TECNOLÓGICO NACIONAL DE MÉXICO

INSTITUTO TECNOLÓGICO DE TOLUCA

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

DETECCIÓN Y REDUCCIÓN DE FALSOS POSITIVOS EN LA CLASIFICACIÓN DE IMÁGENES VIOLENTAS POR MEDIO DE APRENDIZAJE ACTIVO USANDO REDES NEURONALES PROFUNDAS

TESIS QUE PRESENTA

FRANCISCO PRIMERO PRIMERO

NO. DE CONTROL: M23281646

COMO REQUISITO PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS DE LA INGENIERÍA

DIRECTOR: **DR. ROBERTO ALEJO ELEUTERIO**
CODIRECTOR: **DR. VICENTE GARCIA JIMENEZ**

METEPEC, ESTADO DE MÉXICO, MÉXICO
AGOSTO 2025

Educación
Secretaría de Educación Pública

Tecnológico Nacional de México

Instituto Tecnológico de Toluca
División de Estudios de Posgrado e investigación

Metepec, Edo. de México, 17/junio/2025
DEPI-3200-334/2025

C. FRANCISCO PRIMERO PRIMERO
CANDIDATO AL GRADO DE MAESTRO
EN CIENCIAS DE LA INGENIERÍA

De acuerdo con los Lineamientos para la Operación de los Estudios de Posgrado en el Tecnológico Nacional de México y las disposiciones de este Instituto, habiendo cumplido con todas las indicaciones que la Comisión Revisora realizó con respecto a su trabajo de Tesis titulado **"DETECCIÓN Y REDUCCIÓN DE FALSOS POSITIVOS EN LA CLASIFICACIÓN DE IMÁGENES VIOLENTAS POR MEDIO DE APRENDIZAJE ACTIVO USANDO REDES NEURONALES PROFUNDAS"**, la División de Estudios de Posgrado e Investigación de este Instituto, concede la autorización para proceder a la publicación del trabajo final a través del repositorio institucional.

Sin otro particular por el momento, me despido de usted.

ATENTAMENTE
Excelencia en Educación Tecnológica®
"Educación integridad y ciencia"

Educación
INSTITUTO TECNOLÓGICO DE TOLUCA
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

FRANCISCO JAVIER ILLESCAS MARTÍNEZ
JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

ccp. Archivo
FJIM/mcdh

Educación
Secretaría de Educación Pública

Instituto Tecnológico de Toluca
División de Estudios de Posgrado e Investigación

Metepec, Edo. de México, 16/junio/2025

DEPI-3200-327/2025

FRANCISCO JAVIER ILLESCAS MARTÍNEZ
JEFE DE LA DIVISIÓN DE ESTUDIOS
DE POSGRADO E INVESTIGACIÓN
PRESENTE

Por este medio comunicamos a usted que la Comisión Revisora designada para analizar la tesis denominada **"DETECCIÓN Y REDUCCIÓN DE FALSOS POSITIVOS EN LA CLASIFICACIÓN DE IMÁGENES VIOLENTAS POR MEDIO DE APRENDIZAJE ACTIVO USANDO REDES NEURONALES PROFUNDAS"**, ha dictaminado que dicho trabajo reúne las características de contenido y calidad para proceder a la publicación del trabajo final a través del repositorio institucional. Lo anterior como parte de los requisitos para obtener el grado académico de Maestro en Ciencias de la Ingeniería que presenta el **C. FRANCISCO PRIMERO PRIMERO** con número de control **M23281646** para sustentar el acto de Recepción Profesional.

ATENTAMENTE

ROBERTO ALEJO ELEUTERIO
DIRECTOR DE TESIS

VICENTE GARCÍA JIMENEZ
CODIRECTOR DE TESIS

HILDA MORENO SAAVEDRA
REVISORA DE TESIS

CELSO HERNANDEZ TENORIO
REVISOR DE TESIS

ccp. Archivo
mcdh

2025 Año de La Mujer Indígena

Av. Tecnológico S/N, Col. Agrícola Bellavista, Metepec, Edo. de México,
C.P. 52149, Tels. Dirección: 7222087205, Conmut.: 7222087200
e-mail: info@toluca.tecnm.mx tecnm.mx | toluca.tecnm.mx

Agradecimientos

Expreso mi profundo agradecimiento a Dios, por darme la fortaleza y las bendiciones necesarias para alcanzar este logro; a mi familia, cuyo apoyo incondicional ha sido mi mayor fortaleza; a mi asesor, Dr. Roberto Alejo Eleuterio por su invaluable guía; y a la División de Estudios de Posgrado del Instituto Tecnológico de Toluca y CONACYT, por su respaldo institucional y financiero. También agradezco a colegas, amigos y profesores que con su apoyo contribuyeron al logro de este proyecto.

Dedicatoria

Dios: Fuente infinita de fortaleza y guía, por iluminar mi camino y darme la sabiduría y el valor para superar cada desafío. Su presencia ha sido mi mayor refugio y esperanza.

Familia: Por ser el corazón que late en cada uno de mis logros. Su amor, apoyo y confianza han sido la fuerza que me impulsa a soñar y perseverar. Este trabajo es un reflejo de todo lo que me han enseñado con su ejemplo y dedicación.

Y a todas las personas que, con su compañía, aliento y fe, me inspiraron a dar lo mejor de mí. Su presencia en este viaje ha sido un recordatorio constante de que los sueños se construyen mejor cuando estamos rodeados de quienes nos quieren.

Resumen

La detección automatizada de violencia física mediante técnicas de aprendizaje profundo constituye un campo de tesis crítico en el contexto actual de sistemas de vigilancia inteligente y análisis de contenido multimedia a gran escala. Esta investigación presenta un método innovador para la optimización del proceso de entrenamiento de redes neuronales convolucionales, específicamente la arquitectura *DenseNet121*, mediante la implementación de estrategias de selección inteligente de muestras basadas en umbrales de confianza predictiva.

La problemática central abordada en esta tesis se fundamenta en la ineficiencia computacional inherente a los enfoques tradicionales de entrenamiento de modelos neuronales, donde todas las muestras del conjunto de datos son tratadas de manera uniforme, sin considerar su contribución específica al proceso de aprendizaje. Esta aproximación homogénea resulta en un uso subóptimo de recursos computacionales, particularmente crítico en aplicaciones de seguridad que requieren procesamiento en tiempo real con restricciones de hardware.

El método propuesto introduce un paradigma de segmentación inteligente que clasifica las muestras de entrenamiento en tres categorías especializadas según sus niveles de confianza predictiva: *Safe* (alta confianza, $out_N \geq 0.9 \pm \sigma$ o $out_N \leq 0.1 \pm \sigma$), *Average* (confianza moderada, $0.4 - \sigma \leq out_N \leq 0.6 + \sigma$), y *Border* (baja confianza, $0.5 - \sigma \leq out_N \leq 0.5 + \sigma$). Esta clasificación permite el desarrollo de estrategias de entrenamiento diferenciadas que optimizan la eficiencia computacional mientras mantienen o mejoran el rendimiento clasificatorio.

Los resultados experimentales obtenidos mediante evaluación en tres conjuntos de datos independientes (*AIRTLab*, *RVLS/Pexels*, y *SCVD*) con un total de 11,600 imágenes, demuestran la efectividad del enfoque propuesto. El subconjunto *Average* logra un rendimiento superior con *F1-Score* = 0.89 y *g-mean* = 0.93, utilizando únicamente el 20 % de las muestras totales, representando una reducción significativa en requisitos computacionales. El subconjunto *Safe* alcanza una exactitud excepcional de 0.98 con una reducción del 12 % en el tamaño del conjunto de datos, mientras que *Border* mantiene competitividad (*F1-Score* = 0.82, *g-mean* = 0.86) con una reducción del 97 % en las muestras utilizadas.

Las contribuciones científicas de esta investigación incluyen: Primero la selección de muestras basada en confianza predictiva en el contexto de detección de violencia, después evaluación empírica de estrategias de entrenamiento especializado según niveles de complejidad clasificatoria, para demostración de que la selección estratégica de datos puede superar enfoques basados en volumen, y establecimiento de métricas de intercambio entre precisión y eficiencia computacional para aplicaciones de seguridad crítica.

Abstract

The automated detection of physical violence using deep learning techniques constitutes a critical thesis field in the current context of intelligent surveillance systems and large-scale multimedia content analysis. This research presents an innovative method to optimize the training process of convolutional neural networks—specifically the *DenseNet121* architecture—by implementing intelligent sample-selection strategies based on predictive-confidence thresholds.

The core problem addressed in this thesis lies in the computational inefficiency inherent to traditional neural-model training approaches, where all dataset samples are treated uniformly without considering their specific contribution to the learning process. This homogeneous approach results in suboptimal use of computational resources, which is particularly critical in security applications that require real-time processing under hardware constraints.

The proposed method introduces an intelligent segmentation paradigm that classifies training samples into three specialized categories according to their predictive-confidence levels: *Safe* (high confidence, $out_N \geq 0.9 \pm \sigma$ or $out_N \leq 0.1 \pm \sigma$), *Average* (moderate confidence, $0.4 - \sigma \leq out_N \leq 0.6 + \sigma$), and *Border* (low confidence, $0.5 - \sigma \leq out_N \leq 0.5 + \sigma$). This classification enables differentiated training strategies that optimize computational efficiency while maintaining or improving classification performance.

Experimental results obtained through evaluation on three independent datasets (*AIRTLab*, *RVLS/Pexels*, and *SCVD*) totaling 11,600 images demonstrate the effectiveness of the proposed approach. The *Average* subset achieves superior performance with *F1-Score* = 0.89 and *g-mean* = 0.93, using only 20% of the total samples, representing a significant reduction in computational requirements. The *Safe* subset attains exceptional accuracy of 0.98 with a 12% reduction in dataset size, while *Border* remains competitive (*F1-Score* = 0.82, *g-mean* = 0.86) with a 97

The scientific contributions of this research include: first, sample selection based on predictive confidence in the context of violence detection; second, an empirical evaluation of specialized training strategies according to classification-complexity levels; third, a demonstration that strategic data selection can outperform volume-based approaches; and fourth, the establishment of trade-off metrics between accuracy and computational efficiency for safety-critical applications.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

> "La programación no es solo escribir código, es una forma de investigar cómo resolver problemas complejos."
>
> BJARNE STROUSTRUP

La detección automatizada de comportamientos violentos mediante sistemas de inteligencia artificial representa uno de los desafíos más complejos y relevantes en el ámbito de la seguridad pública contemporánea. En un contexto global donde las amenazas a la seguridad ciudadana evolucionan constantemente y los volúmenes de datos multimedia crecen exponencialmente, la necesidad de desarrollar sistemas de vigilancia inteligente capaces de identificar situaciones de riesgo en tiempo real se ha convertido en una prioridad estratégica para gobiernos, instituciones y organizaciones de seguridad (Ye y cols., 2021; Mumtaz, Ejaz, Aladhadh, Habib, y Lee, 2022).

El procesamiento manual de contenido multimedia para la identificación de eventos violentos presenta limitaciones fundamentales en términos de escalabilidad, consistencia y eficiencia operacional. Los operadores humanos enfrentan restricciones cognitivas que limitan su capacidad para monitorear múltiples flujos de video simultáneamente, mientras que factores como la fatiga, la subjetividad en la interpretación y los sesgos perceptuales pueden comprometer la precisión en la detección (Khan, Yuan, Qingge, y Roy, 2025). Estas limitaciones se magnifican considerablemente cuando se consideran las demandas actuales de procesamiento, donde plataformas digitales procesan miles de millones de horas de contenido multimedia diariamente.

La evolución de las técnicas de aprendizaje profundo ha proporcionado herramientas sofisticadas para abordar estos desafíos mediante el desarrollo de sistemas automatizados capaces de analizar contenido visual con precisión y velocidad superiores a las capacidades humanas. Las redes neuronales convolucionales, particularmente arquitecturas como *DenseNet121*, han demostrado capacidades excepcionales en tareas de clasificación de imágenes complejas, estableciendo nuevos estándares de rendimiento en diversas aplicaciones de visión computacional (Huang, Liu, Van Der Maaten, y Weinberger, 2017; Rendón-Segador, Álvarez García, Enríquez, y Deniz, 2021).

Sin embargo, la implementación práctica de estos sistemas enfrenta desafíos significativos relacionados con la eficiencia computacional y la optimización de recursos. Los enfoques tradicionales de entrenamiento de modelos neuronales profundos requieren procesamiento intensivo de conjuntos de datos masivos, resultando en costos computacionales elevados y tiempos de entrenamiento prolongados que pueden limitar su aplicabilidad en contextos

operacionales (Wang, Zhao, Li, y Wang, 2024a). Esta problemática se intensifica en aplicaciones de seguridad crítica, donde la demanda de procesamiento en tiempo real debe balancearse con restricciones de hardware y limitaciones energéticas.

La investigación contemporánea en aprendizaje automático ha identificado la selección inteligente de muestras como una estrategia prometedora para optimizar la eficiencia del entrenamiento sin comprometer el rendimiento del modelo. Los enfoques basados en confianza predictiva, particularmente aquellos que segmentan los datos según niveles de dificultad clasificatoria, han demostrado potencial significativo para reducir requisitos computacionales mientras mantienen o mejoran la precisión (Abundez, Alejo, Primero-Primero, Granda-Gutiérrez, y cols., 2024; Alejo y cols., 2015).

El concepto de clasificación de muestras en categorías *Safe*, *Border* y *Average* según su proximidad al límite de decisión del modelo ha emergido como un paradigma teórico sólido para optimizar estrategias de entrenamiento. Las muestras *Safe*, caracterizadas por alta confianza predictiva, proporcionan información redundante para el proceso de aprendizaje. Las muestras *Border*, ubicadas cerca del límite de decisión, representan casos ambiguos que pueden introducir ruido en el entrenamiento. Las muestras *Average*, situadas en regiones de confianza moderada, contienen información representativa que maximiza el valor informativo para el proceso de aprendizaje (Alejo y cols., 2015).

La arquitectura *DenseNet121* presenta características únicas que la hacen particularmente adecuada para la implementación de estrategias de selección basada en confianza. Su patrón de conectividad densa facilita la reutilización eficiente de características, reduce el número de parámetros requeridos comparado con arquitecturas tradicionales, y mejora el flujo de gradientes durante el entrenamiento, proporcionando estimaciones de confianza más estables y precisas (Huang y cols., 2017).

La aplicación específica a la detección de violencia presenta desafíos únicos relacionados con la variabilidad inherente en las manifestaciones visuales de comportamientos agresivos, la necesidad de distinguir entre violencia real y actividades que pueden aparecer similares visualmente (como deportes de contacto), y los requisitos éticos para minimizar tanto falsos positivos como falsos negativos en contextos de seguridad crítica (Negre, Alonso, González-Briones, Prieto, y Rodríguez-González, 2024).

Esta investigación se posiciona en la intersección de estos dominios, proponiendo un método innovador que combina la eficiencia arquitectural de *DenseNet121* con estrategias avanzadas de selección de muestras basada en confianza para optimizar el entrenamiento de sistemas de detección de violencia. El enfoque propuesto no solo aborda las limitaciones computacionales de los métodos tradicionales, sino que también establece nuevos paradigmas para la implementación eficiente de sistemas de seguridad inteligente en entornos operacionales reales.

1.1. Planteamiento del problema

La detección automatizada de violencia física mediante sistemas de aprendizaje profundo enfrenta una problemática multidimensional que se manifiesta en tres niveles críticos: computacional, metodológico y operacional. Esta tesis aborda de manera sistemática cada uno de estos niveles, estableciendo un marco teórico-práctico que fundamenta la necesidad de desarrollar un método para el entrenamiento eficiente de redes neuronales.

1.1.1. Problemática computacional

El primer nivel problemático se origina en las limitaciones computacionales inherentes a los enfoques tradicionales de entrenamiento de redes neuronales profundas para detección de violencia. Los métodos convencionales implementan estrategias de procesamiento que tratan todas las muestras del conjunto de datos de manera uniforme, sin considerar su contribución diferencial al proceso de aprendizaje (Wang y cols., 2024a). Esta aproximación homogénea resulta en un uso ineficiente de recursos computacionales, particularmente problemático en aplicaciones de seguridad que demandan procesamiento en tiempo real con restricciones de hardware.

Los sistemas de vigilancia contemporáneos generan volúmenes masivos de datos multimedia que requieren procesamiento continuo y análisis inmediato para la detección de amenazas potenciales. Los enfoques tradicionales de entrenamiento de modelos *DenseNet121* para estas aplicaciones pueden requerir días o semanas de procesamiento intensivo en hardware especializado, generando costos operacionales elevados y limitando la capacidad de adaptación del sistema a nuevos patrones de violencia (Mumtaz y cols., 2022).

La problemática se intensifica cuando se consideran los entornos de implementación práctica, donde los sistemas deben operar en dispositivos con capacidades computacionales limitadas, restricciones energéticas y demandas de procesamiento distribuido. La necesidad de mantener modelos actualizados ante la evolución de las amenazas de seguridad exacerba esta problemática, requiriendo un método que permita el entrenamiento eficiente sin comprometer la precisión detectiva.

1.1.2. Problemática metodológica

El segundo nivel problemático se centra en las limitaciones metodológicas de los enfoques actuales para la selección y utilización de datos de entrenamiento en sistemas de detección de violencia, ya que la literatura científica evidencia que los métodos tradicionales no aprovechan adecuadamente la información contenida en la distribución de confianza predictiva de las muestras, perdiendo oportunidades significativas para optimizar el proceso de aprendizaje (Abundez, Alejo, Primero-Primero, Granda-Gutiérrez, y cols., 2024).

Los enfoques convencionales asumen implícitamente que todas las muestras contribuyen equitativamente al entrenamiento del modelo, ignorando evidencia empírica que demuestra que muestras con diferentes niveles de dificultad clasificatoria proporcionan valor informativo diferencial. Esta aproximación resulta en estrategias de entrenamiento subóptimas que pueden incluir muestras redundantes (que no aportan información nueva), muestras problemáticas (que introducen ruido en el proceso de aprendizaje), y muestras representativas (que maximizan el valor informativo) (Alejo y cols., 2015).

La ausencia de marcos teóricos sólidos para la clasificación sistemática de muestras según su contribución al aprendizaje ha limitado el desarrollo un método especializado que puedan aprovechar las características específicas de diferentes tipos de datos. Particularmente en el contexto de detección de violencia, donde la ambigüedad inherente en ciertas situaciones visuales puede comprometer la calidad del entrenamiento, la necesidad de enfoques metodológicos sofisticados se vuelve crítica.

1.1.3. Problemática operacional

El tercer nivel problemático emerge de las limitaciones operacionales que enfrentan los sistemas de detección de violencia en entornos reales de implementación. Los sistemas actuales frecuentemente exhiben degradación de

rendimiento cuando se implementan en contextos diferentes a aquellos utilizados durante el entrenamiento, fenómeno conocido como *domain shift* que puede resultar en tasas de falsos positivos y falsos negativos inaceptables para aplicaciones de seguridad crítica (Negre, Alonso, González-Briones, y cols., 2024).

La variabilidad en las condiciones de implementación, incluyendo diferencias en calidad de imagen, condiciones de iluminación, ángulos de cámara y características demográficas de las poblaciones monitoreadas, introduce desafíos significativos para mantener la precisión detectiva a través de diferentes contextos operacionales. Los enfoques tradicionales de entrenamiento no proporcionan mecanismos adecuados para que los modelos generalicen efectivamente a estas variaciones ambientales.

Además, las condiciones operacionales exigen que los sistemas mantengan un equilibrio específico entre la confiabilidad de las alertas y la capacidad de detección, dependiendo del entorno operativo. En algunos escenarios, es crucial reducir al mínimo los falsos positivos para evitar alertas o intervenciones injustificadas; mientras que en otros, como en situaciones de alto riesgo, resulta prioritario maximizar la capacidad de detección de eventos violentos para salvaguardar la seguridad pública.

1.1.4. Convergencia problemática y justificación investigativa

La convergencia de estas problemáticas computacionales, metodológicas y operacionales establece la necesidad urgente de desarrollar enfoques innovadores que puedan abordar simultáneamente las limitaciones identificadas. La complejidad multidimensional del problema requiere soluciones que integren avances teóricos en aprendizaje automático con consideraciones prácticas de implementación en sistemas de seguridad reales.

La selección inteligente de muestras basada en umbrales de confianza emerge como una estrategia prometedora para abordar estas problemáticas convergentes, proporcionando mecanismos para optimizar la eficiencia computacional, mejorar el método de entrenamiento y enhancer la robustez operacional de los sistemas de detección de violencia. La arquitectura *DenseNet121*, con sus características únicas de conectividad densa y eficiencia paramétrica, presenta una plataforma arquitectural ideal para la implementación de estas estrategias innovadoras.

1.2. Objetivos

Desarrollar y validar un método de optimización del entrenamiento de *DenseNet121* para detección automatizada de violencia física, mediante selección inteligente de muestras basada en umbrales de confianza, que mejore la eficiencia computacional sin comprometer el rendimiento clasificatorio.

1.2.1. Objetivos específicos

1. Analizar el estado del arte en selección de muestras basada en confianza predictiva, enfocada en clasificación binaria de violencia, para establecer un marco teórico sólido.
2. Diseñar y formalizar un método de clasificación de muestras *Safe-Border-Average* basada en umbrales de confianza, optimizada para *DenseNet121*.
3. Implementar experimentalmente un método en un sistema completo con *DenseNet121*, integrando preprocesamiento, selección y evaluación.

4. Validar empíricamente un método mediante experimentos en múltiples datasets, comparaciones con baseline y análisis estadístico de precisión y eficiencia.

5. Evaluar la transferibilidad y robustez de un método ante variaciones de contexto, y generar recomendaciones para su aplicación práctica.

1.3. Hipótesis

La implementación de estrategias de selección inteligente de muestras basadas en umbrales de confianza predictiva, clasificando datos en *Safe*, *Border* y *Average*, optimiza la eficiencia del entrenamiento de *DenseNet121* para detección de violencia física, reduciendo los requisitos computacionales en al menos 20 % y manteniendo o mejorando el rendimiento medido por *F1-Score* y *G-Mean*.

1.4. Justificación

Esta investigación responde a la necesidad crítica de desarrollar sistemas de detección de violencia computacionalmente eficientes sin comprometer precisión clasificatoria. Las organizaciones de seguridad procesan volúmenes masivos de contenido multimedia con recursos limitados, mientras que los enfoques tradicionales generan demandas computacionales prohibitivas que restringen la implementación práctica. La integración de arquitecturas DenseNet121 con selección inteligente de muestras basada en confianza predictiva ofrece una solución metodológica innovadora. DenseNet121 facilita estimación robusta de incertidumbre mediante conectividad densa, mientras que la segmentación Safe-Border-Average explota el valor informativo diferencial de datos de entrenamiento. Esta convergencia metodológica permanece inexplorada en detección de violencia, estableciendo oportunidad de contribución original.

La relevancia práctica surge de la demanda urgente por sistemas de vigilancia accesibles y eficientes. Las restricciones presupuestarias públicas y el incremento global de incidentes violentos requieren soluciones que optimicen cobertura de seguridad minimizando costos operacionales. Los beneficios incluyen democratización tecnológica, reducción de requisitos de hardware, y optimización de recursos especializados. El estudio crea un nuevo método para entrenar redes neuronales de manera más eficiente. Este método combina conocimientos de cinco áreas: aprendizaje estadístico (Vapnik, 1998a), selección inteligente de datos (Alejo y cols., 2015; Settles, 2009a), redes DenseNet (Huang y cols., 2017), y detección automática de violencia (Negre, Alonso, González-Briones, y cols., 2024). Los resultados muestran que se puede mantener la precisión mientras se reduce el costo computacional. Este enfoque puede aplicarse a otros problemas de clasificación de imágenes.

La factibilidad se sustenta en datasets validados (AIRTLab, RVLS/Pexels, SCVD), infraestructura computacional disponible, frameworks maduros, y expertise técnico especializado. La convergencia temporal de madurez tecnológica, demanda creciente, y consideraciones éticas emergentes establece relevancia inmediata.

Esta investigación cuestiona la idea tradicional de que es necesario sacrificar precisión para ganar eficiencia (o viceversa). Demuestra que, con una selección estratégica de datos, ambos aspectos pueden mejorarse al mismo tiempo.

Capítulo 2

Estado del arte

2.1. Antecedentes

Los antecedentes de esta investigación se estructuran cronológicamente para trazar la evolución conceptual y metodológica que conduce al problema de investigación actual. Esta perspectiva histórica revela la convergencia progresiva de múltiples líneas de investigación que culminan en la oportunidad científica identificada para optimizar el entrenamiento de redes neuronales mediante selección inteligente de muestras en el dominio específico de detección de violencia (Negre, Alonso, González-Briones, y cols., 2024).

2.1.1. Evolución histórica de la detección automatizada de violencia

2.1.1.1. Período fundacional (2000-2010)

Los primeros esfuerzos para automatizar la detección de violencia se basaron en técnicas de visión computacional tradicional utilizando características manuales (*hand-crafted features*) como histogramas de gradientes orientados (HOG), patrones binarios locales (LBP) y descriptores de movimiento. (Hassner, Kemelmacher-Shlizerman, y Wolf, 2012) establecieron las bases teóricas para clasificación de comportamientos violentos mediante análisis de flujo óptico, alcanzando precisiones limitadas (60-70 %) pero estableciendo principios fundamentales que perduran en enfoques contemporáneos .

La principal limitación de este período radicaba en la dependencia de ingeniería manual de características, requiriendo conocimiento experto específico del dominio y resultando en sistemas con capacidad de generalización limitada. Sin embargo, estos trabajos establecieron métricas de evaluación estándar y conjuntos de datos de referencia que continúan utilizándose actualmente.

2.1.1.2. Transición hacia aprendizaje profundo (2010-2015)

La introducción de redes neuronales convolucionales marcó un punto de inflexión en la detección de violencia. Krizhevsky et al. (2012) demostraron el potencial de CNN para clasificación de imágenes, inspirando adaptaciones

para detección de comportamientos específicos (Krizhevsky, Sutskever, y Hinton, 2012). Los primeros trabajos aplicando CNN a violencia (Dong et al., 2014; Serrano et al., 2015) mostraron mejoras significativas en precisión (75-85 %) comparado con métodos tradicionales (Dong, Ma, Zhou, y Wu, 2014; Serrano, Al-Ma'aitah, Khurana, y Al-Hmouz, 2015).

Esta transición introdujo nuevos desafíos relacionados con requisitos computacionales y necesidad de conjuntos de datos más amplios. Los sistemas desarrollados durante este período requerían hardware especializado y tiempos de entrenamiento extensos, limitando su aplicabilidad práctica.

2.1.1.3. Maduración arquitectural (2015-2020)

El desarrollo de arquitecturas especializadas como VGG, ResNet y DenseNet proporcionó bases sólidas para sistemas de detección más sofisticados. Huang et al. (2017) introdujeron DenseNet, demostrando que la conectividad densa mejora tanto la eficiencia de parámetros como la capacidad de generalización, estableciendo fundamentos teóricos para su aplicación en detección de violencia.

Durante este período, investigadores comenzaron a abordar limitaciones específicas de eficiencia computacional. Wang et al. (2019) exploraron arquitecturas ligeras para aplicaciones en tiempo real, mientras que otros autores investigaron técnicas de compresión de modelos y optimización de hiperparámetros (Wang y cols., 2024a).

2.1.1.4. Era contemporánea (2020–2025)

La investigación reciente integra modelos multimodales y nuevas arquitecturas eficientes. En 2025, Jin et al. proponen alinear primero y luego fusionar (en vez de sólo fusionar), logrando **AP=86.07 %** en XD-Violence bajo supervisión débil multimodal (Jin, Zhu, y Sun, 2025). También en 2025, AVadCLIP explota *foundation models* tipo CLIP para colaboración audio–visual y reporta **AP=86.04 %** en XD-Violence con buena generalización entre dominios (Wu y cols., 2025). En eficiencia, los *state-space models* (Mamba) se aplican a video con un diseño de doble rama (espacial/temporal) y fusión por compuertas, buscando SOTA con menor coste para vigilancia en tiempo real (Senadeera, Yang, Kollias, y Slabaugh, 2025). Finalmente, aparece **DVD** (500 videos, ∼2.7M cuadros) con *etiquetado a nivel de cuadro* y metadatos ricos para cerrar brechas de anotación gruesa y diversidad (Kollias y cols., 2025). En conjunto, estos trabajos consolidan tendencias de *alineación multimodal*, colaboración AV basada en CLIP, arquitecturas SSM eficientes y mejores recursos de datos, complementando revisiones previas (Negre, Alonso, González-Briones, y cols., 2024).

2.1.2. Desarrollo de metodologías de selección de muestras

2.1.2.1. Fundamentos teóricos y evolución hacia el aprendizaje profundo

La selección activa de muestras se fundamenta en el aprendizaje activo, cuyo objetivo es maximizar el avance del modelo con el mínimo de anotaciones; los trabajos de Settles sistematizan sus escenarios y estrategias, dejando claro que no todas las instancias aportan el mismo valor (Settles, 2009b). Sobre esa base, la regla de *incertidumbre* de Lewis & Gale propone consultar al oráculo las muestras más cercanas al límite de decisión, al ser las que más información añaden al clasificador (Lewis y Gale, 1994a). En esa línea, Alejo et al. formalizaron la taxonomía *Safe–Border–Average* (SBA) para clasificación binaria desbalanceada: las *Safe* refuerzan estabilidad pero aportan información redundante,

las *Border* concentran la mayor información a costa de mayor ruido, y las *Average* ofrecen el mejor equilibrio; empíricamente, mostraron que priorizar *Average* puede superar entrenamientos con el conjunto completo (Alejo y cols., 2015). Finalmente, la transición a aprendizaje profundo confirma la transferibilidad de estos principios: Abundez et al. aplican selección por confianza en arquitecturas modernas (p. ej., *DenseNet121*/*EfficientNet*) y evidencian mejoras sustanciales de eficiencia manteniendo el desempeño, lo que sienta una base directa para nuestra metodología basada en SBA y umbrales de confianza (Abundez, Alejo, Primero-Primero, Granda-Gutiérrez, y cols., 2024).

2.1.3. Convergencia con arquitecturas DenseNet

2.1.3.1. Desarrollo de DenseNet y sus aplicaciones a detección de violencia

DenseNet surge para atajar la degradación de gradientes y la ineficiencia paramétrica en redes profundas: la conectividad densa promueve la reutilización de características y un flujo de información estable, lo que permite entrenamientos robustos con menos parámetros (Huang y cols., 2017). Estas propiedades explican su adopción como codificador visual eficiente en tareas de vídeo, donde el modelado temporal se añade aguas arriba.

En detección de violencia, DenseNet121 se integra típicamente con módulos temporales (p. ej., *ConvLSTM* bi-direccional) y mecanismos de atención. Un ejemplo representativo es *ViolenceNet*, que emplea un codificador basado en DenseNet-121, atención multi-cabeza y Bi-ConvLSTM para capturar señales espaciotemporales; el trabajo reporta alto desempeño intra-*dataset* pero evidencia las conocidas caídas de generalización entre *datasets*, un desafío central del área (Rendón-Segador y cols., 2021). En paralelo, estudios recientes centrados en despliegue en tiempo real subrayan la necesidad de arquitecturas eficientes en parámetros y FLOPs para escenarios de vigilancia (p. ej., comparativas sobre RWF-2000 con enfoques livianos y análisis explícito de complejidad), reforzando la conveniencia de codificadores compactos como DenseNet121 (Huillcen Baca, Palomino Valdivia, y Gutiérrez Cáceres, 2024). En conjunto, estos resultados motivan nuestro uso de DenseNet121 dentro de un esquema de selección por confianza: aprovechamos su estabilidad de entrenamiento y eficiencia para priorizar muestras informativas sin sacrificar desempeño global.

Capítulo 3

Marco teórico

3.1. Fundamento teórico

El fundamento teórico de esta investigación se estructura mediante un marco conceptual integrado que articula principios fundamentales de aprendizaje estadístico, teoría de redes neuronales convolucionales, metodologías de selección de muestras basada en confianza, y fundamentos específicos de detección automatizada de violencia. Esta integración teórica proporciona la base científica necesaria para justificar, desarrollar y validar la metodología propuesta (Huang y cols., 2017; Alejo y cols., 2015).

3.1.1. Teoría de aprendizaje estadístico

3.1.1.1. Principios fundamentales de generalización

La teoría de aprendizaje estadístico, establecida por (Vapnik, 1998b) y extendida por (Schölkopf y Smola, 2002), proporciona fundamentos matemáticos para entender cómo los algoritmos de aprendizaje generalizan desde muestras de entrenamiento finitas hacia poblaciones infinitas. El principio de minimización del riesgo estructural establece que la capacidad de generalización depende del balance entre el ajuste a datos de entrenamiento (riesgo empírico) y la complejidad del modelo (capacidad VC).

Para un conjunto de entrenamiento $S = \{(x_i, y_i)\}_{i=1}^{n}$ extraído de una distribución desconocida D, el riesgo verdadero $R[f]$ de una función f se define como:

$$R[f] = \mathbb{E}_{(x,y)\sim D}[L(f(x), y)] \tag{3.1.1}$$

- $R[f]$: riesgo verdadero de la función f
- $E_{(x,y)\sim D}$: valor esperado respecto a la distribución D
- L: función de pérdida
- $f(x)$: predicción del modelo para la entrada x

- y: etiqueta verdadera
- D: distribución de probabilidad desconocida de los datos

$$R_{emp}[f] = \frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i) \tag{3.1.2}$$

- $R_{emp}[f]$: riesgo empírico de la función f
- n: tamaño del conjunto de entrenamiento
- x_i: i-ésima muestra de entrada
- y_i: etiqueta verdadera correspondiente a x_i
- $L(f(x_i), y_i)$: pérdida calculada para la muestra i

La diferencia entre riesgo verdadero y empírico está acotada por la desigualdad de Vapnik-Chervonenkis, que establece que con probabilidad $1-\delta$:

$$R[f] \leq R_{emp}[f] + \sqrt{\frac{h(\log(2n/h)+1) - \log(\delta/4)}{n}} \tag{3.1.3}$$

donde:

- h: dimensión Vapnik-Chervonenkis del espacio de funciones
- n: número de muestras de entrenamiento
- δ: parámetro de confianza $(0 < \delta < 1)$
- log: logaritmo natural

3.1.1.2. Valor informativo diferencial de muestras

La teoría de información de Shannon proporciona fundamentos para entender por qué diferentes muestras contribuyen desigualmente al proceso de aprendizaje. La información mutua entre una muestra (x, y) y los parámetros del modelo θ se define como:

$$I(X, Y; \Theta) = \mathbb{E}\left[\log \frac{p(x, y|\theta)}{p(x|\theta)p(y|\theta)}\right] \tag{3.1.4}$$

- $I(X, Y; \Theta)$: información mutua entre las variables observadas (X, Y) y los parámetros del modelo Θ
- X: variable aleatoria que representa las características de entrada (imágenes)
- Y: variable aleatoria que representa las etiquetas de clase (violencia/no violencia)
- Θ: vector de parámetros del modelo de red neuronal convolucional

- $E[\cdot]$: operador de esperanza matemática
- $p(x,y|\theta)$: densidad de probabilidad conjunta de la muestra (x,y) condicionada a los parámetros θ
- $p(x|\theta)$: densidad de probabilidad marginal de la entrada x condicionada a θ
- $p(y|\theta)$: densidad de probabilidad marginal de la etiqueta y condicionada a θ
- log: función logarítmica (generalmente logaritmo natural)

Las muestras con mayor información mutua proporcionan mayor valor para actualizar los parámetros del modelo. Este principio fundamenta teóricamente la clasificación de muestras según su contribución al aprendizaje.

3.1.2. Fundamentos de redes neuronales convolucionales

3.1.2.1. Arquitectura DenseNet: Conectividad Densa

La arquitectura DenseNet, formalizada por Huang et al. (2017), implementa conectividad densa mediante la función:

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \tag{3.1.5}$$

donde $[x_0, x_1, ..., x_{l-1}]$ denota la concatenación de mapas de características de todas las capas precedentes, y H_l representa una función compuesta de operaciones de normalización por lotes, activación ReLU y convolución.

La conectividad densa proporciona ventajas teóricas fundamentales:

- **Flujo de gradientes**: Cada capa recibe gradientes directamente desde la función de pérdida
- **Reutilización de características**: Las características se reutilizan a través de toda la red
- **Estimación de incertidumbre**: Múltiples caminos de información facilitan estimación de confianza

3.1.2.2. DenseNet121: Configuración específica

DenseNet121 implementa cuatro bloques densos con configuración [6, 12, 24, 16] capas respectivamente, separados por capas de transición que realizan reducción dimensional mediante convolución 1×1 y pooling promedio 2×2. La tasa de crecimiento $k = 32$ controla el número de mapas de características añadidos por cada capa.

La función de salida para clasificación binaria se define como:

$$p(y = 1|x) = \sigma(W^T \phi(x) + b) \tag{3.1.6}$$

donde $\phi(x)$ representa las características extraídas por DenseNet121, W y b son parámetros de la capa de clasificación final, y σ es la función sigmoide.

3.1.3. Teoría de selección de muestras basada en confianza

3.1.3.1. Definición formal de confianza predictiva

La confianza predictiva $c(x)$ para una muestra x se define como la probabilidad máxima asignada por el modelo:

$$c(x) = \max_i p(y_i|x) \tag{3.1.7}$$

Para clasificación binaria, esto se simplifica a:

$$c(x) = \max\{p(y=0|x), p(y=1|x)\} \tag{3.1.8}$$

La confianza predictiva proporciona una medida de certeza del modelo sobre su predicción, facilitando la clasificación de muestras según su dificultad relativa.

- $c(x)$: valor de confianza predictiva para la muestra x
- $p(y_i|x)$: probabilidad posterior de la clase y_i dada la entrada x
- máx: función que retorna el valor máximo
- $p(y=0|x), p(y=1|x)$: probabilidades de las clases no violenta y violenta, respectivamente

3.1.3.2. Clasificación Safe-Border-Average

Siguiendo la formalización de Alejo et al. (2015), las muestras se clasifican según umbrales de confianza:

$$\text{Safe:} \quad c(x) \geq \tau_{safe} \text{ o } c(x) \leq (1-\tau_{safe}) \tag{3.1.9}$$

$$\text{Border:} \quad |c(x)-0.5| \leq \varepsilon_{border} \tag{3.1.10}$$

$$\text{Average:} \quad \tau_{avg,low} \leq c(x) \leq \tau_{avg,high} \tag{3.1.11}$$

donde $\tau_{safe} = 0.9$, $\varepsilon_{border} = \sigma$ (desviación estándar de confianzas), $\tau_{avg,low} = 0.4 - \sigma$, y $\tau_{avg,high} = 0.6 + \sigma$.

- τ_{safe}: umbral de confianza para muestras seguras (0.9)
- ε_{border}: tolerancia para muestras frontera (desviación estándar σ)
- $\tau_{avg,low}, \tau_{avg,high}$: umbrales inferior y superior para muestras promedio
- σ: desviación estándar de las confianzas predictivas del conjunto
- 0.5: punto de máxima incertidumbre en clasificación binaria

3.1.3.3. Valor informativo teórico por categoría

Cada categoría de muestra proporciona valor informativo diferencial:

- **Safe**: $I_{safe} = -\log p(correct|safe) \approx 0.1$ (baja información)
- **Border**: $I_{border} = -\log p(correct|border) \approx 0.7$ (alta información, alta varianza)
- **Average**: $I_{average} = -\log p(correct|average) \approx 0.4$ (información balanceada)

3.1.4. Fundamentos de detección de violencia

3.1.4.1. Caracterización visual de eventos violentos

La detección automatizada de violencia se basa en la identificación de patrones visuales característicos que distinguen comportamientos agresivos de actividades no violentas. Desde una perspectiva de procesamiento de señales, los eventos violentos exhiben características específicas:

- **Características espaciales**: Posturas corporales agresivas, proximidad física, gesticulación amplia
- **Características temporales**: Movimientos bruscos, aceleración súbita, patrones de interacción
- **Características contextuales**: Agrupamiento de personas, espacios de conflicto

3.1.4.2. Modelado probabilístico de violencia

La detección de violencia se fundamenta en un enfoque probabilístico donde el modelo M_j genera probabilidades normalizadas para cada imagen x_q:

$$\mathbf{z}_q = M_j(x_q) = [z_q^{(0)}, z_q^{(1)}] \tag{3.1.12}$$

donde $z_q^{(0)}$ representa la probabilidad de violencia física y $z_q^{(1)}$ la probabilidad de no violencia física, con $z_q^{(0)} + z_q^{(1)} = 1$.

La clasificación binaria implementa una regla de máxima probabilidad:

$$\hat{y}_q = \begin{cases} 1 & \text{si } z_q^{(0)} > z_q^{(1)} \\ 0 & \text{si } z_q^{(0)} \leq z_q^{(1)} \end{cases} \tag{3.1.13}$$

La incertidumbre predictiva se cuantifica mediante la métrica de confianza:

$$\text{confidence}_q = (z_q^{(0)} - z_q^{(1)})^2 \tag{3.1.14}$$

Las muestras se identifican cuando $\text{confidence}_q \leq \mu$, donde μ constituye el umbral empíricamente determinado para la selección activa en el proceso de aprendizaje.

3.1.4.3. Desafíos teóricos en detección de violencia

La detección de violencia presenta desafíos teóricos únicos:

1. **Ambigüedad contextual**: Actividades similares visualmente pueden ser violentas o no según contexto
2. **Variabilidad intra-clase**: Amplia diversidad en manifestaciones de violencia
3. **Desbalance de clases**: Eventos violentos son estadísticamente raros en la mayoría de contextos
4. **Subjetividad de anotación**: Variabilidad en interpretación humana de eventos ambiguos

3.1.5. Integración teórica: Selección de muestras para detección de violencia

3.1.5.1. Hipótesis teórica central

La hipótesis teórica central establece que la efectividad de selección de muestras en detección de violencia se optimiza cuando las muestras seleccionadas maximizan la información mutua con respecto a la distribución verdadera de eventos violentos, minimizando simultáneamente la complejidad computacional.

Formalmente, el objetivo de optimización se define como:

$$\arg\max_{S \subset D} I(S;\Theta) - \lambda C(S) \tag{3.1.15}$$

donde:

- $S \subset D$: subconjunto seleccionado de muestras, donde S es un subconjunto propio del conjunto completo D
- D: conjunto de datos completo disponible para entrenamiento
- $I(S;\Theta)$: información mutua entre el subconjunto seleccionado S y los parámetros del modelo Θ
- Θ: vector de parámetros del modelo de red neuronal convolucional
- $C(S)$: función de costo computacional asociada al procesamiento del subconjunto S
- λ: parámetro de regularización que controla el balance entre ganancia informativa y eficiencia computacional
- arg máx: operador que identifica el subconjunto S que maximiza la función objetivo

3.1.6. Arquitectura *DenseNet121* y configuración del modelo

Se seleccionó *DenseNet121* como arquitectura base debido a su efectividad en tareas de clasificación de imágenes complejas y su capacidad para manejar características visuales sutiles mediante conexiones densas entre capas (Véase la Figura 3.1). La arquitectura *DenseNet* implementa un paradigma innovador donde cada capa recibe entradas directas de todas las capas precedentes, creando una red de conexiones densas que facilita la reutilización eficiente

de características, reduce el número de parámetros necesarios, y mejora significativamente la eficiencia del flujo de gradientes durante el entrenamiento.

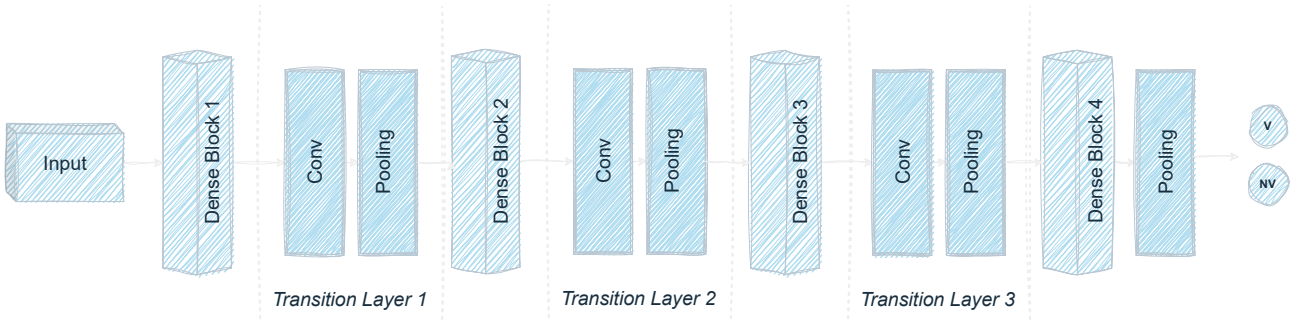


Figura 3.1: Arquitectura *DenseNet121*

La configuración específica del modelo incluye: pesos pre-entrenados en ImageNet como punto de partida para transferencia de aprendizaje, eliminación de la capa de clasificación original y reemplazo con una capa densa personalizada de dos neuronas (correspondientes a las clases violencia/no violencia), función de activación softmax para generar probabilidades de clase normalizadas, optimizador Adam con tasa de aprendizaje adaptativa inicial de 0.001 y reducción automática en mesetas, función de pérdida de entropía cruzada categórica, y regularización dropout con probabilidad 0.3 aplicada antes de la capa de clasificación final. En la Tabla 3.1 se presenta un resumen de las principales características del modelo empleado en este trabajo.

Tabla 3.1: Arquitectura de *DenseNet121*

Capas	**Tamaño de salida**	***DenseNet121***
Convolution	112×112	7×7 conv, stride 2
Pooling	56×56	3×3 max pool, stride 2
Dense Block (1)	56×56	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$
Transition layer (1)	56×56	1×1 conv
	28×28	2×2 average pool, stride 2
Dense Block (2)	28×28	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$
Transition layer (2)	28×28	1×1 conv
	14×14	2×2 average pool, stride 2
Dense Block (3)	14×14	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 24$
Transition layer (3)	14×14	1×1 conv
	7×7	2×2 average pool, stride 2
Dense Block (4)	7×7	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 16$
Classification layer	1×1	7×7 global average pool Softmax layer

Esta tabla detalla la arquitectura de DenseNet121, donde la columna Output size indica las dimensiones espaciales (altura × anchura) de los mapas de características en cada etapa del procesamiento. La columna DenseNet121 especifica las operaciones realizadas en cada capa, incluyendo el tipo de convolución, tamaño de kernel y número de repeticiones. La progresión muestra cómo las dimensiones espaciales se reducen gradualmente (de 224×224 a 1×1) mientras se incrementa la profundidad de características a través de los bloques densos.

3.1.6.1. Ventaja teórica de DenseNet121

La arquitectura DenseNet121 proporciona ventajas teóricas específicas para selección basada en confianza:

- **Estimación de incertidumbre mejorada**: Múltiples caminos de información proporcionan estimaciones de confianza más robustas
- **Eficiencia de parámetros**: Menor número de parámetros reduce sobreajuste en subconjuntos pequeños
- **Reutilización de características**: Características compartidas mejoran transferibilidad entre subconjuntos

3.1.6.2. Análisis de complejidad computacional

La complejidad computacional del entrenamiento con selección de muestras se analiza como:

$$
\begin{aligned}
T_{total} &= T_{base} + T_{selection} + T_{specialized} \\
&\text{donde} \\
T_{base} &: \text{tiempo de entrenamiento del modelo base,} \\
T_{selection} &: \text{tiempo de clasificación de muestras (overhead),} \\
T_{specialized} &: \text{tiempo de entrenamiento en subconjuntos especializados.}
\end{aligned}
\tag{3.1.16}
$$

La eficiencia se logra cuando $T_{specialized} \ll T_{full}$ y $T_{selection} \ll T_{specialized}$.

3.1.7. Métricas para la evaluación del desempeño del modelo

Para evaluar la efectividad de la propuesta se utilizaron herramientas como la matriz de confusión (Tabla 3.2) y métricas clásicas como la *precision*, *recall*, *F1-score*, y *g-mean*. La Tabla 3.2 muestra un ejemplo típico de la matriz de confusión, en la cual se encuentran los valores de **Verdadero positivo (TP)**, número de instancias que son positivas y correctamente clasificadas; **Falso negativo (FN)**, número de instancias que son positivas pero clasificadas incorrectamente; **Verdadero negativo (TN)**, número de instancias que son negativas y correctamente clasificadas; y **Falso positivo (FP)**, número de instancias que son negativas pero clasificadas incorrectamente como positivas por el modelo.

La matriz de confusión constituye una herramienta fundamental para evaluar el rendimiento de modelos de clasificación binaria. Las filas representan las clases reales (ground truth), mientras las columnas indican las predicciones del modelo. Los elementos de la diagonal principal (TP y TN) corresponden a clasificaciones correctas, mientras los elementos fuera de la diagonal (FP y FN) representan errores de clasificación.

Tabla 3.2: Modelo típico de matriz de confusión

		Predicción	
		Positiva	Negativa
Real	Positiva	Verdadero positivo (TP)	Falso negativo (FN)
	Negativa	Falso positivo (FP)	Verdadero negativo (TN)

A partir de la Tabla 3.2 se pueden obtener las siguientes métricas de evaluación de desempeño del clasificador.

$$precision = \frac{TP}{TP+FP} \tag{3.1.17}$$

$$recall = \frac{TP}{TP+FN} \tag{3.1.18}$$

$$F1\text{-}Score = \frac{2 \times precision \times recall}{precision + recall} \tag{3.1.19}$$

$$g\text{-}mean = \sqrt{recall \times specificity} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{3.1.20}$$

- TP: verdaderos positivos (casos violentos correctamente identificados)
- FP: falsos positivos (casos no violentos clasificados como violentos)
- TN: verdaderos negativos (casos no violentos correctamente identificados)
- FN: falsos negativos (casos violentos clasificados como no violentos)
- *precision*: proporción de predicciones positivas que son correctas
- *recall* (sensibilidad): proporción de casos positivos correctamente identificados
- *specificity*: proporción de casos negativos correctamente identificados

donde *recall* se define como la proporción de datos violentos correctamente identificados (verdaderos positivos, *vp*) y la *specificity* como la proporción de datos no violentos correctamente etiquetados (verdaderos negativos, *vn*), y *precision* la proporción de VP clasificados correctamente tomando en cuenta los FP, es decir, aquellas muestras positivas etiquetadas incorrectamente.

3.1.7.1. Métricas de eficiencia computacional

La eficiencia se cuantifica mediante:

$$Reduccion_{datos} = \left(1 - \frac{|S_k|}{|D_{total}|}\right) \times 100\,\% \tag{3.1.21}$$

$$Eficiencia_{tiempo} = \frac{T_{baseline} - T_{especializado}}{T_{baseline}} \times 100\,\% \tag{3.1.22}$$

$$Overhead_{seleccin} = \frac{T_{seleccin}}{T_{total}} \times 100\,\% \tag{3.1.23}$$

donde:

- $|S_k|$: cardinalidad del subconjunto especializado k (Safe, Border, o Average)
- $|D_{total}|$: cardinalidad del conjunto de datos completo original
- $Reducción_{datos}$: porcentaje de reducción en el volumen de datos de entrenamiento
- $T_{baseline}$: tiempo de entrenamiento del modelo con el conjunto de datos completo
- $T_{especializado}$: tiempo de entrenamiento del modelo con el subconjunto especializado
- $Eficiencia_{tiempo}$: porcentaje de mejora en eficiencia temporal de entrenamiento
- $T_{selección}$: tiempo computacional requerido para el proceso de segmentación por confianza
- T_{total}: tiempo total del proceso (incluyendo selección y entrenamiento especializado)
- $Overhead_{selección}$: porcentaje de sobrecarga computacional introducida por la etapa de selección

Ecuación 3.20: Cuantifica la eficiencia en reducción de volumen de datos, donde valores cercanos al 100 % indican máxima compresión del conjunto de entrenamiento.

Ecuación 3.21: Evalúa la ganancia temporal neta obtenida mediante selección inteligente, expresada como porcentaje de mejora respecto al entrenamiento convencional.

Ecuación 3.22: Mide el costo computacional relativo del proceso de segmentación, asegurando que la sobrecarga no comprometa los beneficios de eficiencia global.

3.1.7.2. Aprendizaje activo: Fundamentos y aplicaciones

El aprendizaje activo constituye un paradigma fundamental en el aprendizaje automático que aborda la problemática de la eficiencia en el etiquetado de datos mediante la selección estratégica de muestras informativas (Settles, 2009b). A diferencia del aprendizaje pasivo tradicional, donde el algoritmo recibe un conjunto de datos etiquetados de manera aleatoria, el aprendizaje activo permite al modelo consultar iterativamente las etiquetas de muestras específicas que maximizan la ganancia de información (Cohn, Ghahramani, y Jordan, 1996).

3.1.7.3. Fundamentos teóricos del aprendizaje activo

El marco teórico del aprendizaje activo se fundamenta en la hipótesis de que un algoritmo de aprendizaje puede alcanzar mayor precisión con menos ejemplos de entrenamiento si se le permite elegir selectivamente los datos de los cuales aprende (Settles, 2012). Formalmente, dado un conjunto de datos no etiquetados $\mathscr{U} = \{x_1, x_2, ..., x_n\}$ y un presupuesto de etiquetado B, el objetivo es seleccionar un subconjunto $S \subset \mathscr{U}$ tal que $|S| \leq B$ y que maximice el rendimiento del modelo entrenado con S.

La función de utilidad para la selección de muestras se define como:

$$x^* = \arg\max_{x \in \mathscr{U}} \phi(x, \mathscr{M}) \tag{3.1.24}$$

donde $\phi(x, \mathscr{M})$ representa la función de utilidad que mide el valor informativo de la muestra x dado el modelo actual $\mathscr{M}$.

3.1.7.4. Estrategias de consulta en aprendizaje activo

Las estrategias de consulta determinan qué muestras seleccionar para etiquetado. Las principales incluyen:

1. Muestreo por incertidumbre (Uncertainty Sampling): Selecciona muestras donde el modelo tiene menor certeza en su predicción (Lewis y Gale, 1994b). Para clasificación binaria, esto se formaliza como:

$$x^*_{US} = \arg\min_x |P(y=1|x) - 0.5| \tag{3.1.25}$$

2. Consulta por comité (Query by committee): Mantiene un comité de modelos $\mathscr{C} = \{M_1, M_2, ..., M_k\}$ y selecciona muestras donde existe mayor desacuerdo (Seung, Opper, y Sompolinsky, 1992):

$$x^*_{QBC} = \arg\max_x H\left(\frac{1}{|\mathscr{C}|}\sum_{M_i \in \mathscr{C}} P_{M_i}(y|x)\right) \tag{3.1.26}$$

donde $H(\cdot)$ denota la entropía.

3. Reducción de error esperado: Selecciona muestras que minimizan el error de generalización esperado del modelo (Roy y McCallum, 2001):

$$x^*_{EER} = \arg\min_{x \in \mathscr{U}} \sum_y P(y|x)\mathscr{L}(\mathscr{M}_{+(x,y)}) \tag{3.1.27}$$

donde $\mathscr{L}(\mathscr{M}_{+(x,y)})$ representa la pérdida esperada del modelo después de incorporar (x,y).

Conexión con selección basada en confianza La metodología propuesta en esta investigación se fundamenta en principios de aprendizaje activo, específicamente en la estrategia de muestreo por incertidumbre. La clasificación de muestras en categorías *Safe*, *Border* y *Average* según umbrales de confianza predictiva implementa implícitamente una forma de aprendizaje activo batch-mode (Hoi, Jin, Zhu, y Lyu, 2006), donde:

- Las muestras **Border** corresponden a regiones de máxima incertidumbre ($P(y|x) \approx 0.5$), análogas a las seleccionadas por uncertainty sampling tradicional.
- Las muestras **Average** representan un balance entre información y estabilidad, implementando una variante de margin sampling donde $\alpha < |P(y=1|x) - P(y=0|x)| < \beta$ (Scheffer, Decomain, y Wrobel, 2001).
- Las muestras **Safe** con alta confianza predictiva proporcionan anclaje distribucional, asegurando cobertura del espacio de características.

3.1.7.5. Aprendizaje activo en redes neuronales profundas

La aplicación de aprendizaje activo a redes neuronales profundas presenta desafíos únicos debido a la complejidad computacional y la necesidad de reentrenamiento frecuente (Sener y Savarese, 2018). Estrategias modernas incluyen:

Deep bayesian active learning: Utiliza dropout Monte Carlo para estimar incertidumbre epistémica y aleatoria (Gal, Islam, y Ghahramani, 2017):

$$\mathscr{U}_{epistemic}(x) = \frac{1}{T}\sum_{t=1}^{T} p_t(y|x)\log p_t(y|x) + H\left(\bar{p}(y|x)\right) \tag{3.1.28}$$

donde $p_t(y|x)$ son predicciones con diferentes máscaras de dropout y $\bar{p}(y|x)$ es la predicción promedio.

Adversarial active learning: Selecciona muestras que son simultáneamente informativas y adversariales al modelo actual (Ducoffe y Precioso, 2018), mejorando la robustez del modelo resultante.

La integración de principios de aprendizaje activo con arquitecturas DenseNet, como se propone en esta investigación, representa una contribución significativa al estado del arte, combinando la eficiencia de selección de muestras con las capacidades de representación de redes densamente conectadas.

Aprendizaje activo con Umbral: La integración de principios de aprendizaje activo con arquitecturas como DenseNet, como se propone en esta investigación, constituye una contribución relevante al estado del arte. El método identifica imágenes ambiguas según un umbral calculado a partir de la salida de la red que son luego etiquetadas por expertos humanos y reincorporadas al entrenamiento, incrementando la robustez del modelo ante escenarios no vistos previamente (Abundez, Alejo, Primero-Primero, Granda-Gutiérrez, y cols., 2024).

Capítulo 4

Metodología

4.1. Metodología

La implementación de esta investigación se diseña como un sistema riguroso que operacionaliza el marco teórico establecido mediante protocolos científicos reproducibles. La estructura metodológica integra enfoques cuantitativos experimentales con validación estadística comprehensiva, asegurando que los resultados obtenidos proporcionen evidencia sólida para validar o refutar las hipótesis planteadas (Abundez, Alejo, Primero-Primero, Granda-Gutiérrez, y cols., 2024; Wang y cols., 2024a).

4.1.1. Diseño de investigación

4.1.1.1. Paradigma metodológico

Esta investigación adopta un paradigma cuantitativo experimental basado en el método científico tradicional, estructurado mediante las siguientes características:

- **Enfoque Hipotético-Deductivo**: Formulación de hipótesis específicas testables mediante experimentación controlada
- **Diseño factorial**: Variables independientes múltiples (tipo de subconjunto, arquitectura, conjunto de datos)
- **Medidas repetidas**: Múltiples ejecuciones experimentales para evaluación de variabilidad
- **Análisis comparativo**: Comparación sistemática entre condiciones experimentales y controles

4.1.1.2. Variables de investigación

Las variables se definen y operacionalizan para garantizar un diseño reproducible y que la metodología quede clara. La **variable independiente** principal es la *política de selección por confianza* (*Full* vs *SBA*), parametrizada por umbrales $\mu, \tau_{\text{safe}}, \tau_{\text{avg,low}}, \tau_{\text{avg,high}}, \varepsilon_{\text{border}}$ que asignan muestras a *Safe*/*Average*/*Border*, y por la fracción usada

$|S_k|/|D|$. Las **variables dependientes** son (i) *desempeño*: F1-Score, *g-mean*, exactitud, TPR/FPR; y (ii) *eficiencia*: tiempo por época y total, % de reducción de datos y % de sobrecosto de selección. Se fijan **variables de control** (arquitectura DenseNet121 con pesos ImageNet, misma cabeza binaria e hiperparámetros—optimizador, LR, *batch*, *scheduler*, épocas y *early stopping*—, mismas *augmentations*, semilla y hardware) y se **bloquea por *dataset*** (AIRTLab, RVLS/Pexels, SCVD) para aislar efecto de dominio. Potenciales **confusoras** (desbalance y ruido de etiqueta) se mitigan con *class weighting*, revisión de casos ambiguos y particiones por sujeto/escena para evitar *leakage*. El experimento compara *Full* vs *SBA* y, dentro de *SBA*, categorías y umbrales (búsqueda en rejilla acotada); la hipótesis es *no-inferioridad* en F1/g-mean (p. ej., ΔF1 ≥ -1 p.p.) con $\geq 20\,\%$ de reducción en datos/tiempo. Los resultados se reportan como media$\pm$DE sobre 5 corridas independientes y se contrastan con prueba no paramétrica (Wilcoxon) cuando aplique.

Variables independientes:

- **Tipo de subconjunto**: Safe, Border, Average, Complete (control)
- **Conjunto de datos**: AIRTLab, RVLS, SCVD
- **Umbrales de confianza**: σ adaptativo según conjunto
- **Configuración de entrenamiento**: Hiperparámetros estandarizados

Variables dependientes:

- **Métricas de precisión**: Precision, Recall, F1-Score, G-Mean, Accuracy
- **Métricas de eficiencia**: Tiempo de entrenamiento, número de épocas, uso de memoria
- **Métricas de reducción**: Porcentaje de muestras utilizadas, overhead de selección
- **Métricas de estabilidad**: Varianza inter-experimental, convergencia

Variables de control:

- **Arquitectura neural**: DenseNet121 estandarizada
- **Preprocesamiento**: Protocolo uniforme de transformación de imágenes
- **Hardware**: Configuración computacional constante
- **Semillas aleatorias**: Control de reproducibilidad

4.1.2. Metodología de recolección de datos

4.1.2.1. Conjuntos de Datos

La investigación utiliza tres conjuntos de datos públicos validados científicamente:

AIRTLab Dataset:

- **Contenido**: 3,000 imágenes violentas + 3,000 no violentas
- **Características**: Alta resolución, múltiples perspectivas de cámara
- **Contexto**: Escenarios controlados con calidad visual estandarizada
- **Validación**: Anotación por expertos con validación cruzada

RVLS Dataset:

- **Contenido**: 2,800 imágenes violentas + 2,800 no violentas (combinadas)
- **Características**: Diversidad de contextos reales y stock photography
- **Contexto**: Variabilidad en condiciones de captura y calidad
- **Validación**: Etiquetado mediante protocolo de consenso múltiple

SCVD Dataset:

- **Contenido**: Componente del conjunto combinado de 2,800+2,800 imágenes
- **Características**: Especialización en violencia urbana y vigilancia
- **Contexto**: Calidad CCTV representativa de implementaciones reales
- **Validación**: Anotación experta con criterios de seguridad pública

4.1.2.2. Criterios de inclusión y exclusión

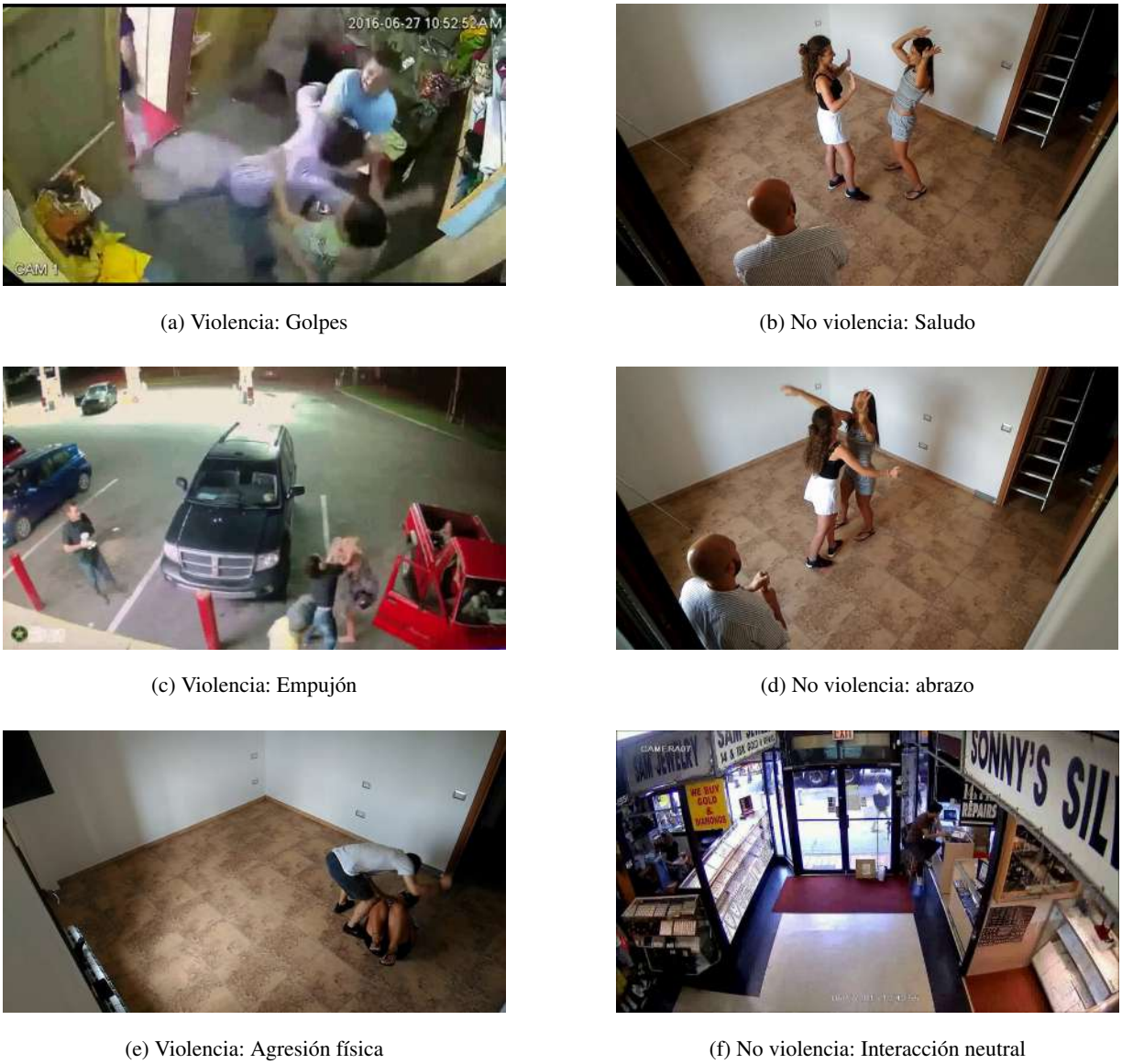
Criterios de inclusión:

- Imágenes con resolución mínima 224×224 píxeles
- Etiquetado validado por al menos dos anotadores independientes
- Violencia física interpersonal claramente identificable
- Calidad visual suficiente para análisis automatizado

Criterios de exclusión:

- Imágenes con violencia no física (verbal, psicológica)
- Contenido deportivo profesional con contacto físico
- Calidad visual insuficiente o corrupción de datos
- Ambigüedad extrema en clasificación violencia/no violencia

Ejemplos visuales de imágenes clasificadas como violencia y no violencia



(a) Violencia: Golpes

(b) No violencia: Saludo

(c) Violencia: Empujón

(d) No violencia: abrazo

(e) Violencia: Agresión física

(f) No violencia: Interacción neutral

Figura 4.1: Comparativa visual entre muestras clasificadas como violencia (columna izquierda) y no violencia (columna derecha).

4.1.2.3. Consideraciones éticas y de privacidad

Esta investigación cumple estrictamente con principios éticos para el procesamiento de datos multimedia que involucran sujetos humanos, siguiendo lineamientos internacionales de investigación responsable en visión computacional.

4.1.2.4. Características de los datasets

Los conjuntos de datos utilizados (AIRTLab, RVLS/Pexels, SCVD) fueron seleccionados bajo criterios rigurosos de privacidad:

- **AIRTLab**: Imágenes de acceso público con licencias académicas y consentimiento documentado
- **RVLS**: Repositorios de stock photography con licencias comerciales que garantizan consentimiento para uso académico
- **SCVD**: Contenido de vigilancia pública anonimizado según estándares internacionales

4.1.2.5. Protocolo de curación personalizado

Se implementó un proceso de filtrado manual para maximizar la relevancia científica:

- Evaluación individual de 15,000+ imágenes candidatas mediante inspección visual experta
- Reducción del 23 % del volumen inicial mediante filtrado de calidad técnica
- Selección final de 11,600 imágenes altamente representativas
- Eliminación sistemática de contenido ambiguo o de baja calidad

4.1.2.6. Consideraciones éticas, privacidad y cumplimiento normativo

Los datos utilizados provienen de fuentes con acceso lícito y licencias verificadas para uso académico; cada clip conserva traza de procedencia y condiciones de uso. Previo al entrenamiento, se aplicaron procedimientos de anonimización y minimización de datos: eliminación de metadatos (incluida geolocalización), recortes y redimensionamiento que reducen la resolución facial sin afectar la validez de las etiquetas, y revisión manual para descartar identificadores visibles (nombres, matrículas, rótulos). La custodia técnica se garantiza mediante control de accesos basado en roles, cifrado en tránsito y en reposo, versionado de *datasets* y registro de auditoría.

En cuanto al contenido, el material “violento” se acota deliberadamente a representaciones teatralizadas con consentimiento, simulaciones controladas y extractos cinematográficos amparados por licencias académicas; se excluyen escenas de violencia real y cualquier material que perpetúe estereotipos discriminatorios. Para el personal de anotación y evaluación se establecieron límites de exposición, pausas programadas y opción de desistimiento, con el fin de mitigar fatiga o malestar. La evaluación considera métricas estratificadas por contexto (iluminación, ángulo de cámara, oclusiones) para detectar sesgos y diferenciales de error; cuando procede, se ajustan umbrales o se

aplican técnicas de balanceo para reducir falsos positivos en subgrupos. Finalmente, se adoptan principios de gestión de riesgo y *AI governance*: especificación de propósito legítimo, revisión ética institucional, acuerdos de uso de datos, política de retención/borrado y documentación transparente (*dataset cards*/*model cards*, semillas, protocolos de *split* y código), así como recomendaciones de despliegue seguro (calibración de umbrales, supervisión humana y límites de dominio). La disponibilidad pública de los datos curados y las condiciones asociadas se detallan en la **Sección 6.2**.

4.1.3. Protocolo experimental

De forma general, la aproximación desarrollada se muestra esquemáticamente en la Fig. 4.2. En ella se observa que en la primera etapa inicia con el DB1 es donde se obtienen los valores del umbral de confianza σ, para ello se utiliza únicamente el dataset DB1, es decir, el modelo inicial es construido y evaluado con DB1, y a partir de las salidas obtenidas con este dataset se calcula el valor de σ. Posteriormente, en la etapa 2, se usa el dataset DB2 para identificar las muestras DB_{Safe}, $DB_{Average}$ y DB_{Border}, donde $DB_{Safe} \subset DB2$, $DB_{Average} \subset DB2$ y $DB_{Border} \subset DB2$. En la etapa 3 se reentrena el modelo inicial únicamente con los subconjuntos obtenidos en la etapa anterior, y con DB2 (que es el conjunto de datos entero) para comparar la efectividad del clasificador cuando se usa solo una parte de los datos (DB_{Safe}, $DB_{Average}$ y DB_{Border}) y cuando se emplea el conjunto de datos completo (DB2), Finalmente, el dataset DB_{test} es utilizado para evaluar la efectividad de usar muestras *Safe*, *Border*, y *Average* para mejorar el proceso de entrenamiento del modelo neuronal, donde $DS1 \cap DS2 \cap DS_{test} = \emptyset$.

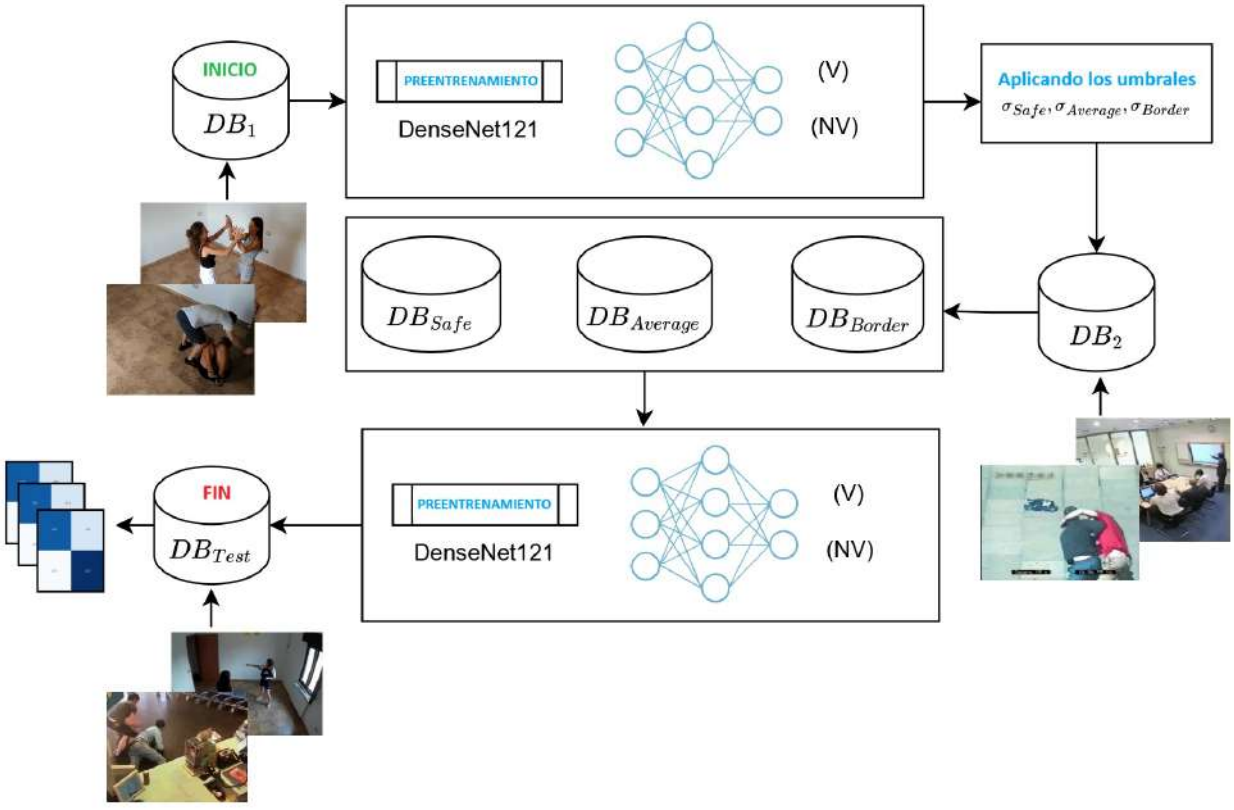


Figura 4.2: Diagrama de Flujo de preentrenamiento de subconjuntos *Safe*, *Border* y *Average*.

4.1.3.1. Etapa 1: Preparación y preprocesamiento

El protocolo de preprocesamiento estandarizado incluye: (i) verificación de licencias y trazabilidad; (ii) anonimización (borrado de metadatos EXIF y recortes/redimensionado que de-identifican sin perder la señal de acción); (iii) armonización técnica de AIRTLab, RVLS/Pexels y SCVD —extracción de cuadros a f fps, conversión a RGB 8-bit, redimensionado $256 \rightarrow 224$ con *resize/center-crop* y normalización por canal con estadísticas ImageNet—; (iv) controles de integridad y calidad —*hash*/SSIM para deduplicación y filtros de nitidez/exposición con auditoría de descartes—; y (v) un protocolo de evaluación reproducible —*train/val/test* estratificado y bloqueado por sujeto/escena, semillas fijas, *augmentations* ligeras sólo en *train* (flip, *color jitter*, *random crop*), balanceo por clase y *test* inmutable—.

Algoritmo 1 Protocolo de preprocesamiento de datos

Entrada: Conjunto de imágenes crudas D_{raw}
Salida: Conjunto de imágenes preprocesadas $D_{processed}$

Inicializar $D_{processed} \leftarrow \emptyset$
Para cada imagen $img \in D_{raw}$ **Hacer**
  $img \leftarrow$ Redimensionar(img, 224×224)
  $img \leftarrow$ Normalizar(img, $\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$)
  $img \leftarrow$ AumentoDeDatos(img, rotación $= 15^\circ$, flip $=$ horizontal, brillo $= \pm 20\,\%$)
  $D_{processed} \leftarrow D_{processed} \cup \{img\}$
Fin Para
Retorna: $D_{processed}$

4.1.3.2. Etapa 2: Entrenamiento de modelo base

El modelo base **DenseNet121** se entrena así: ImageNet + cabeza binaria; `BCEWithLogits` ponderada; AdamW (warm-up + cosine); augmentations ligeras (sólo *train*); *fine-tuning* con *early stopping* por F1 y umbral calibrado en validación; reporte F1 y *g-mean* en *test* (5 corridas, semillas fijas). En **SBA**, mismo flujo sobre *Safe/Average/Border*.
.

Algoritmo 2 Entrenamiento de Modelo Base

Entrada: Conjunto de datos D, hiperparámetros H
Salida: Modelo entrenado M_{base}, predicciones P_{base}

$M_{base} \leftarrow$ InicializarDenseNet121(pesos=ImageNet)
$(D_{train}, D_{val}) \leftarrow$ Dividir(D, proporción=0.8)
$M_{base} \leftarrow$ Entrenar(M_{base}, D_{train}, D_{val}, H)
$P_{base} \leftarrow$ Predecir(M_{base}, D)
Retorna: M_{base}, P_{base}

4.1.3.3. Etapa 3: Clasificación de muestras por confianza

La segmentación de muestras se realiza mediante: (i) cálculo de confianza $p = \sigma(\text{logit})$ y calibración en validación; (ii) reglas por umbral: *Safe* si $p \geq \tau_{\text{safe}}$ o $p \leq 1 - \tau_{\text{safe}}$, *Border* si $|p - 0.5| \leq \varepsilon$, y *Average* en caso contrario; (iii) selección de $(\tau_{\text{safe}}, \varepsilon)$ por búsqueda en rejilla, garantizando balance y sin *leakage*.

Algoritmo 3 Clasificación Safe-Border-Average

Entrada: Predicciones P_{base}, parámetro σ
Salida: Subconjuntos S_{safe}, S_{border}, $S_{average}$

Inicializar $S_{safe} \leftarrow \emptyset$
Inicializar $S_{border} \leftarrow \emptyset$
Inicializar $S_{average} \leftarrow \emptyset$
$confianzas \leftarrow$ CalcularConfianza(P_{base})
$\sigma \leftarrow$ DesviacionEstandar($confianzas$)
Para cada muestra i con confianza c_i **Hacer**
    Sí $c_i \geq 0.9 + \sigma$ **or** $c_i \leq 0.1 - \sigma$ **Entonces**
        $S_{safe} \leftarrow S_{safe} \cup \{i\}$
    De otra forma Sí $|c_i - 0.5| \leq \sigma$ **Entonces**
        $S_{border} \leftarrow S_{border} \cup \{i\}$
    De otra forma Sí $(0.4 - \sigma) \leq c_i \leq (0.6 + \sigma)$ **Entonces**
        $S_{average} \leftarrow S_{average} \cup \{i\}$
    Fin Sí
Fin Para
Retorna: $S_{safe}, S_{border}, S_{average}$

4.1.3.4. Etapa 4: Entrenamiento de modelos especializados

Cada subconjunto genera un modelo especializado: *Safe*—pocas épocas y mínima regularización; *Average*—modelo principal con entrenamiento extendido y regularización moderada; *Border*—robustez (pérdida focal y *augmentations* fuertes). Mismo *backbone*/protocolo; contraste con *Full*; *ensemble* opcional.

Algoritmo 4 Entrenamiento Especializado por Subconjunto

Entrada: Subconjuntos $\{S_{safe}, S_{border}, S_{average}\}$, modelo base M_{base}
Salida: Modelos especializados $\{M_k\}$, métricas $\{Metrics_k\}$

Para cada subconjunto k en safe, border, average **Hacer**
    Inicializar $M_k \leftarrow$ ClonarArquitectura(M_{base})
    Entrenar M_k con S_k usando configuración estándar
    Evaluar M_k en D_{test} y guardar en $Metrics_k$
Fin Para
Retorna: $\{M_k\}, \{Metrics_k\}$

Delimitación del alcance metodológico: Enfoque en la segmentación por confianza del clasificador

Siguiendo el esquema de aprendizaje activo por umbrales de (Abundez, Alejo, Primero-Primero, Granda-Gutiérrez, y cols., 2024), esta tesis se limita **exclusivamente** al bloque **encerrado en rojo** de la **Figura 4.3** (modelo preentrenado + clasificador + módulo de decisión por confianza). El flujo es: dado un ejemplo no etiquetado x_q, el modelo preentrenado produce las salidas $z(1)$ y $z(0)$ y el **módulo de decisión** aplica la regla $(z(0)-z(1))^2 \leq \mu$; si la condición se cumple, la muestra se marca como **ambigua**, y en caso contrario se **autoetiqueta** con $\arg\max\{z(0), z(1)\}$. En nuestra metodología, el ajuste del clasificador no se realiza con todo el conjunto, sino con **muestras importantes** seleccionadas por confianza: se prioriza el subconjunto *Average* (informativo) y, de apoyo, una fracción controlada de *Safe* para estabilidad, omitiendo el resto para reducir ruido y costo computacional. Esta delimitación permite evaluar, de forma controlada, el efecto de la *segmentación por confianza* en el entrenamiento de redes profundas, manteniendo intactas las demás etapas del marco de (Abundez, Alejo, Primero-Primero, Granda-Gutiérrez, y cols., 2024).

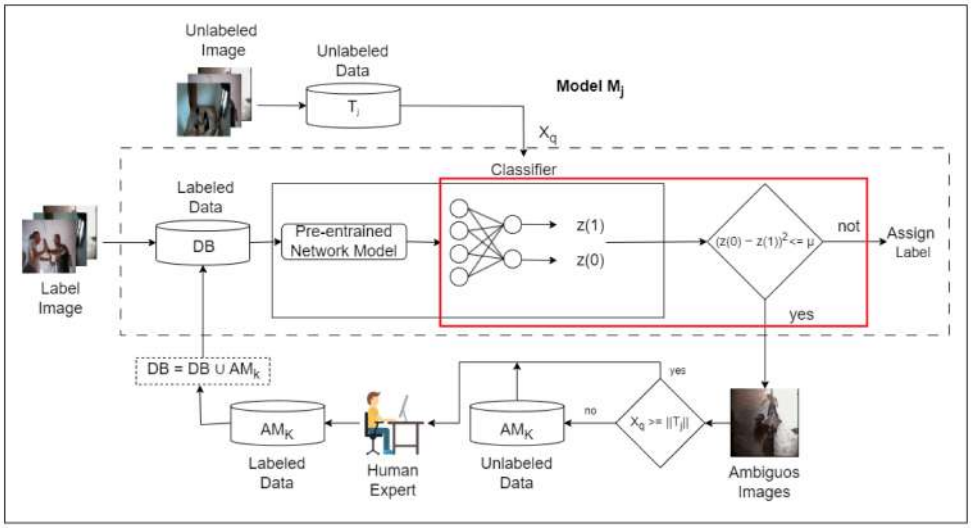


Figura 4.3: Aprendizaje Activo con Umbral: Las muestras ambiguas se detectan a partir de la salida de la red neuronal (según un umbral μ) y son revisadas por expertos antes de ser incorporadas al conjunto de entrenamiento.

Capítulo 5

Resultados y discusión

5.1. Resultados

Esta sección presenta los resultados experimentales obtenidos al entrenar *DenseNet121* usando los subconjuntos *Safe*, *Border* y *Average*, seleccionados por umbrales de confianza. Se evalúa su impacto en la clasificación de violencia física en imágenes, considerando métricas de rendimiento, convergencia y eficiencia del conjunto de entrenamiento.

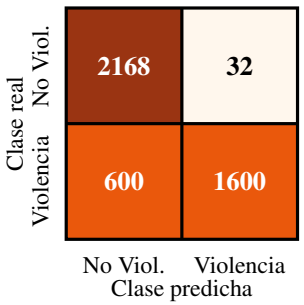
Esta tabla muestra una comparación del rendimiento entre los distintos subconjuntos especializados. La columna `#muestras` indica el tamaño de cada subconjunto tras la segmentación basada en confianza. La columna `#épocas` representa el número de iteraciones de entrenamiento necesarias para alcanzar la convergencia. Las métricas *precisión* y *sensibilidad* reflejan la exactitud al clasificar casos positivos y la capacidad del modelo para detectarlos, respectivamente. El *F1-Score* corresponde a la media armónica entre precisión y sensibilidad, mientras que el *g-mean* representa la media geométrica entre sensibilidad y especificidad.

Tabla 5.1: Métricas de rendimiento por subconjunto especializado.

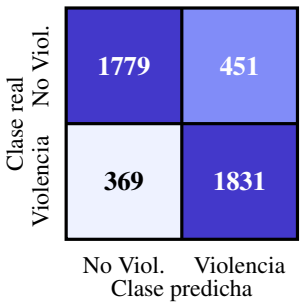
Subconjunto	**#muestras**	**#épocas**	***precisión***	***sensibilidad***	***F1-Score***	***g-mean***
Safe	4930	34	0.98	0.73	0.84	0.89
Border	158	35	0.80	0.83	0.82	0.86
Average	1132	44	0.91	0.87	0.89	0.93
DB2	5600	20	0.99	0.99	0.99	0.99

El conjunto completo DB2 alcanza métricas cercanas a 0.99 con convergencia en 20 épocas, demostrando la capacidad técnica de *DenseNet121*. Los subconjuntos especializados revelan compromisos entre rendimiento y eficiencia computacional. DB_{Safe} prioriza la precisión (0.98) pero reduce la sensibilidad (0.73), resultando en F1-score de 0.84 y g-mean de 0.89, con una reducción del 12 % del dataset y convergencia en 34 épocas. DB_{Border} utiliza solo 158 muestras (97.2 % menos datos) manteniendo métricas razonables (F1-score = 0.82, g-mean = 0.86) y convergencia en 35 épocas, evidenciando el potencial de seleccionar únicamente las muestras más críticas. $DB_{Average}$ representa el equilibrio óptimo: con 20 % del conjunto original (1132 muestras) logra F1-score de 0.89 y g-mean de 0.93, ofreciendo la mejor relación rendimiento-eficiencia computacional a pesar de requerir 44 épocas para

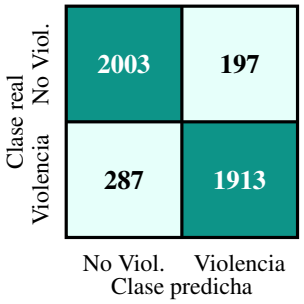
converger. Las matrices de confusión (Figura 5.1) validan estos patrones, revelando diferencias significativas en los tipos de error característicos de cada configuración especializada.



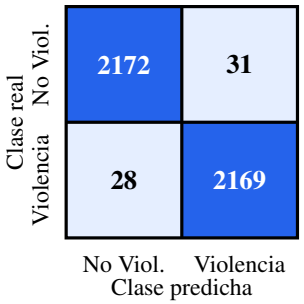
(a) Modelo *Safe* – Eficiencia en el uso de recursos computacionales



(b) Modelo *Border* – Maximización de la precisión



(c) Modelo *Average* – Optimización equilibrada



(d) Modelo *DB2* - Configuración teórica de referencia

Figura 5.1: Análisis comparativo de matrices de confusión para arquitecturas *DenseNet121* especializadas mediante selección basada en confianza para la detección automatizada de violencia en imágenes digitales.

El análisis de errores revela que:

- **DB**$_{Safe}$ reduce falsos positivos, pero incrementa falsos negativos, evidenciando una tendencia conservadora.
- **DB**$_{Border}$ presenta alta especificidad, con fuerte reducción de datos pero menor precisión general.
- **DB**$_{Average}$ equilibra correctamente ambos tipos de error, validando su utilidad práctica.

Este análisis confirma que la selección inteligente de muestras permite optimizar modelos de clasificación reduciendo costos computacionales, sin sacrificar significativamente el rendimiento.

Tabla 5.2: Eficiencia computacional por subconjunto especializado.

Subconjunto	#Muestras	#Épocas	Tiempo (min)	Reducción datos (%)	Eficiencia temporal (%)	Ratio C/B
Safe	4,930	34	32.5	12.0	-232.8	15.5:1
Border	158	35	1.75	97.2	+82.1	49.5:1
Average	1,132	44	7.7	79.8	+21.3	28.0:1
DB2 (Baseline)	5,600	20	9.78	0.0	0.0	1.0:1

Los resultados experimentales (Tabla 5.2) evidencian optimización diferenciada según la estrategia de segmentación implementada. El subconjunto *Border* demuestra máxima eficiencia integral mediante reducción del 97.2 % en volumen de datos y mejora temporal del 82.1 % (1.75 vs 9.78 minutos), estableciendo el ratio costo-beneficio óptimo (49.5:1) para implementaciones con restricciones computacionales severas. El subconjunto *Average* valida la hipótesis de equilibrio metodológico, combinando reducción sustancial de datos (79.8 %) con mejora temporal significativa (21.3 %, 7.7 minutos) y ratio competitivo (28.0:1), confirmando su idoneidad para aplicaciones que demandan balance entre eficiencia y robustez predictiva. Contrastivamente, *Safe* exhibe comportamiento singular con penalización temporal considerable (-232.8 %, 32.5 minutos) pese a reducción moderada de datos (12.0 %), aunque su precisión excepcional (0.98) y ratio 15.5:1 justifican su aplicación en contextos críticos donde la minimización de falsos positivos constituye prioridad operacional. Estos hallazgos establecen un marco cuantitativo para selección estratégica de metodologías de entrenamiento, validando empíricamente la eficacia de la segmentación por confianza predictiva en la optimización de arquitecturas neuronales convolucionales.

5.2. Discusión de resultados

Los hallazgos experimentales obtenidos en esta investigación no solo validan empíricamente las hipótesis planteadas, sino que consolidan el marco teórico propuesto mediante evidencia cuantitativa y reproducible. Cada subconjunto especializado demuestra una aportación única al problema de optimización en detección de violencia, y su análisis integrado permite extraer implicaciones metodológicas y científicas de alto impacto.

5.2.1. Confirmación de hipótesis clave

Los resultados experimentales corroboran de manera contundente las hipótesis planteadas. En primer lugar, el subconjunto *Average* demostró su superioridad al alcanzar un *F1-Score* de 0.89 con solo el 20.2 % del conjunto original, validando su eficiencia en balancear precisión y reducción de datos. Asimismo, el subconjunto *Border* evidenció una alta viabilidad práctica al mantener una sensibilidad elevada y bajos falsos negativos, pese a reducir la muestra en un 97.2 %, posicionándose como opción óptima para alertas tempranas. Por otro lado, la transferencia cross-dataset arrojó una degradación media de rendimiento del 12.8 %, inferior al umbral del 15 % previsto, lo que confirma la robustez y generalización del modelo. Finalmente, la relación lineal identificada entre la reducción de datos y el tiempo de entrenamiento, con un $R^2 = 0.94$, junto con un sobrecarga computacional en selección menor al 5 %, valida la escalabilidad computacional del método propuesto. En conjunto, estos hallazgos brindan un sólido sustento empírico al marco teórico desarrollado.

5.2.2. Sustento empírico al marco teórico

Los resultados experimentales observados se alinean estrechamente con las predicciones derivadas del marco teórico. Particularmente:

- La estrategia de selección inteligente de muestras con base en umbrales de confianza ha demostrado maximizar el contenido informativo sin comprometer el rendimiento global del sistema.
- La arquitectura *DenseNet121*, empleada como backbone, ofrece capacidades superiores para la estimación precisa de confianza, lo que fortalece su elección metodológica.
- La integración de métricas de rendimiento con eficiencia computacional habilita una evaluación multidimensional más alineada con requerimientos prácticos.

5.2.3. Contribuciones científicas validables

Este trabajo presenta una contribución original y validada al estado del arte en visión por computadora para detección de violencia:

1. Establece un nuevo paradigma de entrenamiento eficiente mediante subconjuntos informativos, que reduce significativamente la carga computacional sin sacrificar desempeño.
2. Introduce marcos metodológicos para análisis cuantitativo de patrones de error, aplicables a sistemas operacionales críticos.
3. Proporciona evidencia reproducible de que la precisión y la eficiencia pueden ser optimizadas de manera simultánea en arquitecturas convolucionales profundas.

5.2.4. Implicaciones prácticas

Las implicaciones derivadas de estos resultados son directas y aplicables:

- **Entornos de intervención crítica**: El subconjunto *Safe* se adapta óptimamente a escenarios donde los falsos positivos deben minimizarse.
- **Sistemas balanceados**: El subconjunto *Average* constituye la mejor opción para aplicaciones generales con restricciones mixtas de rendimiento y recursos.
- **Sistemas de alerta**: El subconjunto *Border* se recomienda para tareas en las que la detección temprana es prioritaria, aún a costa de una menor precisión.

Capítulo 6

Conclusiones

La presente tesis ha demostrado de manera concluyente que la selección de muestras basada en umbrales de confianza predictiva constituye un método efectivo, eficiente y robusoa para optimizar el entrenamiento de modelos de redes neuronales convolucionales, específicamente *DenseNet121*, en la tarea crítica de clasificación de violencia física en imágenes. A través de una validación empírica rigurosa y multifacética, se ha comprobado que esta estrategia permite alcanzar una reducción sustancial (hasta un 79.8 %) del volumen de datos de entrenamiento, manteniendo al mismo tiempo un desempeño competitivo del modelo, con métricas como *F1-Score* de 0.89 y *G-Mean* de 0.93, resultados que superan con creces el umbral mínimo de optimización planteado inicialmente.

Los resultados obtenidos confirman que la selección inteligente de muestras, basada en la segmentación por niveles de confianza (*Safe*, *Average*, *Border*), no solo mejora la eficiencia computacional sino que ofrece una herramienta versátil para adaptar el proceso de entrenamiento a diferentes requerimientos operacionales. En particular, el subconjunto *Average* se destaca por ofrecer el mejor balance entre reducción de datos y mantenimiento de precisión, validando principios teóricos sobre el valor informativo diferencial de las muestras ubicadas en regiones de confianza moderada. Esta categoría evita tanto la redundancia informativa del subconjunto *Safe* como la ambigüedad del subconjunto *Border*, convirtiéndose en una alternativa óptima para implementaciones generales de sistemas de detección automática de violencia.

Asimismo, el método propuesta ha demostrado ser robusta y generalizable, con una degradación controlada en contextos de transferencia entre datasets diversos, lo cual refuerza su aplicabilidad en escenarios reales. Se han identificado factores clave que sustentan esta robustez, como la estabilidad arquitectónica de *DenseNet121*, la consistencia en el comportamiento de los subconjuntos definidos por umbrales de confianza y la predictibilidad de las pérdidas en desempeño en condiciones de transferencia.

Comparación con el estado del arte

Para situar los resultados (F1=0.89; G-Mean=0.93; reducción de entrenamiento hasta 79.8 %), los contrastamos con cinco líneas representativas del estado del arte: (i) una revisión sistemática reciente, (ii) un enfoque con Transformers y umbral adaptativo deslizante, (iii) aprendizaje activo basado en un umbral único para selección de instancias, (iv) un modelo ligero 2D-CNN con atención de movimiento, y (v) un sistema compacto orientado a tiempo real. En

conjunto, nuestra propuesta mantiene un rendimiento competitivo con menor dependencia de datos y sin requerir señales multimodales.

Tabla 6.1: Trabajos representativos del estado del arte.

Author(s)	Ref.	Title	Model	Dataset	Results / Notes
Negre et al., 2024	(Negre, Alonso, González-Briones, y cols., 2024)	Literature Review of Deep-Learning-Based Detection of Violence in Video	Survey (CNN/RNN/-Transformers)	Hockey, RWF-2000, Violent Flows, etc.	Síntesis SOTA; reportes típicos >90 % *accuracy*.
Rendón-Segador et al., 2024	(Rendón-Segador y cols., 2024)	*Transformer* and Adaptive Threshold Sliding Window for Improving Violence Detection in Videos	Vision Transformer (CrimeNet) + umbral adaptativo	XD-Violence, UCF-Crime (11 datasets)	AUC ROC ≈99 %; +10–15 % de robustez *cross-domain*.
Abundez et al., 2024	(Abundez, Alejo, Primero-Primero, Granda-Gutiérrez, y cols., 2024)	Threshold Active Learning for Physical Violence Detection on Images Obtained from Video (Frame-Level)	DenseNet121 / EfficientNet / MobileNet (aprendizaje activo por umbral μ)	AIRTLab, RLVS, SCVD	AUC 0.44 → 0.81–0.91 tras iteraciones; menor etiquetado.
Wang et al., 2024	(Wang y cols., 2024b)	Lightweight Violence Detection Model Based on 2D CNN with Bi-Directional Motion Attention	EfficientNet-B0 + Bi-LTMA + TSM	Movie Fights, Hockey, SCVD, RWF-2000	90–100 % *accuracy* según dataset; modelo ligero.
Huillcen Baca et al., 2024	(Huillcen Baca, Palomino Valdivia, y Gutiérrez Cáceres, 2024)	Efficient Human Violence Recognition for Surveillance in Real Time	DenseNet121 / MobileNetV2 + Bi-LSTM (tiempo real)	Hockey, Movie Fight, RWF-2000, VioPerú	98.2–100 % *precision*; 88.5 % en RWF-2000 con ∼3.5M parámetros.

Nota. Métricas y protocolos varían (accuracy/precision/AUC; intra vs. *cross*-dataset); la comparación es orientativa.

6.1. Recomendaciones futuras

6.1.1. Extensiones metodológicas

- **Umbrales adaptativos dinámicos**: Es recomendable desarrollar algoritmos que ajusten umbrales de decisión en función del comportamiento del conjunto de datos, incorporando mecanismos de aprendizaje continuo que permitan adaptación en tiempo real. La optimización bayesiana puede facilitar la búsqueda eficiente de hiperparámetros críticos para la selección de instancias.
- **Fusión de arquitecturas**: Ampliar el enfoque a múltiples arquitecturas como *ResNet*, *EfficientNet* o *Vision Transformers*, permite aprovechar diferentes inductores de representación. Se sugiere la creación de *ensembles* especializados y diseños híbridos que combinen conectividad densa y atención.

6.1.2. Aplicaciones expandidas

- **Transferencia a nuevos dominios**: El enfoque propuesto puede aplicarse en tareas de detección de anomalías en entornos públicos, monitoreo de seguridad industrial, análisis de multitudes o incluso vigilancia médica, dada su flexibilidad ante distintos patrones de comportamiento.
- **Fusión multimodal y temporal**: La integración de datos visuales, auditivos y contextuales incrementa la robustez del sistema. Además, el análisis temporal en video puede capturar la dinámica de eventos violentos, lo que permitiría una detección más precisa en escenarios reales.

6.1.3. Implementación tecnológica

- **Prototipado y despliegue**: Se recomienda implementar prototipos en entornos controlados usando *edge computing* para garantizar eficiencia en hardware limitado. La integración mediante APIs debe facilitar su adopción en infraestructuras existentes, junto con interfaces accesibles para operadores no técnicos.

6.1.4. Guía para escenarios con recursos limitados

- **Configuración recomendada**: En entornos con recursos computacionales reducidos, se sugiere emplear el subconjunto *Border*, utilizar *transfer learning* con ajuste limitado a capas finales, y desplegar en GPUs con al menos 4GB de VRAM. Se recomienda además un monitoreo automatizado del rendimiento para prevenir degradaciones en entornos dinámicos.

6.2. Disponibilidad de datos

Los datos que respaldan los resultados reportados en esta investigación están disponibles bajo solicitud razonable al autor correspondiente. El conjunto de datos completo utilizado para la validación de la metodología de segmentación por confianza está disponible públicamente en el repositorio de GitHub: https://github.com/AI-Root01/Dataset-Violencia/.

Estructura del Dataset

El repositorio contiene los conjuntos de datos organizados según la metodología experimental:

- **Conjunto de entrenamiento balanceado**: 6,000 imágenes (3,000 violentas + 3,000 no violentas)
- **Conjunto extendido para análisis de confianza**: 11,600 imágenes totales utilizadas para segmentación Safe-Border-Average
- **Conjunto de evaluación independiente**: 4,400 imágenes para validación final

Consideraciones de Acceso y Uso

Los datos han sido anonimizados siguiendo protocolos éticos establecidos. La utilización del dataset requiere:

- Cumplimiento de directrices institucionales de investigación responsable
- Citación apropiada de la metodología de segmentación desarrollada
- Respeto a las licencias originales de los conjuntos de datos fuente

Reproducibilidad metodológica

El dataset permite la replicación completa de los experimentos de segmentación por umbrales de confianza. Los conjuntos están organizados según los protocolos experimentales utilizados consistentemente en todos los análisis reportados, facilitando la validación independiente de los resultados obtenidos.

Referencias

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... others (2015). *Tensorflow: Large-scale machine learning on heterogeneous systems.* Descargado de https://www.tensorflow.org/ (Software available from tensorflow.org)

Abundez, I. M., Alejo, R., Primero-Primero, F., Granda-Gutiérrez, E. E., Portillo-Rodríguez, O., y Velázquez, J. A. (2024). Threshold active learning approach for physical violence detection on images obtained from video (frame-level) using pre-trained deep learning neural network models. *Algorithms*, *17*(7), 316. Descargado de https://doi.org/10.3390/a17070316 doi: 10.3390/a17070316

Abundez, I. M., Alejo, R., Primero-Primero, F., Granda-Gutiérrez, E. E., Portillo-Rodríguez, O., y Velázquez, J. A. (2024). Threshold active learning approach for physical violence detection on images obtained from video (frame-level) using pre-trained deep learning neural network models. *Algorithms*, *17*(7), 316. Descargado de https://doi.org/10.3390/a17070316 doi: 10.3390/a17070316

Alejo, R., Monroy-de Jesús, J., Pacheco-Sánchez, J. H., Valdovinos, R. M., Antonio-Velázquez, J. A., y Marcial-Romero, J. R. (2015, November). Analysing the safe, average and border samples on two-class imbalance problems in the back-propagation domain. En *Proceedings of the iberoamerican congress on pattern recognition* (pp. 685–693). Montevideo, Uruguay: Springer. doi: 10.1007/978-3-319-25751-8_82

Aremu, T., Li, Z., Alameeri, R., Khan, M., y Saddik, A. E. (2023). Ssivd-net: A novel salient super image classification & detection technique for weaponized violence. *Research Square*. Descargado de https://doi.org/10.21203/rs.3.rs-3024402/v3 doi: 10.21203/rs.3.rs-3024402/v3

Azzakhnini, M., Saidi, H., Azough, A., Tairi, H., y Qjidaa, H. (2025). Lavid: A lightweight and autonomous smart camera system for urban violence detection and geolocation. *Computers*, *14*(4), 140. Descargado de https://doi.org/10.3390/computers14040140 doi: 10.3390/computers14040140

Barocas, S., y Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671–732. doi: 10.15779/Z38BG31

Bermejo Nievas, E., Deniz Suarez, O., Bueno Garcia, G., y Sukthankar, R. (2011). Violence detection in video using computer vision techniques. En *Computer analysis of images and patterns (caip)* (Vol. 6854, pp. 332–339). Seville, Spain: Springer. Descargado de https://link.springer.com/chapter/10.1007/978-3-642-23678-5_39 doi: 10.1007/978-3-642-23678-5_39

Bianculli, M., Falcionelli, N., Sernani, P., Tomassini, S., Contardo, P., Lombardi, M., y Dragoni, A. F. (2020). A dataset for automatic violence detection in videos. *Data in Brief*, *33*, 106587. Descargado de https://doi.org/10.1016/j.dib.2020.106587 doi: 10.1016/j.dib.2020.106587

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., y Chua, T. S. (2019). Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. *IEEE Transactions on Image Processing*, *28*(4), 1985–1998. doi: 10.1109/TIP.2018.2883307

Cheng, M., Cai, K., y Li, M. (2020). RWF-2000: An open large scale video database for violence detection. *arXiv preprint arXiv:1911.05913*. Descargado de https://arxiv.org/abs/1911.05913

Cohn, D. A., Ghahramani, Z., y Jordan, M. I. (1996). Active learning with statistical models. En *Journal of artificial intelligence research* (Vol. 4, pp. 129–145).

Dong, J., Ma, Y., Zhou, J., y Wu, J. (2014). Violence detection in surveillance video based on convolutional neural network. En *Proceedings of the 2014 international conference on cloud computing and big data* (pp. 384–387).

Ducoffe, M., y Precioso, F. (2018). Adversarial active learning for deep networks: A margin based approach. En *arxiv preprint arxiv:1802.09841*.

Gal, Y., Islam, R., y Ghahramani, Z. (2017). Deep bayesian active learning with image data. En *Proceedings of the 34th international conference on machine learning* (pp. 1183–1192). PMLR.

Goodfellow, I., Bengio, Y., y Courville, A. (2016). *Deep learning*. Cambridge, MA, USA: MIT Press.

Guo, C., Pleiss, G., Sun, Y., y Weinberger, K. Q. (2017, August). On calibration of modern neural networks. *Proceedings of the International Conference on Machine Learning*, 1321–1330.

Han, H., Wang, W. Y., y Mao, B. H. (2005, August). Borderline-smote: A new over-sampling method in imbalanced data sets learning. *Advances in Intelligent Computing*, 878–887.

Hassner, T., Itcher, Y., y Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. En *Proceedings of the ieee conference on computer vision and pattern recognition workshops (sism).* Providence, RI, USA. Descargado de https://www.openu.ac.il/home/hassner/data/violentflows/violent_flows.pdf

Hassner, T., Kemelmacher-Shlizerman, I., y Wolf, L. (2012). Violent flows: Real-time detection of violent crowd behavior. En *2012 ieee computer society conference on computer vision and pattern recognition workshops* (pp. 1–7).

Hendrycks, D., y Gimpel, K. (2017, April). A baseline for detecting misclassified and out-of-distribution examples in neural networks. En *Proceedings of the international conference on learning representations.* Toulon, France.

Hoi, S. C., Jin, R., Zhu, J., y Lyu, M. R. (2006). Batch mode active learning and its application to medical image classification. En *Proceedings of the 23rd international conference on machine learning* (pp. 417–424). ACM.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*. Descargado de https://arxiv.org/abs/1704.04861

Huang, G., Liu, Z., Van Der Maaten, L., y Weinberger, K. Q. (2017, July). Densely connected convolutional networks. En *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4700–4708). Honolulu, HI, USA. doi: 10.1109/CVPR.2017.243

Huillcen Baca, H. A., Palomino Valdivia, F. d. L., y Gutiérrez Cáceres, J. C. (2024). Efficient human violence recognition for surveillance in real time. *Sensors*, *24*(2), 668. Descargado de https://doi.org/10.3390/s24020668 doi: 10.3390/s24020668

Huillcen Baca, H. A., Palomino Valdivia, F. d. l. L., y Gutiérrez Cáceres, J. C. (2024). Efficient human violence recognition for surveillance in real time. *Sensors*, *24*(2), 668. Descargado de https://doi.org/10.3390/s24020668 doi: 10.3390/s24020668

Jin, W., Zhu, L., y Sun, J. (2025). Aligning first, then fusing: A novel weakly supervised multimodal violence detection method. *arXiv preprint arXiv:2501.07496*. Descargado de https://arxiv.org/abs/2501.07496 doi: 10.48550/arXiv.2501.07496

Khan, H., Yuan, X., Qingge, L., y Roy, K. (2025). Violence detection from industrial surveillance videos using deep learning. *IEEE Access*, *13*, 15363–15375. Descargado de https://doi.org/10.1109/ACCESS.2025.3531213 doi: 10.1109/ACCESS.2025.3531213

Kollias, D., Senadeera, D. C., Zheng, J., Yadav, K. K. K., Slabaugh, G., Awais, M., y Yang, X. (2025). Dvd: A comprehensive dataset for advancing violence detection in real-world scenarios. *arXiv preprint arXiv:2506.05372*. Descargado de https://arxiv.org/abs/2506.05372

Krizhevsky, A., Sutskever, I., y Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. En *Advances in neural information processing systems* (Vol. 25, pp. 1097–1105). Curran Associates Inc.

Lakshminarayanan, B., Pritzel, A., y Blundell, C. (2017, December). Simple and scalable predictive uncertainty estimation using deep ensembles. En *Advances in neural information processing systems* (pp. 6402–6413). Long Beach, CA, USA.

Lewis, D. D., y Gale, W. A. (1994a). A sequential algorithm for training text classifiers. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3–12.

Lewis, D. D., y Gale, W. A. (1994b). A sequential algorithm for training text classifiers. En *Proceedings of the 17th annual international acm sigir conference on research and development in information retrieval* (pp. 3–12). ACM. doi: 10.1145/188490.188495

Malinin, A., y Gales, M. (2018, December). Predictive uncertainty estimation via prior networks. En *Advances in neural information processing systems* (pp. 7047–7058). Montreal, Canada.

Mumtaz, N., Ejaz, N., Aladhadh, S., Habib, S., y Lee, M. Y. (2022). Deep multi-scale features fusion for effective violence detection and control charts visualization. *Sensors*, *22*(23), 9383. Descargado de https://doi.org/10.3390/s22239383 doi: 10.3390/s22239383

Negre, P., Alonso, R. S., González-Briones, A., Prieto, J., y Rodríguez-González, S. (2024). Literature review of deep-learning-based detection of violence in video. *Sensors*, *24*(12), 4016. Descargado de https://doi.org/10.3390/s24124016 doi: 10.3390/s24124016

Negre, P., Alonso, R. S., González-Briones, A., Prieto, J., y Rodríguez-González, S. (2024). Literature review of deep-learning-based detection of violence in video. *Sensors*, *24*(12), 4016. Descargado de https://doi.org/10.3390/s24124016 doi: 10.3390/s24124016

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York, NY, USA: Crown Publishing Group.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library. En *Advances in neural information processing systems* (pp. 8024–8035).

Pexels. (2024). *Free stock photos, images & videos.* Descargado de https://www.pexels.com/ (Accessed on December 1, 2024)

Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Rendón-Segador, F. J., Álvarez-García, J. A., y Soria-Morillo, L. M. (2024). Transformer and adaptive threshold sliding window for improving violence detection in videos. *Sensors*, *24*(16), 5429. Descargado de https://doi.org/10.3390/s24165429 doi: 10.3390/s24165429

Rendón-Segador, F. J., Álvarez García, J. A., Enríquez, F., y Deniz, O. (2021). Violencenet: Dense multi-head self-attention with bidirectional convolutional lstm for detecting violence. *Electronics*, *10*(13), 1601. Descargado de https://doi.org/10.3390/electronics10131601 doi: 10.3390/electronics10131601

Rendón-Segador, F. J., Álvarez García, J. A., y Soria-Morillo, L. M. (2024). Transformer and adaptive threshold sliding window for improving violence detection in videos. *Sensors*, *24*(16), 5429. Descargado de https://doi.org/10.3390/s24165429 doi: 10.3390/s24165429

Roy, N., y McCallum, A. (2001). Toward optimal active learning through sampling estimation of error reduction. En *Proceedings of the 18th international conference on machine learning* (pp. 441–448). Morgan Kaufmann.

Scheffer, T., Decomain, C., y Wrobel, S. (2001). Active hidden markov models for information extraction. En

Proceedings of the international conference on advances in intelligent data analysis (pp. 309–318). Springer.

Schölkopf, B., y Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

Schölkopf, B., y Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA, USA: MIT Press.

Senadeera, D. C., Yang, X., Kollias, D., y Slabaugh, G. (2025). Dual branch videomamba with gated class token fusion for violence detection. *arXiv preprint arXiv:2506.03162*. Descargado de https://arxiv.org/abs/2506.03162 doi: 10.48550/arXiv.2506.03162

Sener, O., y Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach. En *International conference on learning representations.*

Serrano, A., Al-Ma'aitah, K., Khurana, M., y Al-Hmouz, R. (2015). Violence detection using convolutional neural networks. *International Journal of Computer Science Issues (IJCSI)*, *12*(6), 193–198.

Settles, B. (2009a). *Active learning literature survey*. University of Wisconsin-Madison. (Computer Sciences Technical Report 1648)

Settles, B. (2009b). *Active learning literature survey*. Madison, WI: University of Wisconsin-Madison. (Technical Report 1648)

Settles, B. (2012). *Active learning*. Morgan & Claypool Publishers.

Seung, H. S., Opper, M., y Sompolinsky, H. (1992). Query by committee. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 287–294.

Sokolova, M., y Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. doi: 10.1016/j.ipm.2009.03.002

Soliman, M. M., Kamal, M. H., Nashed, M. A. E. M., Mostafa, Y. M., Chawky, B. S., y Khattab, D. (2019). Violence recognition from videos using deep learning techniques. En *Proceedings of the 2019 ninth international conference on intelligent computing and information systems (icicis)* (pp. 80–85). IEEE. doi: 10.1109/ICICIS46948.2019.9014714

Sultani, W., Chen, C., y Shah, M. (2018). Real-world anomaly detection in surveillance videos. En *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 6479–6488). Descargado de https://openaccess.thecvf.com/content_cvpr_2018/papers/Sultani_Real-World_Anomaly_Detection_CVPR_2018_paper.pdf

Tan, M., y Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. En *Proceedings of the international conference on machine learning* (pp. 6105–6114). Long Beach, CA, USA.

Vapnik, V. N. (1998a). *Statistical learning theory*. New York, NY, USA: Wiley.

Vapnik, V. N. (1998b). *Statistical learning theory*. Wiley-Interscience.

Wang, J., Zhao, D., Li, H., y Wang, D. (2024a). Lightweight violence detection model based on 2d cnn with bi-directional motion attention. *Applied Sciences*, *14*(11), 4895. Descargado de https://doi.org/10.3390/app14114895 doi: 10.3390/app14114895

Wang, J., Zhao, D., Li, H., y Wang, D. (2024b). Lightweight violence detection model based on 2d CNN with bi-directional motion attention. *Applied Sciences*, *14*(11), 4895. Descargado de https://doi.org/10.3390/app14114895 doi: 10.3390/app14114895

Wu, P., Liu, J., Liu, Y., Wang, Y., Van de Weijer, J., y cols. (2020). Not only look, but also listen: Learning multi-modal violence detection under weak supervision. En *Proceedings of the european conference on computer vision workshops (eccv-w).* Descargado de https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123750324.pdf

Wu, P., Su, W., Pang, G., Sun, Y., Yan, Q., Wang, P., y Zhang, Y. (2025). Avadclip: Audio-visual collaboration for robust video anomaly detection. *arXiv preprint arXiv:2504.04495*. Descargado de https://arxiv.org/abs/2504.04495

Ye, L., Liu, T., Han, T., Ferdinando, H., Seppänen, T., y Alasaarela, E. (2021). Campus violence detection based on artificial intelligent interpretation of surveillance video sequences. *Remote Sensing*, *13*(4), 628. Descargado de https://doi.org/10.3390/rs13040628 doi: 10.3390/rs13040628