doctoral Thesis has shown how the presented architecture using explainable artificial intelligence algorithms provides understanding to the violence detection process. Therefore, a future research line is proposed to carry out an exhaustive analysis based on the explainable predictions made by the developed architecture in the detection of potentially violent videos. This analysis consists of understanding whether there is a pattern in which violence detection is erroneously made under certain types of scenes.

(**FR4**) *OB4 - Implement and demonstrate the suitability of violence detection architectures by comparing them with each other and with other architectures proposed in the state of the art.* Based on the conclusions of this Doctoral Thesis, the need to advance in the creation and adoption of more precise and objective metrics to evaluate model explainability in the context of violence detection is evident. Quantifying explainability is crucial to ensure that the decisions made by artificial intelligence models are understandable and trustworthy both for model evaluation and for end users.

Apéndice A

Evidencias y Resultados

Apéndice A

Evidencias y Resultados

En este Capítulo se presentan las publicaciones en revistas científicas y congresos internacionales que contribuyen al desarrollo y resultados de esta Tesis Doctoral. En este sentido, la estructura de este Capítulo es la siguiente. La Sección A.1 detalla las publicaciones realizadas en revistas científicas y conferencias internacionales. La Sección A.2 describe las estancias llevadas a cabo en organismos de investigación internacionales relacionadas con las líneas de investigación de esta Tesis Doctoral.

A.1. Publicaciones

En primer lugar, se detallan las diferentes publicaciones desarolladas en revistas científicas internacionales o repositorios abiertos, así como congresos internacionales. Asimismo, se refleja en azul el factor de impacto[1] y el cuartil al que pertenece cada revista internacional según el año de publicación o, en su defecto, el último factor de impacto disponible antes de la publicación del artículo o, en caso de que la revista no contara aún con factor de impacto en la fecha de publicación, el factor de impacto del primer año calculado para la revista tras la publicación del artículo.

A.1.1. Publicaciones en revistas científicas internacionales y repositorios abiertos

Esta Sección se divide en dos subapartados. El subapartado *a* contiene las publicaciones en revistas internacionales y repositorios abiertos relacionadas con el tema de esta Tesis Doctoral. El subapartado *b* contiene las publicaciones en revistas internacionales y repositorios abiertos no relacionadas con el tema de esta Tesis Doctoral.

[1] Factor de impacto según el JCR – *Journal Citation Reports*.

a) Publicaciones relacionadas con el tema de esta Tesis Doctoral

1. Negre, P., Alonso, R. S., González-Briones, A., Prieto, J., & Rodríguez-González, S. (2024). Literature Review of Deep-Learning-Based Detection of Violence in Video. *Sensors, 24* (12), 4016. https://doi.org/10.3390/s24124016. [JCR 3.4 – Q2 (*2023*)]

2. Negre, P., Alonso, R. S., Prieto, J., Dang, C. N., & Corchado, J. M. (2024). Systematic Mapping Study on Violence Detection in vídeo by Means of Trustworthy Artificial Intelligence. *SSRN*. https://dx.doi.org/10.2139/ssrn.4757631.

3. Negre, P., Alonso, R. S., Prieto, J., Novais, P., & Corchado, J. M. (2024). Violence Detection in Video Models Implementation Using Pre-trained VGG19 Combined With Manual Logic, LSTM Layers and Bi-LSTM layers. LSTM Layers and Bi-LSTM layers. *SSRN*. https://dx.doi.org/10.2139/ssrn.4832475.

4. Negre, P., Prieto, J., Alonso, R. S., Chamoso, P., & Corchado, J. M. (2024). eXplainable Artificial Intelligence Combining CNN and RNN Layers for Violence Detection in Video. *SSRN*

b) Publicaciones no relacionadas con el tema de la Tesis Doctoral

1. Negre, P., Alonso, R. S., Prieto, J., García, Ó., de-la-Fuente-Valentín, L. (2024). Prediction of footwear demand using Prophet and SARIMA. *Expert Systems with Applications*, 124512. https://doi.org/10.1016/j.eswa.2024.124512 [JCR 7.5 – Q1 (*2023*)]

A.1.2. Publicaciones en congresos internacionales

Esta Sección se divide en dos subapartados. El subapartado *a* contiene las publicaciones en congresos internacionales relacionadas con el tema de esta Tesis Doctoral. El subapartado *b* contiene las publicaciones en congresos internacionales no relacionadas con el tema de esta Tesis Doctoral, publicadas durante su desarrollo.

a) Publicaciones relacionadas con el tema de esta Tesis Doctoral

1. Negre, P., Alonso, R. S., Prieto, J., Arrieta, A. G., & Corchado, J. M. (2023, julio). Review of Physical Aggression Detection Techniques in vídeo Using Explainable Artificial Intelligence. In *International Symposium on Ambient Intelligence* (pp. 53-62). Cham: Springer Nature Switzerland.

2. Negre, P., Alonso, R. S., Prieto, J., Novais, P., & Corchado, J. M. (2024, junio). Integrating Pretrained VGG19 and Bi-LSTM for Violence Detection in video.

b) Publicaciones no relacionadas con el tema de la Tesis Doctoral

1. Negre, P., Alonso, R. S., Prieto, J., García, Ó., & de-la-Fuente-Valentín, L. (2024, junio). KPIs for the evaluation of footwear supply forecasting using Prophet.

A.2. Estancias internacionales

Finalmente, se detallan las dos estancias internacionales llevadas a cabo fuera de España en dos instituciones de enseñanza superior de prestigio, donde se realizan trabajos de investigación relacionados con esta Tesis Doctoral.

- Centro ALGORITMI, Universidade do Minho.
  - Ciudad de la entidad: Braga, Portugal.
  - Fecha de inicio – Fecha de fin: 04/03/2024 – 04/06/2024
  - Tareas realizadas: Implementación y análisis de las técnicas de inteligencia artificial explicables fijadas como objetivos de esta Tesis Doctoral
  - Publicaciones desarrolladas:
    - Negre, P., Alonso, R. S., Prieto, J., Novais, P., & Corchado, J. M. (2024). Violence Detection in Video Models Implementation Using Pre-trained VGG19 Combined With Manual Logic, LSTM Layers and Bi-LSTM layers. *SSRN* (2024, mayo).
    - Negre, P., Alonso, R. S., Prieto, J., Novais, P., & Corchado, J. M. (2024, junio). Integrating Pretrained VGG19 and Bi-LSTM for Violence Detection in video.

Como resultado de dicha estancia se originan publicaciones en conjunto con los investigadores del Centro ALGORITMI, como el paper presentado al congreso internacional DCAI 2024 expuesto en la Sección A.1.2, como parte de la Tesis Doctoral.

– Ho Chi Minh University of Transport.

- Ciudad de la entidad: Ho Chi Minh, Vietnam.
- Fecha de inicio – Fecha de fin: 04/07/2023 – 15/09/2023
- Tareas realizadas: Desarrollo del estudio de mapeo sistemático de artículos que tratan la detección de violencia en vídeo mediante el uso de inteligencia artificial que forma parte de esta Tesis Doctoral.
- Publicaciones desarrolladas:
  - Negre, P., Alonso, R. S., Prieto, J., Dang, C. N., Corchado, J. M. (2024). Systematic Mapping Study on Violence Detection in Video by Means of Trustworthy Artificial Intelligence. *SSRN 4757631*.

Como resultado de dicha estancia se originan publicaciones en conjunto con los investigadores de la *Ho Chi Minh University of Transport*, como el paper preprinteado a la espera de ser aceptado titulado *Systematic Mapping Study on Violence Detection in vídeo by Means of Trustworthy Artificial Intelligence*, como parte del estudio sistemático realizado en esta Tesis Doctoral.

Bibliografía

Aarthy, K., & Nithya, A. A. (2022). Crowd Violence Detection in Videos Using Deep Learning Architecture. *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, 1-6.

Abhishek, K., & Kamath, D. (2022). Attribution-based XAI methods in computer vision: A review. *arXiv preprint arXiv:2211.14736*.

Adithya, H., Lekhashree, H., & Raghuram, S. (2023). Violence Detection in Drone Surveillance Videos. *International Conference on Smart Computing and Communication*, 703-713.

Afra, S., & Alhajj, R. (2020). Early warning system: From face recognition by surveillance cameras to social media analysis to detecting suspicious people. *Physica A: Statistical Mechanics and its Applications*, *540*, 123151.

Ageed, Z., & Zeebaree, S. (2021). A Comprehensive Survey of Big Data Mining Approaches in Cloud Systems. *1*, 29-38.

Ahmed, M., Ramzan, M., Khan, H. U., Iqbal, S., Khan, M. A., Choi, J.-I., Nam, Y., & Kadry, S. (2021). Real-time violent action recognition using key frames extraction and deep learning.

Aktı, Ş., Ofli, F., Imran, M., & Ekenel, H. K. (2022). Fight detection from still images in the wild. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 550-559.

Aktı, Ş., Tataroğlu, G. A., & Ekenel, H. K. (2019). Vision-based fight detection from surveillance cameras. *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 1-6.

Ali, O., Shrestha, A., Soar, J., & Wamba, S. F. (2018). Cloud computing-enabled healthcare opportunities, issues, and applications: A systematic review. *International Journal of Information Management*, *43*, 146-158.

Alonso, R. S., Sittón-Candanedo, I., Casado-Vara, R., Prieto, J., & Corchado, J. M. (2020). Deep reinforcement learning for the management of software-defined networks and network function virtualization in an edge-IoT architecture. *Sustainability*, *12*(14), 5706.

Appavu, N., et al. (2023). Violence Detection Based on Multisource Deep CNN with Handcraft Features. *2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*, 1-6.

Asad, M., Yang, J., He, J., Shamsolmoali, P., & He, X. (2021). Multi-frame feature-fusion-based model for violence detection. *The Visual Computer*, *37*, 1415-1431.

Baskerville, R. L. (1999). Investigating information systems with action research. *Communications of the Association for Information Systems*, *2*(1), 19.

Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., & Sukthankar, R. (2011). Violence detection in video using computer vision techniques. *Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14*, 332-339.

Bi, Y., Li, D., & Luo, Y. (2022). Combining keyframes and image classification for violent behavior recognition. *Applied Sciences*, *12*(16), 8014.

Blunsden, S., & Fisher, R. (2010). The BEHAVE video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, *4*(1-12), 4.

Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A review of trustworthy and explainable artificial intelligence (xai). *IEEe Access*.

Chen, Y., Zhang, B., & Liu, Y. (2021). ESTN: Exacter Spatiotemporal Networks for Violent Action Recognition. *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*, 44-48.

Cheng, M., Cai, K., & Li, M. (2021a). RWF-2000: An Open Large Scale Video Database for Violence Detection. *2020 25th International Conference on Pattern Recognition (ICPR)*, 4183-4190.

Cheng, M., Cai, K., & Li, M. (2021b). RWF-2000: an open large scale video database for violence detection. *2020 25th International Conference on Pattern Recognition (ICPR)*, 4183-4190.

Cheng, S.-T., Hsu, C.-W., Horng, G.-J., & Jiang, C.-R. (2021c). Video reasoning for conflict events through feature extraction. *The Journal of Supercomputing*, *77*, 6435-6455.

Collins, C., Dennehy, D. D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, *60*, 102383.

Contardo, P., Tomassini, S., Falcionelli, N., Dragoni, A. F., & Sernani, P. (2023). Combining a mobile deep neural network and a recurrent layer for violence detection in videos.

Demarty, C.-H., Penet, C., Soleymani, M., & Gravier, G. (2015). VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation. *Multimedia Tools and Applications*, *74*, 7379-7404.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248-255.

Ding, D., Ma, Z., Chen, D., Chen, Q., Liu, Z., & Zhu, F. (2021). Advances in video compression system using deep neural network: A review and case studies. *Proceedings of the IEEE*, *109*(9), 1494-1520.

Diwan, T., Anirudh, G., & Tembhurne, J. V. (2023). Object detection using YOLO: Challenges, architectural successors, datasets and applications. *multimedia Tools and Applications*, *82*(6), 9243-9275.

Eden, C., & Ackermann, F. (2018). Theory into practice, practice to theory: Action research in method development. *European Journal of Operational Research*, *271*(3), 1145-1155.

Ehsan, T. Z., & Mohtavipour, S. M. (2020). Vi-Net: a deep violent flow network for violence detection in video sequences. *2020 11th International Conference on Information and Knowledge Technology (IKT)*, 88-92.

Ehsan, T. Z., Nahvi, M., & Mohtavipour, S. M. (2023). An accurate violence detection framework using unsupervised spatial–temporal action translation network. *The Visual Computer*, 1-21.

European Commission. (2019). Ethics Guidelines for Trustworthy AI.

Fekih-Romdhane, F., Malaeb, D., Sarray El Dine, A., Obeid, S., & Hallit, S. (2022). The relationship between smartphone addiction and aggression among Lebanese adolescents: the indirect effect of cognitive function. *BMC pediatrics*, *22*(1), 735.

FRA, E. (2023, febrero). Crime, safety and victims' rights. Fundamental Rights Survey 2023.

Freire-Obregón, D., Barra, P., Castrillón-Santana, M., & Marsico, M. D. (2022). Inflated 3D ConvNet context analysis for violence detection. *Machine Vision and Applications*, *33*, 1-13.

Gkountakos, K., Ioannidis, K., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2020). A crowd analysis framework for detecting violence scenes. *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 276-280.

Gupta, H., & Ali, S. T. (2022). Violence Detection using Deep Learning Techniques. *2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, 121-124.

Halder, R., & Chatterjee, R. (2020). CNN-BiLSTM model for violence detection in smart surveillance. *SN Computer science*, *1*(4), 201.

Hassner, T., Itcher, Y., & Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, 1-6.

Hillis, S., Mercy, J., Amobi, A., & Kress, H. (2016). Global prevalence of past-year violence against children: a systematic review and minimum estimates. *Pediatrics*, *137*(3).

Hipp, J. R., Lee, S., Ki, D., & Kim, J. H. (2021). Measuring the built environment with google street view and machine learning: Consequences for crime on street segments. *Journal of Quantitative Criminology*, 1-29.

Hu, X., Fan, Z., Jiang, L., Xu, J., Li, G., Chen, W., Zeng, X., Yang, G., & Zhang, D. (2022). TOP-ALCM: A novel video analysis method for violence detection in crowded scenes. *Information Sciences*, *606*, 313-327.

Hung, L.-P., Yang, C.-W., Lee, L.-H., & Chen, C.-L. (2021). Constructing a Violence Recognition Technique for Elderly Patients with Lower Limb Disability. *International Conference on Smart Grid and Internet of Things*, 24-37.

Huszár, V. D., Adhikarla, V. K., Négyesi, I., & Krasznay, C. (2023). Toward Fast and Accurate Violence Detection for Automated Video Surveillance Applications. *IEEE Access*, *11*, 18772-18793.

Islam, M. S., Hasan, M. M., Abdullah, S., Akbar, J. U. M., Arafat, N., & Murad, S. A. (2021a). A deep Spatio-temporal network for vision-based sexual harassment detection. *2021 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, 1-6.

Islam, Z., Rukonuzzaman, M., Ahmed, R., Kabir, M. H., & Farazi, M. (2021b). Efficient two-stream network for violence detection using separable convolutional lstm. *2021 International Joint Conference on Neural Networks (IJCNN)*, 1-8.

ISO/IEC. (2020). *Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence* (Technical Report N.º ISO/IEC TR 24028:2020).

Jaafar, N., & Lachiri, Z. (2023). Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. *Expert Systems with Applications*, *211*, 118523.

Jahlan, H. M. B., & Elrefaei, L. A. (2021). Mobile neural architecture search network and convolutional long short-term memory-based deep features toward detecting violence from video. *Arabian Journal for Science and Engineering*, *46*(9), 8549-8563.

Jain, A., & Vishwakarma, D. K. (2020). Deep NeuralNet for violence detection using motion features from dynamic images. *2020 third international conference on smart systems and inventive technology (ICSSIT)*, 826-831.

Jaiswal, S. G., & Mohod, S. W. (2021). CLASSIFICATION OF VIOLENT VIDEOS USING ENSEMBLE BOOSTING MACHINE LEARNING APPROACH WITH LOW LEVEL FEATURES.

Jayasimhan, A., & Pabitha, P. (2022). A hybrid model using 2D and 3D Convolutional Neural Networks for violence detection in a video dataset. *2022 3rd International Conference on Communication, Computing and Industry 4.0 (C2I4)*, 1-5.

Ji, Y., Wang, Y., Kato, J., & Mori, K. (2020). Predicting Violence Rating Based on Pairwise Comparison. *IEICE TRANSACTIONS on Information and Systems*, *103*(12), 2578-2589.

Jing, F., Liu, L., Zhou, S., Song, J., Wang, L., Zhou, H., Wang, Y., & Ma, R. (2021). Assessing the impact of street-view greenery on fear of neighborhood crime in Guangzhou, China. *International journal of environmental research and public health*, *18*(1), 311.

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys (CSUR)*, *55*, 1-38.

Kaur, G., & Singh, S. (2022). Violence detection in videos using deep learning: A survey. *Advances in Information Communication Technology and Computing: Proceedings of AICTC 2021*, 165-173.

Killgore, W. D., Cloonan, S. A., Taylor, E. C., Anlap, I., & Dailey, N. S. (2021). Increasing aggression during the COVID-19 lockdowns. *Journal of Affective Disorders Reports*, *5*, 100163.

Kim, H., Jeon, H., Kim, D., & Kim, J. (2022). Lightweight framework for the violence and falling-down event occurrence detection for surveillance videos. *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, 1629-1634.

Kitchenham, B. A., Budgen, D., & Brereton, O. P. (2011). Using mapping studies as the basis for further research–a participant-observer case study. *Information and Software Technology*, *53*(6), 638-651.

Kostka, G., Steinacker, L., & Meckel, M. (2020). Between privacy and convenience: Facial recognition technology in the eyes of citizens in China, Germany, the UK and the US. *Germany, the UK and the US (February 10, 2020)*.

Kumar, A., Shetty, A., Sagar, A., Charushree, A., & Kanwal, P. (2023). Indoor Violence Detection using Lightweight Transformer Model. *2023 4th International Conference for Emerging Technology (INCET)*, 1-6.

Li, A., Thotakuri, M., Ross, D. A., Carreira, J., Vostrikov, A., & Zisserman, A. (2020). The AVA-Kinetics Localized Human Actions Video Dataset.

Liang, Q., Cheng, C., Li, Y., Yang, K., & Chen, B. (2021). Fusion and visualization design of violence detection and geographic video. *Theoretical Computer Science: 39th National Conference of Theoretical Computer Science, NCTCS 2021, Yinchuan, China, July 23–25, 2021, Revised Selected Papers 39*, 33-46.

Lohithashva, B., & Aradhya, V. M. (2021). Violent video event detection: a local optimal oriented pattern based approach. *Applied Intelligence and Informatics: First International Conference, AII 2021, Nottingham, UK, July 30–31, 2021, Proceedings 1*, 268-280.

Long, D., Liu, L., Xu, M., Feng, J., Chen, J., & He, L. (2021). Ambient population and surveillance cameras: The guardianship role in street robbers' crime location choice. *Cities*, *115*, 103223.

López-Blanco, R., Alonso, R. S., Rodríguez-González, S., Prieto, J., & Corchado, J. M. (2024). Trustworthy Artificial Intelligence-based federated architecture for symptomatic disease detection. *Neurocomputing*, 127415.

Madhavan, R., Utkarsh & Vidhya, J. (2021). Violence Detection from CCTV Footage Using Optical Flow and Deep Learning in Inconsistent Weather and Lighting

Conditions. *Advances in Computing and Data Sciences: 5th International Conference, ICACDS 2021, Nashik, India, April 23–24, 2021, Revised Selected Papers, Part I 5*, 638-647.

Magdy, M., Fakhr, M. W., & Maghraby, F. A. (2023). Violence 4D: Violence detection in surveillance using 4D convolutional neural networks. *IET Computer Vision*.

Mahalle, M. D., & Rojatkar, D. V. (2021a). Audio Based Violent Scene Detection Using Extreme Learning Machine Algorithm. *2021 6th International Conference for Convergence in Technology (I2CT)*, 1-8.

Mahalle, M. D., & Rojatkar, D. V. (2021b). Audio based violent scene detection using extreme learning machine algorithm. *2021 6th international conference for convergence in technology (I2CT)*, 1-8.

Mahmoodi, J., Nezamabadi-pour, H., & Abbasi-Moghadam, D. (2022). Violence detection in videos using interest frame extraction and 3D convolutional neural network. *Multimedia tools and applications*, *81*(15), 20945-20961.

Mahmoudi, S. A., Amel, O., Stassin, S., Liagre, M., Benkedadra, M., & Mancas, M. (2023). A Review and Comparative Study of Explainable Deep Learning Models Applied on Action Recognition in Real Time. *Electronics*, *12*(9), 2027.

Martínez-González, M. B., Turizo-Palencia, Y., Arenas-Rivera, C., Acuña-Rodríguez, M., Gómez-López, Y., & Clemente-Suárez, V. J. (2021). Gender, Anxiety, and Legitimation of Violence in Adolescents Facing Simulated Physical Aggression at School. *Brain Sciences*, *11*(4).

Mohtavipour, S. M., Saeidi, M., & Arabsorkhi, A. (2022). A multi-stream CNN for deep violence detection in video sequences using handcrafted features. *The Visual Computer*, 1-16.

Monteiro, C., & Durães, D. (2022). Modelling a Framework to Obtain Violence Detection with Spatial-Temporal Action Localization. *World Conference on Information Systems and Technologies*, 630-639.

Muarifah, A., Mashar, R., Hashim, I. H. M., Rofiah, N. H., & Oktaviani, F. (2022). Aggression in adolescents: the role of mother-child attachment and self-esteem. *Behavioral Sciences*, *12*(5).

Mugunga, I., Dong, J., Rigall, E., Guo, S., Madessa, A., Hirpa, N., & Hafiza, S. (2021a). Un modelo de funciones basado en fotogramas para la detección de la violencia desde cámaras de vigilancia que utilizan la red ConvLSTM. *2021*

Sexta Conferencia Internacional sobre Imagen, Visión y Computación (ICIVC), 55-60.

Mugunga, I., Dong, J., Rigall, E., Guo, S., Madessa, A. H., & Nawaz, H. S. (2021b). A frame-based feature model for violence detection from surveillance cameras using ConvLSTM network. *2021 6th International Conference on Image, Vision and Computing (ICIVC)*, 55-60.

Mumtaz, A., Bux Sargano, A., & Habib, Z. (2022a). Fast learning through deep multi-net CNN model for violence recognition in video surveillance. *The Computer Journal*, *65*(3), 457-472.

Mumtaz, N., Ejaz, N., Aladhadh, S., Habib, S., & Lee, M. Y. (2022b). Deep multi-scale features fusion for effective violence detection and control charts visualization. *Sensors*, *22*(23), 9383.

Nadeem, M. S., Kurugollu, F., Saravi, S., Atlam, H. F., & Franqueira, V. N. (2023). Deep labeller: automatic bounding box generation for synthetic violence detection datasets. *Multimedia Tools and Applications*, 1-18.

Naik, A. J., & Gopalakrishna, M. (2021). Deep-violence: individual person violent activity detection in video. *Multimedia Tools and Applications*, *80*(12), 18365-18380.

Narynov, S., Zhumanov, Z., Gumar, A., Khassanova, M., & Omarov, B. (2021). Detecting School Violence Using Artificial Intelligence to Interpret Surveillance Video Sequences. *Advances in Computational Collective Intelligence: 13th International Conference, ICCCI 2021, Kallithea, Rhodes, Greece, September 29–October 1, 2021, Proceedings 13*, 401-412.

Negre, P., Alonso, R. S., González-Briones, A., Prieto, J., & Rodríguez-González, S. (2024a). Literature Review of Deep-Learning-Based Detection of Violence in Video. *Sensors*, *24*(12), 4016.

Negre, P., Alonso, R. S., Prieto, J., Dang, C. N., & Corchado, J. M. (2024b). Systematic Mapping Study on Violence Detection in Video by Means of Trustworthy Artificial Intelligence. *Available at SSRN 4757631*.

Negre, P., Alonso, R. S., Prieto, J., Novais, P., & Corchado, J. M. (2024c). Violence Detection in Video Models Implementation Using Pre-trained VGG19 Combined With Manual Logic, LSTM Layers and Bi-LSTM layers. *LSTM Layers and Bi-LSTM layers (May 18, 2024)*.

Omarov, B., Narynov, S., Zhumanov, Z., Gumar, A., & Khassanova, M. (2022). State-of-the-art violence detection techniques in video surveillance security systems: a systematic review. *PeerJ Computer Science*, *8*, e920.

Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008). Systematic mapping studies in software engineering. EASE'08 Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, 68–77.

Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and software technology*, *64*, 1-18.

Qu, W., Zhu, T., Liu, J., & Li, J. (2022). A time sequence location method of long video violence based on improved C3D network. *The Journal of Supercomputing*, *78*(18), 19545-19565.

Rachna, U., Guruprasad, V., Shindhe, S. D., & Omkar, S. (2022). Real-Time Violence Detection Using Deep Neural Networks and DTW. *International Conference on Computer Vision and Image Processing*, 316-327.

Sahatova, K., & Balabaeva, K. (2022). An overview and comparison of XAI methods for object detection in computer tomography. *Procedia Computer Science*, *212*, 209-219.

Santos, F., Durães, D., Marcondes, F. S., Lange, S., Machado, J., & Novais, P. (2021). Efficient violence detection using transfer learning. *International Conference on Practical Applications of Agents and Multi-Agent Systems*, 65-75.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local SVM approach. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, *3*, 32-36.

Sernani, P., Falcionelli, N., Tomassini, S., Contardo, P., & Dragoni, A. F. (2021a). Deep Learning for Automatic Violence Detection: Tests on the AIRTLab Dataset. *IEEE Access*, *9*, 160580-160595.

Sernani, P., Falcionelli, N., Tomassini, S., Contardo, P., & Dragoni, A. F. (2021b). Deep learning for automatic violence detection: Tests on the AIRTLab dataset. *IEEE Access*, *9*, 160580-160595.

Shang, Y., Wu, X., & Liu, R. (2022). Multimodal Violent Video Recognition Based on Mutual Distillation. *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 623-637.

Sharma, S., Sudharsan, B., Naraharisetti, S., Trehan, V., & Jayavel, K. (2021). A fully integrated violence detection system using CNN and LSTM. *International Journal of Electrical & Computer Engineering (2088-8708)*, *11*(4).

Shubber, M. S. M., & Al-Ta'i, Z. T. M. (2022). A review on video violence detection approaches. *International Journal of Nonlinear Analysis and Applications*, *13*(2), 1117-1130.

Shukla, H., & Pandey, M. (2020). Human Suspicious Activity Recognition. *IIRJET*, *5*(4).

Siddique, M., Islam, M. S., Sinthy, R., Mohima, K., Kabir, M., Jibon, A. H., & Biswas, M. (2022). State-of-the-Art Violence Detection Techniques: A review. *Asian Journal of Research in Computer Science*, 29-42.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, N., Prasad, O., & Sujithra, T. (2022). Deep Learning-Based Violence Detection from Videos. *Intelligent Data Engineering and Analytics: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021)*, 323-332.

Sjoberg, M., Baveye, Y., Wang, H., Quang, V. L., Ionescu, B., Dellandréa, E., Schedl, M., Demarty, C.-H., & Chen, L. (2015). The MediaEval 2015 Affective Impact of Movies Task. *MediaEval*, *1436*.

Soliman, M. M., Kamal, M. H., El-Massih Nashed, M. A., Mostafa, Y. M., Chawky, B. S., & Khattab, D. (2019). Violence Recognition from Videos using Deep Learning Techniques. *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, 80-85.

Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (XAI) methods. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2239-2250.

Srivastava, A., Badal, T., Garg, A., Vidyarthi, A., & Singh, R. (2021). Recognizing human violent action using drone surveillance within real-time proximity. *Journal of Real-Time Image Processing*, *18*, 1851-1863.

Srivastava, A., Badal, T., Saxena, P., Vidyarthi, A., & Singh, R. (2022). UAV surveillance for violence detection and individual identification. *Automated Software Engineering*, *29*(1), 28.

Su, Y., Lin, G., Zhu, J., & Wu, Q. (2020). Human interaction learning on 3d skeleton point clouds for video violence recognition. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 74-90.

Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6479-6488.

Swamy, V., Radmehr, B., Krco, N., Marras, M., & Käser, T. (2022). Evaluating the explainers: black-box explainable machine learning for student success prediction in MOOCs. *arXiv preprint arXiv:2207.00551*.

Talha, K. R., Bandapadya, K., & Khan, M. M. (2022). Violence Detection Using Computer Vision Approaches. *2022 IEEE World AI IoT Congress (AIIoT)*, 544-550.

Traoré, A., & Akhloufi, M. A. (2020a). 2D bidirectional gated recurrent unit convolutional neural networks for end-to-end violence detection in videos. *International Conference on Image Analysis and Recognition*, 152-160.

Traoré, A., & Akhloufi, M. A. (2020b). Violence detection in videos using deep recurrent and convolutional neural networks. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 154-159.

Tüzün, E., Tekinerdogan, B., Macit, Y., & İnce, K. (2019). Adopting integrated application lifecycle management within a large-scale software company: An action research approach. *Journal of Systems and Software*, *149*, 63-82.

Ullah, F. U. M., Muhammad, K., Haq, I. U., Khan, N., Heidari, A. A., Baik, S. W., & de Albuquerque, V. H. C. (2021). AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks. *IEEE Transactions on Industrial Informatics*, *18*(8), 5359-5370.

Ullah, F. U. M., Obaidat, M. S., Muhammad, K., Ullah, A., Baik, S. W., Cuzzolin, F., Rodrigues, J. J., & de Albuquerque, V. H. C. (2022). An intelligent system for complex violence pattern analysis and detection. *International Journal of Intelligent Systems*, *37*(12), 10400-10422.

Vidhya, J., & Uthra, R. (2021). Violence detection in videos using Conv2D VGG-19 architecture and LSTM network. *Proceedings of the Algorithms, Computing and Mathematics Conference, Chennai, India*, 19-20.

Vijeikis, R., Raudonis, V., & Dervinis, G. (2022). Efficient violence detection in surveillance. *Sensors*, *22*(6), 2216.

Vomfell, L., Härdle, W. K., & Lessmann, S. (2018). Improving crime count forecasts using Twitter and taxi data. *Decision Support Systems*, *113*, 73-85.

Vosta, S., & Yow, K.-C. (2022). A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras. *Applied Sciences*, *12*(3).

White, S. J., Sin, J., Sweeney, A., Salisbury, T., Wahlich, C., Montesinos Guevara, C. M., Gillard, S., Brett, E., Allwright, L., Iqbal, N., et al. (2024). Global prevalence and mental health outcomes of intimate partner violence among women: a systematic review and meta-analysis. *Trauma, Violence, & Abuse*, *25*(1), 494-511.

Wintarti, A., Puspitasari, R. D. I., & Imah, E. M. (2022). Violent Videos Classification Using Wavelet and Support Vector Machine. *2022 International Conference on ICT for Smart Society (ICISS)*, 01-05.

Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., & Yang, Z. (2020). Not only look, but also listen: Learning multimodal violence detection under weak supervision. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, 322-339.

Yao, H., & Hu, X. (2023). A survey of video violence detection. *Cyber-Physical Systems*, *9*(1), 1-24.

Yue, H., Xie, H., Liu, L., & Chen, J. (2022). Detecting people on the street and the streetscape physical environment from Baidu street view images and their effects on community-level street crime in a Chinese city. *ISPRS International Journal of Geo-Information*, *11*(3), 151.

Zhang, Z., Yuan, D., Li, X., & Su, S. (2022). Violent Target Detection Based on Improved YOLO Network. *International Conference on Artificial Intelligence and Security*, 480-492.

Zheng, Z., Zhong, W., Ye, L., Fang, L., & Zhang, Q. (2021). Violent scene detection of film videos based on multi-task learning of temporal-spatial features. *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 360-365.

Zhou, L. (2022). End-to-end video violence detection with transformer. *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, 880-884.

Zhou, W., Min, X., Zhao, Y., Pang, Y., & Yi, J. (2023). A Multi-Scale Spatio-Temporal Network for Violence Behavior Detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.