Aplicación de técnicas de aprendizaje de máquina al análisis de archivos de vídeo para la detección de delitos

Juan Camilo Londoño Lopera

Universidad Nacional de Colombia
Facultad de Minas, Departamento de Energía Eléctrica y Automática
Medellín, Colombia
2024

Aplicación de técnicas de aprendizaje de máquina al análisis de archivos de vídeo para la detección de delitos

Juan Camilo Londoño Lopera

Tesis presentada como requisito parcial para optar por el título de:
Magister en Ingeniería - Automatización Industrial

Director:
Dr. Freddy Bolaños Martinez

Co-Director:
Dr. Luis Alejandro Fletscher Bocanegra

Línea de Investigación:
Sistemas de ingeniería inteligentes

Universidad Nacional de Colombia
Facultad de Minas, Departamento de Energía Eléctrica y Automática
Medellín, Colombia
2024

Agradecimientos

Expreso mi sincero agradecimiento al Doctor Freddy Bolaños Martínez y al Doctor Luis Alejandro Fletcher Bocanegra, quienes desempeñaron los roles de director y co-director de este proyecto de tesis de maestría. Su guía y apoyo continuo fueron fundamentales en cada etapa de este proceso académico. Además, quiero extender mi gratitud a la Doctora Mónica Ayde Vallejo Velásquez por su valioso acompañamiento, sus aportes significativos y el tiempo generosamente dedicado a mi formación de posgrado. Su compromiso y orientación han sido pilares fundamentales en mi desarrollo académico y profesional.

Agradezco profundamente a mis padres y hermanos por su apoyo constante y aliento a lo largo de mi carrera académica. Su respaldo emocional y comprensión han sido fundamentales en los momentos desafiantes. Este logro no habría sido posible sin su apoyo incondicional y sacrificios.

This work was supported by the Science, Technology, and Innovation Fund (FCTeI) of the General Royalties System (SGR) under the project identified by code BPIN 2020000100044, Universidad de Antioquia and Universidad Nacional de Colombia.

Resumen

Esta tesis se centra en el desarrollo de una aplicación para seguridad ciudadana mediante técnicas de aprendizaje de máquina, con el objetivo principal de detectar delitos a través del análisis de archivos de video. La investigación comienza con una revisión sistemática de las técnicas más relevantes, estableciendo criterios de selección que priorizan estructuras capaces de integrar eficientemente la dimensión temporal. Se favorecen modelos de aprendizaje de máquina, que ofrecen versatilidad para la incorporación de nuevos parámetros, especialmente aquellos basados en esquemas espacio-temporales, fundamentales para el análisis de video y la consideración del contexto temporal de los eventos.

Dado que la recolección de datos extensos y etiquetados resulta inviable en el marco temporal del proyecto, se opta por utilizar simulaciones basadas en conjuntos de datos públicos en línea diseñados específicamente para la detección de delitos. Se selecciona cuidadosamente al menos un tipo de delito para la investigación, considerando su relevancia y disponibilidad de repeticiones para el desarrollo efectivo del modelo de predicción. La validación del modelo se lleva a cabo mediante una evaluación exhaustiva, utilizando diversos conjuntos de datos previamente seleccionados y parámetros clave de desempeño, como la curva ROC - AUC. Este enfoque integral busca garantizar la eficacia y aplicabilidad del modelo en entornos prácticos y del mundo real.

Palabras clave: Seguridad ciudadana, Aprendizaje de máquina, Detección de delitos, Modelos LSTM, Redes Neuronales Convolucionales 3D, Predicción de eventos

Application of machine learning techniques to video file analysis for crime detection

Juan Camilo Londoño Lopera

Director:
Dr. Freddy Bolaños Martinez

Co-Director:
Dr. Luis Alejandro Fletscher Bocanegra

Universidad Nacional de Colombia
Facultad de Minas, Departamento de Energía Eléctrica y Automática
Medellín, Colombia
2024

Abstract

This thesis focuses on developing an application for public safety through machine learning techniques, with the primary goal of crime detection by analyzing video files. The research begins with a systematic review of the most relevant techniques, establishing selection criteria that prioritize structures capable of efficiently integrating the temporal dimension. Machine learning models are favored for their versatility in incorporating new parameters, especially those based on spatiotemporal schemes, crucial for video analysis and considering the temporal context of events.

Since the collection of extensive and labeled data is impractical within the project's timeframe, simulations based on publicly available online datasets specifically designed for crime detection are used. At least one type of crime is carefully selected for investigation, considering its relevance and the availability of repetitions for the effective development of the prediction model. Model validation is conducted through a comprehensive evaluation, utilizing various pre-selected datasets and key performance parameters, such as the ROC-AUC curve. This holistic approach seeks to ensure the effectiveness and applicability of the model in practical and real-world settings.

Keywords: Public safety, Machine learning, Crime detection, LSTM Models, 3D Convolutional Neural Networks, Event Prediction

Contenido

Lista de Figuras

Lista de Tablas

1 Introducción

El mundo está evolucionando a gran velocidad y la mayor parte de los elementos que conforman una ciudad están comenzando a interactuar entre ellos y a tener cierto nivel de autonomía. Elementos como las viviendas, el tránsito vehicular, los sistemas de videovigilancia, el monitoreo del clima, entre muchos otros aspectos, se pueden interconectar y operar de manera colaborativa. Este fenómeno se atribuye a los avances en el manejo de grandes volúmenes de datos (*Big Data*) y a los sistemas relacionados con internet de las cosas (*IoT por sus siglas en inglés*)[Rathore et al., 2016]. Estas tecnologías tienen la capacidad de procesar toda la información generada por los dispositivos y sensores inteligentes para construir o proponer arquitecturas que contribuyen no solo a los elementos individuales, sino a la operación integral de todo el conjunto. De esta manera, se conforma una ciudad inteligente con la capacidad de tomar decisiones referentes a temas de seguridad, desastres, gestión de recursos, entre otros aspectos fundamentales para el mejoramiento de la calidad de vida [Simić et al., 2020].

Uno de los aspectos más importantes para el mejoramiento de la calidad de vida de las personas es la prevención de delitos. Regiones con gran población suelen estar relacionadas con tasas de criminalidad alta, por lo que garantizar la seguridad pública es una tarea cada vez más difícil. Según la OCDE (Organización para la Cooperación y el Desarrollo Económico) la seguridad ciudadana es un factor determinante para el bienestar de las personas por lo que este tema se ha vuelto una prioridad para las administraciones públicas en la mayoría de los territorios a nivel internacional [OECD, 2021]. Adicionalmente, las estadísticas delictivas generadas por SIEDCO (Sistema de Información Estadístico, Delincuencial Contravencional y Operativo de la Policía Nacional) [SIEDCO, 2021] mencionan que los delitos que presentan las mayores tasas en el país son:

- Hurto a personas
- Violencia intrafamiliar
- Hurto de vehículos
- Delitos sexuales
- Homicidios

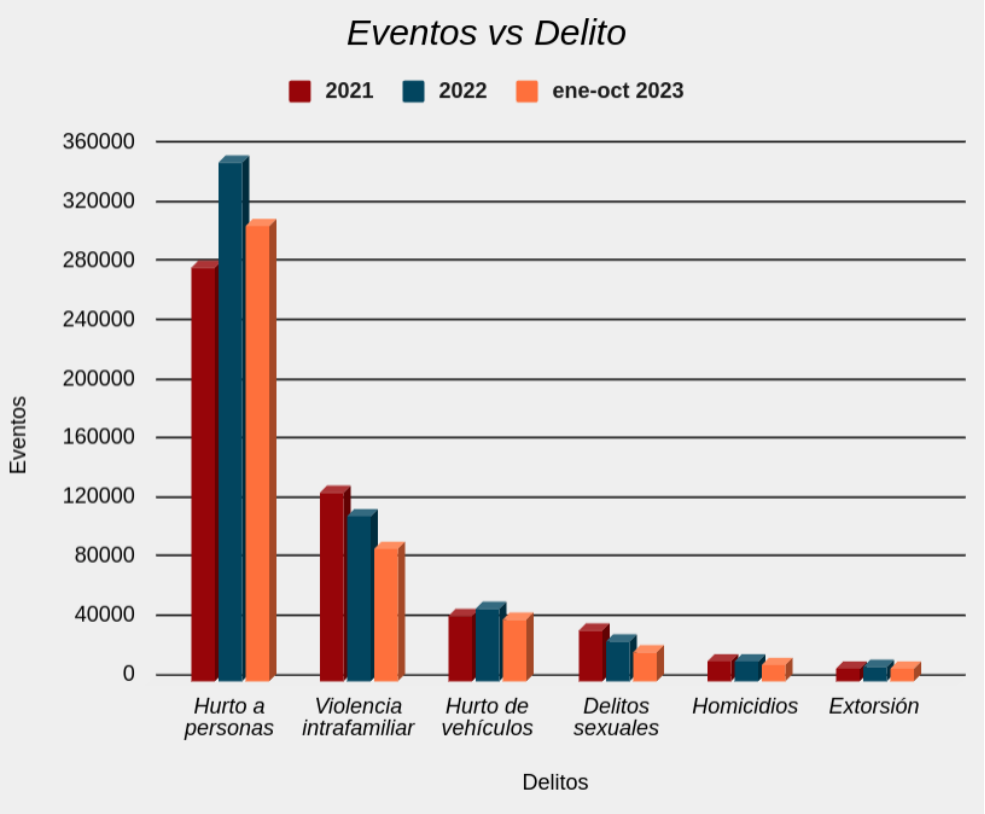


Figura 1-1: Cifras de delitos en Colombia, fuente: SIEDCO

- Extorsión

En la Figura **1-1** se observan las cifras de delitos que se han reportado en los últimos 3 años en todo el territorio Colombiano. En esta se evidencia el problema de criminalidad por el que atraviesa el país, siendo el hurto a personas la mayor preocupación con alrededor de 310 mil eventos en promedio. Por esta razón, son de vital importancia el desarrollo de trabajos que han estado enfocando sus esfuerzos en mejorar el despliegue de los recursos policiales de cada región para prevenir delitos. Algunos de estos se relacionan con sistemas de re-identificación de personas usando bases de datos de rostros y redes neuronales convolucionales con resultados que alcanzan valores de desempeño del 98 % [Medapati et al., 2019]. Aun así se menciona que es posible que se puedan realizar mejoras en este tipo de sistemas implementando diversos métodos de optimización. Adicionalmente, se han implementado modelos de aprendizaje profundo con el objetivo de fortalecer el proceso de reconocimiento facial de los sistemas de monitoreo de las ciudades, demostrando que es posible la obtención de resultados en tiempo real [Zhu and Yang, 2018], con tiempos de respuesta del orden de los 20 milisegundos de modo que se logre dar un apoyo acertado a las fuerzas policiales para contribuir al tema seguridad pública.

Algunos de los trabajos realizados se han enfocado en proponer arquitecturas que puedan contribuir a la tarea de procesar la cantidad masiva de información que se puede recolectar de una ciudad. Estas propuestas no solo benefician a los ciudadanos, sino que también contribuyen al bienestar general de toda la ciudad [Rathore et al., 2016]. En primer lugar, se utiliza el concepto del Internet de las Cosas para conectar una amplia gama de dispositivos y sensores en diferentes áreas. La idea es recopilar información sobre los principales aspectos urbanos, como son: la calidad del aire, el consumo de energía, los índices de tráfico, sistemas de videovigilancia, entre otros elementos que generan grandes volúmenes de datos. Posteriormente, se realiza un análisis de *Big Data*, correlaciones, reconocimiento de patrones, algoritmos de predicción y detección, con el fin de extraer información relevante de los datos para tener un contexto de cómo funciona la ciudad y así tomar decisiones acertadas.

Es necesario tener presente que la administración de grandes volúmenes de datos de los ciudadanos requiere de un tratamiento especial referente a la privacidad de la información, por lo que es necesario brindar soluciones que contemplen este aspecto. Para complementar lo anterior, se han realizado desarrollos que buscan proporcionar seguridad a la información electrónica de las aplicaciones de una ciudad inteligente, pero que al mismo tiempo puedan trabajar en tiempo real para no afectar la velocidad en la toma de decisiones. El objetivo es que el sistema logre ser seguro y presente buen rendimiento en cuanto a costo computacional, contribuyendo a las etapas críticas de este tipo de aplicaciones [Rathore et al., 2018]. La velocidad de procesamiento, la seguridad de la información y los sistemas de comunicaciones son elementos críticos en cuanto a la toma de decisiones, por lo que se han planteado diversas arquitecturas con el objetivo de dar una idea respecto a las implicaciones que tiene el manejo de grandes conjuntos de datos [Simić et al., 2020].

Hoy en día las ciudades cuentan con numerosos sistemas de videovigilancia que comúnmente no se logran aprovechar al máximo, pues la cantidad de información es tan grande que es difícil de procesar para un operador humano. Sin embargo, es de mucho interés investigativo y práctico darle uso a esta información, por lo que actualmente es el insumo principal para el desarrollo de conjuntos de datos como LAD (*Large-scale Anomaly Detection*)[Wan et al., 2021] que son utilizados para la creación de modelos basados en aprendizaje de máquina para contribuir a la identificación y prevención de delitos. Con base en lo anterior, se ha logrado identificar un tópico relacionado con el tema de prevención de delitos llamado detección de eventos anormales o detección de anomalías en video (*VAD por sus siglas en inglés*), donde el objetivo es identificar comportamientos o patrones de apariencia que están fuera de lo común. Actualmente, es un tema de investigación que sigue siendo estudiado, pues contribuye a la seguridad pública usando la información suministrada por los sistemas de videovigilancia [Ullah et al., 2021b].

Normalmente en un video la ocurrencia de eventos anormales es algo poco común por lo que esta tarea de detección de anomalías se ha convertido en un desafío en cuanto a investigación debido a la falta de datos. Otro concepto relacionado que se ha ido desarrollando es el término de Aprendizaje de Múltiples Instancias (*MIL por sus siglas en inglés*) que se basa en un video que se transmite como un paquete de fragmentos con el objetivo de aprender a identificar el suceso o la etiqueta de cada fragmento a través de la anotación del video. Se han implementado modelos de aprendizaje profundo basados en redes neuronales convolucionales 3D en conjunto con aprendizaje de múltiples instancias donde se han alcanzado mejoras de alrededor de un 8 % comparado con modelos convencionales [Ullah et al., 2021a]. Con base en lo anterior, se puede decir que la detección de anomalías en video sigue siendo un importante tópico de investigación en el que es crucial realizar nuevos desarrollos que logren ser implementados en la vida real y que contribuyan con el mejoramiento de la calidad de vida.

En el contexto descrito previamente, donde las ciudades inteligentes emergen como entidades dinámicas y conectadas, se enfrentan desafíos críticos relacionados con la seguridad ciudadana. La interconexión de diversos elementos urbanos, respaldada por tecnologías de *Big Data* e *IoT*, proporciona una oportunidad para optimizar la gestión de recursos, prevenir delitos y mejorar la calidad de vida. En este escenario, la prevención del crimen se presenta como una prioridad, respaldada por estadísticas que destacan la tasa de delitos como hurto a personas, violencia intrafamiliar, hurto de vehículos, delitos sexuales, homicidios y extorsión. Con base en lo anterior, este trabajo de tesis cobra una importancia estratégica al abordar la detección de delitos mediante la aplicación de técnicas de aprendizaje de máquina en el análisis de archivos de video.

1.1. Justificación

Elementos como la tasa de fecundidad, el aumento de la longevidad y la migración internacional son factores que han influido directamente en el aumento de la población en el mundo. Este aumento poblacional ha generado una gran evolución en los procesos urbanísticos y se han ido desarrollando grandes ciudades, donde debido al gran número de personas se presentan problemas de desigualdad, falta de oportunidades, diferencias culturales, entre otros aspectos que provocan que estén relacionadas con índices de criminalidad altos. Es por esto que la seguridad pública se ha convertido en uno de los retos más importantes a trabajar en territorios con grandes poblaciones, pues es una de las principales características que contribuye con el mejoramiento de la calidad de vida de las personas. Por ejemplo, según cifras del Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo (SIEDCO) de la Policía Nacional en el año 2021 en la ciudad de Bogotá el delito más frecuente fue el hurto a personas, se presentaron 166.858 casos, un promedio de 687 por día. Es una cifra que ha venido incrementando por lo que es fundamental ahondar esfuerzos en cuanto a detección y prevención de delitos con el fin de mejorar y contribuir al tema de seguridad pública.

Para atacar este problema en diferentes partes del mundo se han implementado sistemas de reconocimiento facial, seguimiento de sospechosos, sistemas de vigilancia urbana, entre otro tipo de aplicaciones [Simić et al., 2020]. Aun así, la seguridad ciudadana sigue siendo considerada uno de los grandes retos a trabajar pues se requiere de eficacia y una toma de decisiones rápida para atacar el problema, debido a que frecuentemente los operadores humanos presentan tiempos de reacción limitados. Por tal motivo, este trabajo de tesis busca explorar las diferentes alternativas para detectar delitos a través de los sistemas de videovigilancia instalados con el fin de reducir los tiempos de respuesta de las autoridades policiales y de este modo contribuir a la seguridad pública.

1.2. Objetivos

1.2.1. Objetivo General

Desarrollar un modelo basado en aprendizaje de máquina que detecte delitos a partir de la información suministrada por los sistemas de videovigilancia.

1.2.2. Objetivos específicos

- Seleccionar e identificar técnicas relevantes para el análisis de video con aprendizaje de máquina aplicadas a la detección de delitos.
- Definir los conjuntos de datos, las métricas de desempeño y al menos un tipo de delito con base en las anomalías presentes en las bases de datos públicas.
- Desarrollar un modelo para detección de delitos basado en aprendizaje de máquina a partir de las técnicas seleccionadas.
- Evaluar el modelo en diferentes conjuntos de datos y realizar un análisis de los resultados.

1.3. Antecedentes

Actividades tan simples como lo es una caminata por la ciudad pueden llegar a representar grandes desafíos en torno a la seguridad ciudadana. Por este motivo diseñaron la aplicación SPATH (Safest PATH)[Pang et al., 2019] que busca proporcionar una navegación segura a pie. En este desarrollo consideran que la mayoría de recursos informáticos disponibles en el área urbana como las cámaras, la infraestructura celular y los vehículos con recursos informáticos están infrautilizados por lo que los usan como arquitectura principal para procesar y transmitir videos. De este modo, los usuarios pueden acceder previamente y visualizar el estado de los senderos que desean transitar. Su principal problema es que la transmisión a larga distancia de un gran volumen de videos es una carga grande para la red. Por esta razón, plantean el algoritmo *Fast Iterative Matching (FIM)* de baja complejidad para resolver eficazmente el problema de minimización de latencia. Los autores realizaron un trabajo comparativo mediante un esquema de 20 nodos de cámaras con los enfoques *Greedy assignment scheme (GRE)* y *Random assignment scheme (RAN)*, donde FIM se presenta una latencia de aproximadamente 10 segundos mientras que los enfoques GRE y RAN están alrededor de los 90 segundos demostrando así la eficacia de la solución propuesta.

En otros estudios se han planteado métodos que detectan áreas sensibles y que buscan predecir las tendencias delictivas. Estos se enfocan en modelos de predicción espacio-temporal donde se busca estimar el número de delitos que es probable que ocurran en una región asociada [Catlett et al., 2019]. Sus modelos fueron evaluados con conjuntos de datos reales de grandes ciudades como son Nueva York y Chicago. Los conjuntos de datos incluyen aproximadamente 2 millones de eventos delictivos de 16 años de Chicago y 1.5 millones de 11 años de Nueva York. El enfoque propuesto presenta resultados alrededor de un 5 % mejores que los propuestos en un trabajo similar [Gorr et al., 2003], donde implementaron el enfoque de suavizado exponencial de Holt para desarrollar un modelo de pronóstico preciso para series de delitos. Se menciona que falta avanzar en modelos que logren correlación de eventos de la ciudad con los delitos para comprender la relación entre ellos.

Por otra parte, se presenta la plataforma *CriClust*[Isafiade and Bagula, 2020], cuyo objetivo principal es la detección de patrones delictivos. Este enfoque es de relevancia debido a la persistente preocupación por el control del crimen en numerosas regiones del mundo. El estudio de este fenómeno es esencial, ya que los delitos rara vez son eventos aleatorios, con frecuencia implican planificación. Este estudio se concentra particularmente en el análisis de delincuentes en serie, ya que estos suelen ser responsables de una proporción significativa de los delitos en algunas regiones. La carencia de herramientas inteligentes adecuadas para apoyar a las agencias encargadas del orden público resalta la importancia de soluciones como *CriClust*. Si bien su enfoque principal radica en identificar patrones en delitos sexuales y robos, su aplicabilidad se extiende a una variedad de delitos en diversas áreas. Un componente fundamental

de esta plataforma es la colaboración cercana con expertos en inteligencia criminal. Esta colaboración es esencial para definir adecuadamente los parámetros de búsqueda y garantizar que la herramienta funcione de manera efectiva en la identificación de patrones delictivos.

A pesar de los esfuerzos realizados por los investigadores en este campo, la seguridad pública sigue siendo un gran desafío para las ciudades inteligentes. Actualmente, se ha observado un enfoque cada vez más prominente en el desarrollo de soluciones con un enfoque espacio-temporal utilizando redes neuronales 3D. Un ejemplo notable es la creación de un modelo de aprendizaje débilmente supervisado [Sultani et al., 2019] basado en el Aprendizaje de Múltiples Instancias (MIL). Este enfoque fue probado con el conjunto de datos *UCF-Crime*, que abarca 128 horas de material de video e involucra 13 categorías de anomalías relacionadas con situaciones como peleas, accidentes y robos, entre otros eventos. Estos esfuerzos y conjuntos de datos contribuyen ampliamente a la comunidad académica debido a que ofrecen una base sólida para el desarrollo de soluciones que aborden las problemáticas de la seguridad pública. Además, se han realizado investigaciones destinadas a explorar nuevas perspectivas en el aprendizaje de múltiples instancias, incluso mediante el empleo de redes neuronales convolucionales gráficas, que buscan disminuir el ruido en las etiquetas a través de técnicas de aprendizaje supervisado [Zhong et al., 2019]. Estos enfoques han demostrado obtener valores de precisión superiores al 80 %, pero se menciona la posibilidad de mejorar aún más los resultados mediante optimizaciones en los métodos de clasificación. En una línea similar, en otro estudio [Zhang and Yu, 2018], se desarrolló un modelo que integra un módulo de convolución deformable, con el propósito de abordar las limitaciones asociadas con las transformaciones geométricas en las redes neuronales convolucionales. Estos esquemas de aprendizaje automático resultan esenciales para el análisis y correlación de eventos relacionados con delitos en una región específica.

El enfoque MIL es un tópico que se ha venido trabajando de manera recurrente en los últimos años, métodos como RTFM (Aprendizaje robusto de magnitud de característica temporal) [Tian et al., 2021], se han utilizado para detectar de una forma efectiva las instancias positivas y han alcanzado resultados satisfactorios en el rango del 75 % y el 98.6 % de precisión demostrando ser una solución que puede ser implementada en la vida real y brindar características positivas a los sistemas actuales de videovigilancia. Existen otros métodos como *Deep Auto-Encoder (AE)* que son populares en los sistemas de detección de anomalías, se han realizado trabajos para mejorar el proceso mediante una unidad de prototipo dinámico y meta-aprendizaje alcanzando resultados arriba del 85 % superando los métodos tradicionales [Lv et al., 2021a]. Otros investigadores han implementado modelos de redes neuronales en conjunto con señales de audio para la detección de anomalías, este tipo de modelos multimodal presenta efectos muy positivos en la detección. Sin embargo, aún se considera un problema el tema de la detección en línea [Wu et al., 2020], por lo que es importante implementar mejoras que aporten en ese aspecto.

La detección de anomalías con etiquetas débiles ha sido en gran parte explorada en el contexto de delitos. Sin embargo, esta técnica puede extrapolarse a otros dominios, como la movilidad, para identificar problemas de tráfico, accidentes y otros elementos que inciden en la seguridad pública. Un ejemplo ilustrativo es una implementación exitosa [Lv et al., 2021b], que logró un rendimiento superior al 85 % y fue evaluada en una tarjeta de video 2080Ti, alcanzando una velocidad de procesamiento de 44 FPS, lo que sugiere su aplicabilidad en escenarios en tiempo real.

Varios trabajos se han centrado en dotar a los investigadores de herramientas para abordar esta tarea de detección. Por ejemplo, el desarrollo del conjunto de datos Benchmark LAD [Wan et al., 2021] para la detección de anomalías en videos es notable. Este conjunto incluye 2000 secuencias de video con 14 categorías de anomalías. Su singularidad radica en las etiquetas, tanto a nivel de video como a nivel de cuadro, permitiendo la implementación de modelos de aprendizaje completamente supervisados. Utilizando una red neuronal profunda 3D, se logró una precisión superior al 86 %, probada en cinco conjuntos de datos diferentes. A pesar de estos avances, persiste la necesidad de investigar más en la detección de anomalías en secuencias de video. Un enfoque similar es evidenciado en el conjunto de datos UBnormal [Acsintoae et al., 2021], que no solo ofrece anotaciones a nivel de píxel, sino también un análisis comparativo de los métodos desarrollados en este campo. Estos desarrollos no solo proporcionan una base esencial para la seguridad pública, sino también para la implementación efectiva en aplicaciones de ciudades inteligentes.

Con base en la revisión del estado del arte, se ha observado que los métodos utilizados para la detección de delitos en videos se pueden clasificar en tres categorías principales (ver Tabla: **1-1**). En donde se ha comprobado que las técnicas de aprendizaje profundo, en general, superan a aquellas basadas en clasificación por un margen que varía entre el 10 % y el 20 %.

Tabla 1-1: Categorías métodos detección de delitos en videos

Métodos basados en:	**Descripción**
Características estadísticas	Utilizan técnicas para analizar los datos y detectar patrones fuera de lo común. Típicamente incluyen la detección de desviaciones, la identificación de tendencias y la detección de datos atípicos. Ejemplo: [Catlett et al., 2019]
Clasificación	Se basan en un sistema de puntuación. Estos métodos son usados para detectar eventos en tiempo real, combinan técnicas de aprendizaje profundo y estadísticas para identificar los eventos más relevantes. Ejemplo: [Sultani et al., 2019]
Aprendizaje profundo	Utilizan redes neuronales para detectar patrones anómalos. Son métodos que tienden a ser más precisos que los basados en características estadísticas, pero son más complejos y requieren más datos para entrenar. Ejemplo: [Cheng et al., 2021]

Se presenta una clasificación adicional de técnicas en [Ramzan et al., 2019], donde se realiza una revisión exhaustiva de investigaciones enfocadas en la detección de violencia, estableciendo el contexto para identificar este tipo de eventos y resaltando su relevancia en términos de seguridad pública y prevención del crimen. También se abordan los diversos tipos de violencia y los desafíos intrínsecos a la detección automatizada, tales como la variabilidad en la apariencia de objetos y la complejidad de las escenas. La clasificación está enfocada en cuatro categorías primordiales relacionadas con técnicas de detección de violencia:

1. Basadas en características manuales: Estas técnicas se apoyan en el uso de información visual y de metadatos asociados para identificar patrones específicos de violencia. Los descriptores de bajo nivel, como el color, la textura y la forma, se extraen de los fotogramas de video para ayudar a clasificar las escenas. Estas características pueden ser utilizadas posteriormente en algoritmos de aprendizaje automático clásicos para la detección de violencia.

2. Basadas en aprendizaje profundo: Estas técnicas se centran en el uso de redes neuronales profundas, como las redes convolucionales o las redes recurrentes, para aprender directamente patrones complejos de violencia en los datos de video. Estas redes pue-

den capturar relaciones y dependencias a diferentes niveles de abstracción, lo que les permite identificar características sutiles en los videos relacionados con la violencia.

3. Fusión de características: Esta categoría se refiere a la combinación de múltiples tipos de características extraídas de videos, como características de bajo nivel, audio y metadatos, para mejorar la capacidad de detección de violencia. La fusión de características busca aprovechar la información heterogénea de los videos para obtener una representación más completa de las escenas y así aumentar la precisión en la detección de la violencia.

4. Aprendizaje activo: Esta categoría se basa en estrategias que permiten la selección inteligente de muestras de video para mejorar la eficacia del modelo de detección de violencia. Estas técnicas seleccionan de manera activa los videos que presentan características ambiguas o difíciles de clasificar, lo que permite al sistema centrar su atención en las instancias más informativas y complejas, lo que finalmente aumenta la capacidad de detección y clasificación del modelo.

Esté trabajo proporciona una revisión de las técnicas de detección de violencia en videos, resalta los principales desafíos que son propios de esta tarea y examina soluciones potenciales. Además, destaca la necesidad de propuestas más sólidas y eficientes para la detección de violencia en videos. Asimismo, el estudio presenta una base sólida y exhaustiva para futuras investigaciones, señalando la importancia de abordar la complejidad y la diversidad de los datos de video para lograr una detección de violencia más precisa.

Uno recurso relevante en el tema de detección de delitos en videos es el trabajo presentado por [Vahdani and Tian, 2023]. Este documento presenta un compendio de conjuntos de datos, métricas y métodos utilizados en el campo de la detección de acciones en videos no recortados, que pueden ser extrapolados a la problemática de detección de delitos. La detección de acciones en videos representa un desafío significativo, ya que en la mayoría de los videos, las acciones de interés ocurren en intervalos de tiempo muy cortos, lo que limita la cantidad de ejemplos disponibles para entrenar un modelo. Como solución a esta problemática, la comunidad académica ha estado trabajando en el desarrollo de diversos conjuntos de datos que desempeñan un papel crucial en el entrenamiento y evaluación de algoritmos de detección de acciones. Estos conjuntos de datos están diseñados para proporcionar ejemplos diversificados y representativos de acciones en una amplia variedad de contextos.

Además de los conjuntos de datos, se ha realizado un progreso significativo en el desarrollo de técnicas y métodos que abordan la detección de acciones desde diversas perspectivas. Esta diversidad de enfoques es esencial porque en muchos casos, los datos se organizan de manera que no son compatibles con un solo método. Por lo tanto, esta investigación ha llevado a una clasificación de los métodos en determinadas categorías (ver Tabla: **1-2**), lo que facilita la comprensión y el uso de estas técnicas.

Tabla 1-2: Métodos según el tipo de aprendizaje

Métodos	**Descripción**
Métodos Totalmente Supervisados	Requieren etiquetas precisas para cada acción en los datos de entrenamiento. Aunque son altamente precisos, en la mayoría de los casos se enfrentan a la limitación de la disponibilidad de etiquetas precisas para grandes conjuntos de datos.
Métodos Débilmente Supervisados	Las etiquetas son menos precisas o pueden estar disponibles sólo a nivel de video, lo que conlleva a mayor disponibilidad de datos y por ende a un entrenamiento en conjuntos de datos más grandes.
Métodos No Supervisados	El modelo aprende patrones de manera no supervisada, es decir, sin etiquetas explícitas. Estos métodos son útiles cuando no se dispone de etiquetas precisas o cuando se desea explorar patrones inesperados en los datos.
Métodos Semi Supervisados	Estos métodos combinan datos etiquetados y no etiquetados para mejorar el rendimiento del modelo. Son efectivos cuando se dispone de una cantidad limitada de datos etiquetados.
Métodos Autosupervisados	El modelo se entrena para aprender a partir de sus propias predicciones, lo que puede ser útil cuando las etiquetas precisas son escasas.

Otro aspecto importante es la importancia de las métricas de evaluación en la medición del desempeño de los métodos. Algunas de las métricas clave incluyen las definiciones presentadas en la Tabla **1-3**.

Tabla 1-3: Métricas de evaluación

Métrica	**Descripción**
Average Recall (AR)	Evalúa la capacidad de un modelo para para identificar correctamente los eventos positivos en un conjunto de datos.
AN curve (AUC)	Representa la capacidad del modelo para discriminar entre clases en las categorías.
Mean Average Precision (mAP)	Calcula el promedio de las puntuaciones de precisión para diferentes acciones.
Average Precision (AP)	Mide la precisión promedio para una acción específica.
Frame-AP	Se centra en la precisión en el nivel de cuadros dentro de un video.
Video-AP	Evalúa la precisión en el nivel de video.

Estos avances y clasificaciones en la detección de acciones en videos no recortados marcan un avance significativo en el campo de la visión por computadora y el aprendizaje automático. Al proporcionar un enfoque integral para comprender y analizar el contenido visual en videos complejos, estos desarrollos están determinando el camino a seguir para una gama diversa de aplicaciones prácticas y de investigación. No solo están contribuyendo de manera importante a la mejora de la seguridad pública y la videovigilancia, sino que también a la implementación de sistemas de detección de acciones inteligentes y adaptativos en una variedad de entornos y contextos.

Además, estas investigaciones permiten la implementación de métodos que ayudan con una identificación más precisa y una comprensión más profunda de las acciones humanas y los eventos en tiempo real. Y tienen el potencial de impulsar avances significativos en áreas como la seguridad cibernética, la investigación forense, la salud y más. A medida que se perfeccionan y refinan aún más estas técnicas, se espera que su impacto en la sociedad y la industria sea aún más profundo y transformador.

1.4. Marco teórico

1.4.1. Ciudad inteligente

En las últimas décadas, el rápido avance de la tecnología ha impulsado un rápido crecimiento en la urbanización global. Este fenómeno ha dado lugar a un aumento significativo en la densidad poblacional de las áreas urbanas, generando la necesidad de desarrollar infraestructuras y servicios vitales, como redes eléctricas, suministro de agua, sistemas de transporte y otras comodidades con el objetivo de elevar la calidad de vida en estas regiones. El manejo eficiente de esta diversidad de recursos es crucial para satisfacer las necesidades de la población en constante expansión, por lo que es importante la implementación de estrategias que aborden las distintas problemáticas específicas de cada localidad o región. En respuesta a estos desafíos, ha surgido el concepto de ciudades inteligentes (ver Figura **1-2**), el cual propone la utilización de herramientas tecnológicas para una administración más eficiente de los recursos urbanos, con el fin de garantizar la sostenibilidad a largo plazo y, sobre todo, mejorar el bienestar y la comodidad de los residentes [Anthopoulos, 2015].

Figura 1-2: Ciudad Inteligente. Tomada de https://sysman.com.co

1.4.2. Análisis de datos

Debido a las tecnologías usadas en una ciudad inteligente como son los sensores, las redes de comunicación, el uso del *Big Data*, los sistemas de control y las plataformas de participación

ciudadana, se genera una gran cantidad de datos importantes que deben ser procesados para tomar decisiones acertadas en beneficio de la comunidad. De ahí la importancia del análisis de datos que desempeña un papel fundamental en la gestión de una ciudad inteligente. El contexto de esta investigación, se centra en el análisis de datos para fortalecer la seguridad en entornos urbanos inteligentes, siendo este un pilar esencial en el concepto mismo de ciudad inteligente. El análisis de datos permite la recolección y procesamiento de información proveniente de dispositivos como cámaras de videovigilancia, redes sociales y sistemas de información geográfica. A través de técnicas de minería y análisis de datos, aprendizaje automático y análisis de imágenes, se pueden identificar patrones, anomalías y tendencias relevantes para la seguridad pública.

1.4.3. Seguridad Ciudadana

Las administraciones gubernamentales en distintas regiones del mundo suelen establecer políticas públicas, como la Política de Seguridad y Convivencia (PSC) en Colombia, que busca atender las necesidades relacionadas con la seguridad ciudadana, y en la que participan diversos actores, incluida la Policía Nacional. Es esencial proveer a estos agentes del orden público de herramientas que faciliten el cumplimiento de sus responsabilidades. Algunas de estas herramientas son:

- Sistemas de información de mando y control: Estas herramientas centralizan la información de la región con el propósito de permitir a los agentes tomar decisiones en temas relacionados con la seguridad ciudadana.
- Sistemas de videovigilancia: Incluyen redes de cámaras, sistemas de monitoreo, grabación y almacenamiento de videos, y otros recursos tecnológicos que generan una gran cantidad de información crítica para el análisis y la toma de decisiones.
- Modelos de predicción del delito: Utilizan técnicas de aprendizaje automático para predecir la ocurrencia de delitos. Se basan en datos históricos de incidentes delictivos y factores contextuales, como la densidad poblacional, la ubicación de puntos de interés y los patrones de movilidad.
- Modelos de detección de eventos anómalos: Se centran en la detección de eventos inusuales o anómalos que puedan indicar situaciones de riesgo o potenciales amenazas. Utilizan técnicas de aprendizaje automático para analizar datos en tiempo real, como flujos de video, datos de sensores y registros de actividad, y generan alertas tempranas para una respuesta rápida.
- Arquitecturas de sistemas de seguridad integrados: Se basan en la interconexión de diferentes sistemas y dispositivos de seguridad, como cámaras de videovigilancia, sistemas de control de acceso, alarmas y sensores. La integración de estos componentes permite

una gestión centralizada y eficiente de la seguridad, con capacidades de monitoreo, análisis y respuesta automatizada.

1.4.4. Aprendizaje de máquina

Durante los últimos años se han ido desarrollando varios trabaos enfocados en la detección de delitos en videos, sin embargo se menciona que este tema aún representa un desafío fundamental en la búsqueda de soluciones que fortalezcan la seguridad pública y prevengan actividades criminales. Estos trabajos presentan resultados satisfactorios y la mayoría están centrados en el desarrollo de métodos de aprendizaje de maquina. Estos métodos son un apartado de la Inteligencia artificial que busca generar sistemas que aprendan a partir de datos.

El aprendizaje de máquina hoy en día es aplicado a diferentes tipos de contextos como el procesamiento de lenguaje natural, el reconocimiento de imágenes, la minería de datos, entre otros casos que se han vuelto fundamentales para la humanidad. Principalmente se dice que es un área de estudio que utiliza datos para que las máquinas aprendan a partir de ellos, estos datos se podrían ver como experiencias, y la idea es que a medida que el sistema obtenga más experiencias, más precisos serán los resultados. Este tema también es conocido como aprendizaje automático y está basado en modelos matemáticos creados a partir de un entrenamiento o aprendizaje basado en datos, el modelo es creado a partir de los patrones que se encuentran en los datos y con esto se busca obtener las predicciones. Algunas de las tareas que se pueden realizar mediante técnicas de aprendizaje automático son:

1. Predicción de valores.
2. Identificación de anomalías.
3. Encontrar la estructura ó agrupamiento en clusters.
4. Predecir categorías.

1.4.5. Extractores de características

Una de las etapas iniciales que se desarrolla en gran parte de los métodos de aprendizaje de máquina para la detección de delitos en video es la implementación de extractores de características, también conocidos como descriptores de video. Los descriptores son un instrumento clave para el análisis de patrones y están diseñados para capturar la información mas determinante y distintiva de los videos, con el fin de poder usar esta información posteriormente en los algoritmos de aprendizaje automático. Estos descriptores buscan extraer características relacionadas con texturas, formas, colores, movimiento, propiedades visuales que permitan representar el video en forma de datos de una manera más compacta y a su vez

estructurada. Sus principales aplicaciones se resaltan en el análisis y comprensión de contenido visual, la detección de eventos, la identificación de objetos y la clasificación de acciones en videos. Durante la búsqueda bibliográfica se encontraron 2 descriptores muy utilizados C3D [Tran et al., 2015] e I3D [Carreira and Zisserman, 2017].

C3D

Es una red neuronal convolucional 3D que se usa comúnmente para la extracción de características de video, debido a que está diseñada para trabajar con secuencias de video capturando información espacial y temporal con ayuda de filtros convolucionales. Este algoritmo utiliza filtros tridimensionales para capturar información de manera mas precisa. C3D surge como un desarrollo de Berkeley Vision and Learning Center (BVLC) que es un grupo de investigación de aprendizaje profundo en visión ubicado en la Universidad de California, Berkeley. Este desarrollo esta basado en Caffe, un marco de aprendizaje profundo desarrollado teniendo en cuenta la limpieza, la legibilidad y la velocidad. Fue creado por Yangqing Jia y está en desarrollo activo por parte del Berkeley Vision and Learning Center (BVLC). Es conocido como un enfoque innovador para la extracción de características de vídeos, se entrena para aprender patrones en el movimiento mediante la incorporación de la componente temporal.

I3D

Principalmente es una red neuronal convolucional 3D que surge a parir de la arquitectura 2D del modelo inception pero utiliza una técnica llamada inflación para operar videos. Se diseñó para capturar características espaciales y temporales en videos. Este descriptor ha sido muy utilizado para realizar tareas de reconocimiento de acciones en vídeos. Combina aprendizaje supervisado y no supervisado para entrenar la red neuronal y una de las características clave de I3D es su capacidad para aprender representaciones espaciales y temporales a diferentes escalas y niveles de abstracción. Esta capacidad de capturar detalles finos y patrones de movimiento en videos la hace especialmente adecuada para tareas de reconocimiento de acciones humanas y análisis de videos complejos en aplicaciones que requieren una comprensión profunda del contenido visual.

1.4.6. Detección de delitos en video

En este contexto, las técnicas de aprendizaje automático, particularmente en el campo del aprendizaje de máquina, han emergido como herramientas esenciales para abordar el problema de detección de delitos en video de manera eficiente. A diferencia de los métodos tradicionales que dependen en gran medida de reglas manuales y enfoques heurísticos, el aprendizaje de máquina permite que los sistemas aprendan automáticamente patrones y

características relevantes directamente de los datos. Esta capacidad de adaptación y su habilidad para identificar sutilezas y patrones ocultos en los videos hacen que las técnicas sean adecuadas para la detección de eventos delictivos en una amplia variedad de escenarios. Un aspecto muy importante es la capacidad de las técnicas de aprendizaje de máquina para lidiar con la naturaleza dinámica de los datos en videos. Las situaciones delictivas pueden ocurrir en entornos muy diversos, a diferentes escalas de tiempo y con múltiples variaciones visuales. Los modelos tradicionales a menudo tienen dificultades para capturar la complejidad de estos escenarios cambiantes, mientras que este tipo de técnicas pueden aprender y adaptarse a estas variaciones, lo que mejora la precisión y la robustez de la detección.

El entrenamiento de modelos de detección de delitos en videos implica el procesamiento de grandes volúmenes de datos. Las técnicas de aprendizaje de máquina permiten a estos modelos aprovechar estas grandes colecciones de información, aprender de ellas y mejorar su rendimiento con el tiempo. La capacidad de generalización de los modelos entrenados les permite detectar patrones no solo en los datos utilizados para el entrenamiento, sino también en situaciones nuevas y desconocidas. Esto es crucial en la prevención y detección de delitos en tiempo real, lo que contribuye significativamente a la seguridad pública.

Uno de los aspectos fundamentales para analizar las técnicas que se han ido desarrollando con el objetivo de detectar comportamientos, eventos o actividades en video está relacionado principalmente con las características de los datos disponibles con los que se va a desarrollar el modelo. Es decir, dependiendo del nivel de etiquetado de cada video es posible implementar diferentes tipos de técnicas. Por lo que se plantea una clasificación de las técnicas utilizadas en esta área basadas en el tipo de conjunto de datos utilizado para el entrenamiento.

Durante la búsqueda bibliográfica se encontró que los conjuntos de datos públicos disponibles que se han ido desarrollando para utilizar en tareas de detección de eventos enfocados en delitos buscan llegar a un etiquetado a nivel de cuadro por lo que son aplicables para métodos de aprendizaje supervisado y uno de los métodos mas usados para esta categoría son las redes o modelos LSTM, una variación de las redes neuronales recurrentes, puesto que son especialmente eficaces en tareas donde la secuencia y la dependencia temporal son importantes.

1.4.7. Modelo LSTM

Los modelos LSTM (*Long Short-Term Memory*) constituyen una evolución de las arquitecturas de redes neuronales, específicamente diseñadas para abordar las complejidades de las secuencias de datos, permitiendo la captura eficiente de patrones tanto espaciales como temporales. Este enfoque innovador combina características clave de las redes neuronales convolucionales (CNN) y las redes neuronales recurrentes (RNN), lo cual lo distingue por su

capacidad única de modelar dependencias temporales a largo plazo, superando las limitaciones de las RNN tradicionales en la retención de información relevante.

La representación matemática de una red neuronal LSTM puede parecer compleja debido a la integración de operaciones de convolución y estructuras recurrentes. En términos simples, las LSTM abordan el problema de la desaparición del gradiente, común en las RNN convencionales, al introducir compuertas especializadas que regulan el flujo de información a través de la red. Estas compuertas permiten a las LSTM aprender qué información retener y cuál desechar, posibilitando así el modelar de forma más efectiva secuencias temporales complejas.

Operaciones que se llevan a cabo en una celda LSTM:

1. Puertas Sigmoidales:
   - $i_t = \sigma(W_{ix} * X_t + W_{ih} * h_{t-1} + b_i)$
   - $f_t = \sigma(W_{fx} * X_t + W_{fh} * h_{t-1} + b_f)$
   - $o_t = \sigma(W_{ox} * X_t + W_{oh} * h_{t-1} + b_o)$
   - $g_t = \tanh(W_{gx} * X_t + W_{gh} * h_{t-1} + b_g)$
2. Actualización de la Memoria Celular
   - $C_t = f_t.C_{t-1} + i_t.g_t$
   - Las puertas f_t y i_t determinan cuánta de la memoria celular se debe olvidar y cuánta se debe agregar a la memoria en el instante t
3. Cálculo de la salida
   - $h_t = o_t \tanh(C_t)$
   - La puerta o_t determina cuánta de la memoria celular se debe transmitir como salida en el instante t

Notación:

- X_t es la entrada en el instante de tiempo t
- C_t es la memoria celular en el instante de tiempo t
- h_t es la salida en el instante de tiempo t
- W y U son matrices de peso
- b es el vector de sesgo

1.4.8. Redes Neuronales Convolucionales 3D (CNN 3D)

Las Redes Neuronales Convolucionales 3D (CNN 3D) representan una extensión de las arquitecturas convolucionales tradicionales, se han diseñado con el objetivo de abordar la complejidad de datos tridimensionales, como secuencias de video o volúmenes de imágenes. Están especialmente diseñadas para capturar patrones no solo en el espacio bidimensional, como en las imágenes convencionales, sino también a lo largo de la dimensión temporal.

La representación matemática de una CNN 3D implica la aplicación de operaciones de convolución tridimensional sobre los datos de entrada, lo que permite detectar patrones en tres dimensiones. De forma similar a las redes convolucionales 2D, las redes convolucionales 3D utilizan filtros que se deslizan a lo largo de la dimensión espacial y temporal de los datos, extrayendo características relevantes en cada paso. Esta capacidad tridimensional las hace ideales para tareas que involucran secuencias temporales de imágenes, como la acción en videos o la detección de eventos en datos volumétricos.

Las CNN 3D han demostrado su eficacia en una variedad de aplicaciones, que incluyen el reconocimiento de acciones en videos, el análisis de imágenes 3D, entre otros. Tienen una gran capacidad para capturar patrones espacio-temporales y esto las convierte en una herramienta esencial en el campo de la visión por computadora y el procesamiento de video, donde la comprensión de la evolución temporal de los datos es fundamental.

Las operaciones en una capa de convolución 3D son:

1. Convolución 3D:
   - $C_{lmn} = \sum_{i=-a}^{a} \sum_{j=-b}^{b} \sum_{k=-c}^{c} W_{ijk} . X_{(l+i)(m+j)(n+k)}$
   - Donde a, b, c son los tamaños del kernel 3D en las 3 dimensiones
2. Agregar sesgo y aplicar función de activación.
   - $Y_{lmn} = \sigma(C_{lmn} + b)$
   - Donde σ representa la función de activación, como la función sigmoide o la función ReLU.
3. Agrupación 3D
   En algunas capas, podría aplicarse una operación de agrupación 3D para reducir las dimensiones espaciales de la característica.

Ecuaciones Generales

En general, para una capa de convolución 3D, la operación se puede expresar matemáticamente como:

$Y_{lmn} = \sigma(\sum_{i=-a}^{a} \sum_{j=-b}^{b} \sum_{k=-c}^{c} W_{ijk}.X_{(l+i)(m+j)(n+k)} + b)$

Notación:

- X_{ijk} es el valor en la posición (i, j, k) de la entrada 3D
- W es el tensor de pesos 3D
- b es el sesgo
- C_{lmn} es el valor en la posición (l, m, n) de las características en el mapa tridimensional

1.5. Técnicas

En cuanto a las técnicas utilizadas en la detección de delitos en videos, se opta por realizar una clasificación en dos etapas principales: los métodos basados en el aprendizaje débilmente supervisado y los métodos basados en el aprendizaje completamente supervisado. Esta elección se fundamenta en la tendencia observada en los últimos años en la comunidad académica, donde se ha centrado significativamente la investigación y se han asignado considerables recursos a estas dos categorías.

Los métodos basados en el aprendizaje débilmente supervisado han adquirido especial relevancia debido a su capacidad para aprovechar conjuntos de datos más amplios y variados. Estos enfoques son esenciales en situaciones en las que la disponibilidad de etiquetas precisas es limitada o costosa de obtener. Además, su flexibilidad para aprender patrones incluso a partir de etiquetas menos precisas o a nivel de video ha demostrado ser una ventaja significativa.

Por otro lado, los métodos basados en el aprendizaje completamente supervisado siguen siendo esenciales y altamente precisos, pero a menudo enfrentan la limitación de depender de etiquetas precisas para cada acción en los datos de entrenamiento. No obstante, a pesar de esta limitación, su uso persiste debido a su capacidad para lograr una precisión excepcional en situaciones donde se dispone de etiquetas de alta calidad.

Esta clasificación permite explorar y comparar de manera más detallada cómo estas dos categorías de métodos se comportan en diferentes contextos y conjuntos de datos. Además, al enfocarse en estas categorías principales, se pueden destacar las ventajas y desafíos específicos asociados con cada una de ellas. Este enfoque permite comprender mejor el panorama actual de la detección de delitos en videos y respaldar la elección de utilizar técnicas basadas en aprendizaje supervisado en el modelo propuesto.

1.6. Resultados de métodos de aprendizaje débilmente supervisado

En esta sección, se destacan los resultados más importantes derivados de los esfuerzos realizados por la comunidad académica, concentrándose especialmente en los métodos basados en aprendizaje débilmente supervisado. Estos desarrollos representan una contribución significativa al campo, ya que han demostrado eficacia en la resolución de problemas complejos con disponibilidad limitada de datos etiquetados. Este enfoque se caracteriza por la utilización de datos que contienen información parcial o ruidosa sobre las etiquetas, eliminando en gran medida la necesidad de grandes conjuntos de datos totalmente etiquetados. Los resultados obtenidos mediante este enfoque reflejan avances sustanciales en la capacidad de los modelos para aprender patrones y características relevantes, incluso cuando se enfrentan a la limitación de datos etiquetados.

1.6.1. Redes Neuronales Recurrentes

Se presentan tres propuestas relacionadas con modelos LSTM y RNN que buscan abordar la tarea de detección de eventos anómalos en videos de videovigilancia. La primera propuesta exhibe valores de AUC superiores al 80 %, sugiriendo que han logrado desarrollar un modelo con una capacidad predictiva notable y la habilidad de discriminar entre distintas clases. En esta propuesta, se emplearon los conjuntos de datos UCFCrime y UCFCrime2Local, los cuales se centran en la detección de delitos mediante el uso de 1900 muestras distribuidas en 13 categorías. Además, llevaron a cabo una comparativa con tres modelos preentrenados de CNN como extractores de características, revelando que ResNet-50 muestra una eficacia superior en la predicción.

La segunda propuesta sigue un enfoque similar; no obstante, utilizaron el modelo preentrenado MobileNetV2 y obtuvieron un valor cercano al 80 % con UCFCrime. Es importante tener en cuenta que en esta propuesta se utilizó la métrica "Accuracy". Por lo tanto, no es posible realizar una comparación directa de los valores, ya que, en este contexto, el "Accuracy" representa la proporción de predicciones correctas en un conjunto de datos específico y puede estar sesgado por el tamaño del conjunto de datos o por un desbalance en las clases. En contraste, el AUC mide la capacidad para distinguir entre clases y representa la probabilidad de que el modelo clasifique un ejemplo positivo aleatorio más alto que un ejemplo negativo aleatorio.

Finalmente, se presenta un enfoque de codificador-decodificador utilizando redes neuronales recurrentes, logrando un valor del 60.30 % utilizando el conjunto de datos HR-Crime que es un subconjunto de UCFCrime. Este resultado sugiere que el modelo enfrenta dificultades para distinguir entre las distintas clases.

Con base en los resultados observados y los tiempos de procesamiento que rondan los 200 ms, se puede concluir que los modelos LSTM tienen el potencial de ser implementados en un entorno real, ofreciendo una contribución valiosa a la seguridad pública en la detección de delitos. Estos modelos presentan la capacidad de proporcionar respuestas rápidas y precisas en la mayoría de los casos, lo que refuerza su idoneidad para aplicaciones prácticas en situaciones de la vida real.

1. CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks
   - Descripción: Se extraen 1000 características por cada fotograma usando el modelo de CNN preentrenado ResNet-50 para luego pasarlas por una arquitectura de red LSTM bidireccional que identifica la secuencia de fotogramas como evento anómalo o normal.
   - Conjuntos de datos: UCFCrime y UCFCrime2Local
   - Procesamiento: 0.20 seg para procesar una secuencia de 15 fotogramas.
   - Métricas: ROC - AUC
   - Resultados:
     a) UCFCrime (AUC)
        - VGG19 82.00 %
        - InceptionV3 80.00 %
        - ResNet-50 85.53 %

     b) UCFCrime2Local (AUC)
        - VGG19 87.50 %
        - InceptionV3 88.00 %
        - ResNet-50 89.05 %
   - Ref : [Ullah et al., 2021a]
2. An Efficient Anomaly Recognition Framework Using an Attention Residual LSTM in Surveillance Videos
   - Descripción: Mediante un modelo de CNN se busca aprender características visuales de secuencias de fotogramas para generar información espacio temporal y reconocer la actividad anómala. Se usa la red preentrenada MobileNetV2 [Sandler et al., 2018] para extraer 1000 características por cada fotograma. Implementan el concepto residual attention-based long short-term memory (LSTM) que puede aprender información de contexto temporal y reconocer con precisión la actividad anómala. Se genera un vector de características de 30 frames del video. Este vector alimenta la red LSTM que reconoce la actividad anomala.

- Conjuntos de datos: UCFCrime, UMN, Avenue
- Procesamiento: Tiempo de detección de 0.263 seg
- Métricas: Matrix confusion, F1 score, recall Precision, clas-wise accuracy, AUC y ROC curve
- Resultados:
  a) UCFCrime (Accuracy)
    - 78.43 %
  b) UMN [Mehran et al., 2009] (Accuracy)
    - 98.20 %
  c) Avenue [Lu et al., 2013] (Accuracy)
    - 98.80 %
- Ref: [Ullah et al., 2021b]

3. HR-Crime: Human-Related Anomaly Detection in Surveillance Videos
   - Descripción: Los autores proponen MPED-RNN de modo que sigue una arquitectura de codificador-decodificador: el decodificador aprende aproximaciones cercanas de la trayectoria normal que se decodifican con alta precisión; esto implica que, cuando se presenta con trayectorias anormales, la arquitectura del codificador-decodificador obtiene reconstrucciones inexactas que resultan en puntajes altos de anomalías.
   - Conjuntos de datos: HR-Crime
   - Métricas: ROC-AUC
   - Resultados:
     a) HR-Crime (AUC)
       - 60.30 %
   - Ref: [Boekhoudt et al., 2021]

1.6.2. Redes Neuronales Convolucionales 3D

Uno de los enfoques más ampliamente adoptados para abordar el desafío de la detección de anomalías en videos es el uso de redes neuronales convolucionales 3D (CNN 3D). Este segmento destaca 11 propuestas distintas, cada una con arquitecturas y conjuntos de datos diversos. La mayoría de estos estudios evalúan sus resultados mediante la métrica AUC, lo que motiva la elección de esta métrica como indicador principal en el presente trabajo. Al examinar estos resultados, es evidente que la mayoría supera el umbral del 80 %, y algunos

incluso alcanzan un rendimiento AUC que sobrepasa el 90 %. Este fenómeno sugiere que las arquitecturas basadas en CNN 3D se perfilan como una estrategia altamente prometedora para concebir soluciones efectivas en el ámbito de la detección de anomalías en videos. La consistente excelencia en el rendimiento AUC resalta la eficacia de estos modelos en la identificación y discriminación de patrones anómalos en secuencias visuales, consolidando su importancia en la creación de sistemas robustos para abordar problemáticas relacionadas con la seguridad y la videovigilancia.

1. Real-world Anomaly Detection in Surveillance Videos
    - Descripción: Se propone realizar el aprendizaje a través de clasificación de múltiples instancias profundas usando etiquetas débiles. Los videos son considerados como bolsas y los segmentos de video como instancias en el enfoque MIL. Se busca localizar las anomalías en el entrenamiento, mediante la implementación de restricciones de escasez y suavidad temporal en la función de pérdida. Se extraen las características de FC6 (4096) del descriptor de video C3D mediante fotogramas de 240x320 píxeles a 30 FPS, las secuencias de video se componen por 16 fotogramas.
    - Conjuntos de datos: UCFCrime
    - Métricas: ROC - AUC
    - Resultados:
        a) UCFCrime (AUC)
            - 75.41 %
    - Ref: [Sultani et al., 2019]
2. Iterative weak/self-supervised classification framework for abnormal events detection
    - Descripción: Implementan 2 enfoques de aprendizaje profundo, uno mediante una red débilmente supervisada y otro con una red auto supervisada, luego usan un clasificador *Random Forest* para fusionar las puntuaciones y obtener mejores resultados. Mencionan que la detección de eventos anormales en imágenes sigue siendo considerado un gran desafío en la investigación.
    - Conjuntos de datos: UCFCrime, UBI-Fights y UCSD
    - Métricas: ROC - AUC
    - Resultados:
        a) UCFCrime (AUC)
            - 76.90 %
        b) UBI-Fights (AUC)

- 84.60 %

- Ref: [Degardin and Proença, 2021]

3. Multiple Instance-Based Video Anomaly Detection using Deep Temporal Encoding-Decoding
   - Descripción: Modelo de codificación-decodificación temporal profunda débilmente supervisada utilizando aprendizaje de instancias múltiples. Proponen una nueva función de pérdida que maximiza la distancia media entre las predicciones de instancias normales y anormales. La principal contribución se basa en que consideran relaciones temporales entre instancias de video empleando una red de codificación-decodificación temporal profunda que está diseñada para capturar la evolución espacio temporal de instancias de video.Las características se extraen mediante C3D.
   - Conjuntos de datos: UCFCrime, ShanghaiTech
   - Métricas: ROC - AUC
   - Resultados:
     a) UCFCrime (AUC)
     - 79.49 %
   - Ref: [Kamoona et al., 2023]

4. 3D ResNet with Ranking Loss Function for Abnormal Activity Detection in Videos
   - Descripción: El método busca minimizar la tasa de falsas alarmas mientras se realiza una tarea de detección de actividad anómala. Implementan una red neuronal profunda 3D ResNet3D para extraer características espaciotemporales. Posteriormente, mediante Aprendizaje de instancias múltiples a nivel de video, realizan la clasificación y proponen *3D Deep Multiple Instance Learning with ResNet (MILR)*
   - Conjuntos de datos: UCFCrime
   - Métricas: ROC - AUC
   - Resultados:
     a) UCFCrime (AUC)
     - 76.67 %
   - Ref: [Dubey et al., 2019]

5. Weakly-Supervised Spatio-Temporal Anomaly Detection in Surveillance Video
   - Descripción: El objetivo es localizar un segmento espacio temporal con solo anotaciones a nivel de video con supervisión durante el entrenamiento mediante la

detección de anomalías espacio temporales débilmente supervisadas (WSSTAD). Formulan el problema como una tarea de MIL. MGPR que tiene como objetivo transferir las abstracciones aprendidas entre ramas para realizar un proceso de refinamiento progresivo.

- Conjuntos de datos: UCFCrime
- Métricas: ROC - AUC
- Resultados:
  a) UCFCrime (AUC)
    - 87.65 %
- Ref: [Wu et al., 2021]

6. Weakly-supervised Joint Anomaly Detection and Classification
   - Descripción: Implementan un método conjunto de clasificación y detección de anomalías mediante aprendizaje débilmente supervisado. El modelo está compuesto por 4 etapas que comprenden la división del vídeo en segmentos, la extracción de características mediante una red neuronal convolucional 3D, un modelo temporal y un proceso de detección y clasificación.
   - Conjuntos de datos: UCFCrime
   - Métricas: ROC - AUC
   - Resultados:
     a) UCFCrime (AUC)
       - 82.12 %
   - Ref: [Majhi et al., 2021]

7. Anomaly Recognition from surveillance videos using 3D Convolutional Neural Networks
   - Descripción: Proporcionar un marco eficaz para el reconocimiento de diferentes anomalías del mundo real a partir de videos mediante un método relacionado con características espaciotemporales utilizando una red neuronal convolucional 3D. Extraen las características con 3D ConvNets.
   - Conjuntos de datos: UCFCrime
   - Métricas: ROC - AUC
   - Resultados:
     a) UCFCrime (AUC)
       - 82.00 %

- Ref: [Maqsood et al., 2021]

8. Dance With Self-Attention: A New Look of Conditional Random Fields on Anomaly Detection in Videos
   - Descripción: El objetivo fue diseñar un esquema MIL contractivo efectivo para ampliar el margen entre las instancias normales y anormales en los videos. Esta tarea fue realizada usando un enfoque débilmente supervisado mediante un extractor de características y una red neuronal convolucional 3D multiescala.
   - Conjuntos de datos: UCFCrime y ShanghaiTech
   - Métricas: ROC - AUC
   - Resultados:
     a) UCFCrime (AUC)
        - 85.00 %
     b) ShanghaiTech [Luo et al., 2017] (AUC)
        - 96.85 %
   - Ref: [Maqsood et al., 2021]
9. Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos
   - Descripción: Se propone una red neuronal convolucional profunda unificada que sea capaz de reconocer y localizar la acción en función de las características de convolución 3D. La idea se basa en dividir el video en clips de igual longitud, luego para cada clip se genera un conjunto de propuestas de "Tube" basadas en características de ConvNet. Luego, las propuestas se vinculan entre sí empleando flujo de red y se realiza la detección de acciones espacio-temporales.
   - Conjuntos de datos: UCF101
   - Métricas: ROC - AUC
   - Resultados:
     a) UCF101 [Soomro et al., 2012](AUC)
        - 86.70 %
   - Ref: [Maqsood et al., 2021]
10. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset
    - Descripción: Implementan un esquema *Two stream* utilizando como extractor de características I3D preentrenado con Kinetics
    - Conjuntos de datos: UCF101

- Métricas: ROC - AUC
- Resultados:
  a) UCF101 (accuracy)
    - 97.80 %
- Ref: [Carreira and Zisserman, 2017]

11. Anomaly Localy in Video Surveillance
    - Descripción: Implementan *Tube Extraction* para la extracción de características de secuencias de fotogramas para pasarlos por I3D y luego por un modelo de regresión con capas de convolución 3D, este resultado se fusiona con las coordenadas del evento dentro de cada fotograma para poder obtener una predicción.
    - Conjuntos de datos: UCFCrime
    - Métricas: ROC - AUC
    - Resultados:
      a) UCFCrime (AUC)
        - 77.52 %
    - Ref: [Landi et al., 2019]

1.6.3. Redes Neuronales de grafos de convolución

1. Graph Convolutional Label Noise Cleaner: Train a Plug-and-play Action Classifier for Anomaly Detection
    - Descripción: Método para limpiar el ruido de las etiquetas en los datos de clasificación utilizando grafos de convolución que permite entrenar un clasificador de acciones para detectar anomalías en el ruido de las etiquetas de los datos. El método de GCLNC se basa en la idea de que las anomalías en las etiquetas de los datos pueden ser representadas como grafos. Implementan una red neuronal de gráficos y consideran un enfoque para el método de aprendizaje de instancias múltiples donde las etiquetas del video son ruidosas o están mal etiquetadas.
    - Conjuntos de datos: UCFCrime, ShanghaiTech y UCSD-Peds
    - Métricas: ROC - AUC
    - Resultados:
      a) UCFCrime (AUC)
        - TSN RGB 82.12 %
        - TSN Optical Flow 78.08 %
    - Ref: [Zhong et al., 2019]

1.6.4. Aprendizaje robusto de magnitud de característica temporal

1. Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning
   - Descripción: Se entrena una función de aprendizaje de magnitud de característica para reconocer de manera efectiva las instancias positivas, mejorando sustancialmente la solidez del enfoque MIL para las instancias negativas. RTMF consiste en que se confía en la magnitud de la característica temporal de los fragmentos de modo que las características con baja magnitud corresponden a eventos normales y las de alta magnitud a eventos anormales. En el método RTMF calculan las características con C3D y con I3D, con estos datos utilizan un clasificador MIL y se realiza una optimización basada en la magnitud media de las características de los k fragmentos principales de un video. Los videos son divididos en 32 segmentos. De C3D extraen 4096 características de fc6 y de I3D extraen 2048 características de mix5c
   - Conjuntos de datos: UCFCrime, ShanghaiTech, XD-Violence y UCSD-Peds
   - Métricas: ROC - AUC, AP (average precision)
   - Resultados:
     a) UCFCrime (AUC)
        - C3D RGB 83.28 %
        - I3D RGB 83.28 %
   - Ref: [Tian et al., 2021]

1.6.5. Enfoque Multimodal Audio y Video

1. Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision
   - Descripción: Detección de violencia poco supervisada como una tarea de aprendizaje de múltiples instancias (MIL) mediante un enfoque multimodal donde consideran audio y video. Principalmente explota la relación entre fragmentos y aprende poderosas representaciones basadas en estas relaciones mediante una red holística y localizada (HL-Net). Además, implementan un aproximador holistic and localized cue (HLC) para la detección de violencia en línea. Usan C3D e I3D como extractores de características e implementan una red neuronal de 3 ramas que se alimenta de información de video y audio para la detección de anomalías, a través de etiquetas débiles. Como extractor de características audio usan VGGish ([Gemmeke et al., 2017], [Hershey et al., 2016])
   - Conjuntos de datos: UCFCrime, ShanghaiTech, XD-Violence y UCSD-Peds

- Métricas: Frame-level precision-recall curve (PRC), ROC-AUC, Average precision (AP)
- Resultados:
  a) UCFCrime (AUC)
    - C3D RGB 82.44 %
- Ref: [Wu et al., 2020]

1.6.6. Generador de Pseudo Etiquetas

1. MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection

   - Descripción: Implementan un *MIL-pseudo label generator*. La idea es generar pseudo etiquetas con mayor precisión que aquellos que simplemente asignan etiquetas de nivel de video a cada clip. Adicionalmente se adopta una estrategia de escasez de muestreo continuo para obligar a la red a prestar más atención al contexto. Implementan también un Codificador de funciones de atención auto-guiada potenciada, los eventos anómalos pueden ocurrir en cualquier lugar y con cualquier tamaño. Además usan C3D e I3D como extractores de características.
   - Conjuntos de datos: UCFCrime, ShanghaiTech
   - Métricas: Frame-level precision-recall curve (PRC), ROC-AUC, Average precision (AP)
   - Resultados:
     a) ShanghaiTech (AUC)
       - C3D RGB 93.13 %
       - I3D RGB 94.83 %

     b) UCFCrime (AUC)
       - C3D RGB 81.40 %
       - I3D RGB 82.30 %
   - Ref: [Feng et al., 2021]

1.6.7. Modelo de codificación de contexto

1. Localizing Anomalies from Weakly-Labeled Videos

   - Descripción: Se propone un modelo de codificación de contexto de alto orden para no solo extraer representaciones semánticas sino también medir las variaciones dinámicas para que el contexto temporal pueda utilizarse de manera efectiva. Se

resalta que la detección de anomalías en video sigue siendo un desafío. Además, el modelo tiene la capacidad de trabajar a 44 FPS por lo que es aplicable a un sistema de detección en linea.

- Conjuntos de datos: UCFCrime, TAD
- Métricas: ROC-AUC
- Resultados:
  a) TAD [Xu et al., 2022](AUC)
    - 89.64 %
  b) UCFCrime (AUC)
    - 85.38 %
- Ref: [Lv et al., 2021b]

1.6.8. Autoencoders

1. Learning Temporal Regularity in Video Sequences

   - Descripción: Se presenta un enfoque basado en autoencoders, la idea principal es que su función objetivo es computacionalmente más eficiente que la codificación escasa y conserva la información espacio-temporal mientras codifica la dinámica. El autoencoder reconstruye el movimiento regular con un error bajo pero incurre en un error de reconstrucción grande para los movimientos irregulares. Se propone un autoencoder para la regularidad temporal basado en 2 características. Características de movimiento "artesanales" y un aprendizaje basado en deep autoencoder con 7 capas completamente conectadas.
   - Conjuntos de datos: CUHK Avenue, UCSD Ped1, UCSD Ped2, Subway Entrance y Subway Exit
   - Métricas: ROC-AUC
   - Resultados:
     a) CUHK Avenue (AUC)
       - 70.20 %
     b) UCSD Ped1 [Mahadevan et al., 2010](AUC)
       - 81.00 %
     c) UCSD Ped2 (AUC)
       - 90.00 %
     d) Subway Entrance [Adam et al., 2008](AUC)

- 94.30 %

e) Subway Exit (AUC)

- 80.70 %

- Ref: [Hasan et al., 2016]

En la Tabla **1-4** se muestra un resumen de los resultados más importantes en cada una de las categorías de trabajos realizados en el ámbito de modelos basados en aprendizaje débilmente supervisado.

Tabla 1-4: Resumen resultados aprendizaje débilmente supervisado

Técnica	**Item**	**Conjunto de datos**	**AUC más alto**
Redes Neuronales Recurrentes	1	UCFCrime2Local	89.05 %
Redes Neuronales Recurrentes	3	HR-Crime	60.30 %
Redes Neuronales Convolucionales 3D	1	UCFCrime	75.41 %
Redes Neuronales Convolucionales 3D	2	UBI-Fights	84.60 %
Redes Neuronales Convolucionales 3D	3	UCFCrime	79.49 %
Redes Neuronales Convolucionales 3D	5	UCFCrime	87.65 %
Redes Neuronales Convolucionales 3D	8	ShanghaiTech	96.85 %
Redes Neuronales Convolucionales 3D	9	UCF101	86.70 %
Redes Neuronales de grafos de convolución	1	UCFCrime	82.12 %
Aprendizaje robusto de magnitud de característica temporal	1	UCFCrime	83.28 %
Enfoque multimodal Audio y Video	1	UCFCrime	82.44 %
Generador de pseudo etiquetas	1	ShanghaiTech	94.83 %
Generador de pseudo etiquetas	1	UCFCrime	82.30 %
Modelo de codificación de contexto	1	TAD	89.64 %
Autoencoders	1	Subway Entrance	94.30 %

1.7. Resultados de métodos de aprendizaje supervisado

Este segmento de análisis se concentra en los enfoques de aprendizaje supervisado, particularmente en propuestas que se fundamentan en modelos que incorporan redes neuronales convolucionales 3D (CNN 3D). Estas propuestas emplean las CNN 3D ya sea como parte del proceso de extracción de características o como la columna vertebral central para la clasificación. Los resultados obtenidos hasta ahora revelan un desempeño favorable en términos del valor de AUC. No obstante, algunos de estos estudios señalan la posibilidad de mejoras adicionales para alcanzar resultados aún más precisos. Además, se destaca la necesidad de conjuntos de datos más detallados y extensamente etiquetados. Estos conjuntos de datos son concebidos como fuente de datos para futuras investigaciones, están diseñados para contribuir al perfeccionamiento de los métodos de aprendizaje supervisado, abordando así las áreas donde aún se pueden fortalecer los modelos.

1. Learning Spatiotemporal Features with 3D Convolutional Networks
   - Descripción: Proponen aprendizaje de características espaciotemporales usando un modelo de red profunda 3D ConvNet. Usan C3D que tiene propiedades como descriptor de videos para extraer las características. Con base en diversos experimentos encontraron los ajustes adecuados para la arquitectura de la red C3D que fue entrenado con Sport1M. Para usarlo como descriptor de video se basa en tomar 16 fotogramas de un video solapados con 8 fotogramas entre un video y el siguiente, estos datos se pasan por C3D y se extraen las características de fc6 que entrega un vector de dimensión 4096.
   - Conjuntos de datos: Sport1M, UCF101, ASLAN, YUPENN y UMD
   - Métricas: ROC-AUC
   - Resultados:
     a) Sport1M [Karpathy et al., 2014](AUC) (Reconocimiento de acciones)
        - 85.20 %

     b) UCF101 (AUC) (Reconocimiento de acciones)
        - 85.20 %

     c) ASLAN [Kliper-Gross et al., 2012](AUC) (Etiquetado de similitud de acción)
        - 96.50 %

     d) YUPENN [Ullah and Petrosino, 2017](AUC) (Clasificación de escenas)
        - 98.10 %

     e) UMD (AUC) (Clasificación de escenas)
        - 87.70 %

- Ref: [Tran et al., 2015]

2. RWF-2000: An Open Large Scale Video Database for Violence Detection
   - Descripción: Proponen un método de 3D CNN en conjunto con flujo óptico para el reconocimiento de comportamientos violentos y lo llaman *Flow Gated Network*. Se toma el reconocimiento de comportamientos violentos como una sub-área del tema de reconocimiento de la acción humana. En este presentan una estructura del modelo que cuenta con 2 entradas el canal RGB y el canal de flujo óptico, además de un bloque de fusión y una capa completamente conectada. La entrada al sistema es de la siguiente forma: 64*224*224*5 donde 5 es el número de canales, 3 corresponden a RGB y 2 a los canales de flujo óptico. En primer lugar, emplearon el método de Gunner Farneback [Farnebäck, 2003] *Two-frame motion estimation based on polynomial expansion* para calcular el flujo óptico denso entre fotogramas vecinos. Con esto se obtiene un vector 2D de desplazamiento que genera un mapa de calor para indicar la intensidad de movimiento. Luego se suman los mapas de calor para obtener un mapa de intensidad de movimiento final.
   - Conjuntos de datos: RWF-2000
   - Métricas: Accuracy
   - Resultados:
     a) RWF-2000 (Accuracy)
        - C3D 85.20 %

     b) RWF-2000 (Accuracy)
        - Flow Gated Network 87.25 %
   - Ref: [Cheng et al., 2021]

3. Anomaly Detection in Video Sequences: A Benchmark and Computational Model
   - Descripción: Mencionan que unos de los principales problemas en las tareas de detección de anomalías en video es la ausencia de conjuntos de datos con anotaciones finas. Por lo que presentan un Dataset Benchmark LAD con Etiquetas a nivel de fotograma y gracias a esto proponen un método de aprendizaje completamente supervisado mediante una red neuronal profunda. El modelo se basa en extraer características usando I3D(Kinetics400), estas se pasan por un *global context-aware stream* para aprender características de alto nivel y posteriormente por 2 capas de convolución LSTM para aprender las características espacio temporales. Los datos de entrada son 16 fotogramas no solapados de 224x224
   - Conjuntos de datos: UCFCrime
   - Métricas: ROC-AUC

- Resultados:
  a) UCFCrime (AUC)
     - 74.98 %
- Ref: [Wan et al., 2021]

4. UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection
   - Descripción: CycleGAN introducen eventos anormales anotados a nivel de píxel en el momento del entrenamiento, permitiendo por primera vez el uso de métodos de aprendizaje totalmente supervisados para la detección de eventos anormales.
   - Conjuntos de datos: UBnormal
   - Métricas: ROC-AUC
   - Resultados:
     a) UBnormal (AUC)
        - 86.50 %
   - Ref: [Acsintoae et al., 2021]

5. Detection of Real-world Fights in Surveillance Videos
   - Descripción: Evalúan diferentes extractores de características como:CNN de flujos, 3D CNN, Descriptor de punto de interés local, Diversos clasificadores como: CNN, LSTM, SVM
   - Conjuntos de datos: CCTV-Fights.
   - Métricas: mAP y F-Measure
   - Resultados:
     a) CCTV-Fights. (mAP) (Two-Stream clasificador CNN)
        - 79.50 %
     b) CCTV-Fights. (F-Measure)
        - 75.00 %
     c) CCTV-Fights. (mAP) (Two-Stream clasificador LSTM)
        - 76.00 %
     d) CCTV-Fights. (F-Measure)
        - 75.90 %
   - Ref: [Perez et al., 2019]

En la Tabla **1-5** se muestra un resumen de los resultados más importantes en cada una de las categorías de trabajos realizados en el ámbito de modelos basados en aprendizaje supervisado.

Tabla 1-5: Resumen resultados aprendizaje supervisado

Item	**Conjunto de datos**	**AUC más alto**
1	YUPENN	98.10 %
2	RWF2000	87.25 %
3	UCFCrime	74.98 %
4	UBnormal	86.90 %

A partir de los resultados observados en el análisis de los modelos LSTM, RNN y CNN 3D para la detección de eventos anómalos en videos de videovigilancia, se observa una variada gama de enfoques, arquitecturas y métricas de evaluación. La diversidad de conjuntos de datos utilizados, como UCFCrime, XD-Violence, LAD, entre otros, proporcionan una base para la exploración y validación de modelos. Las propuestas presentadas destacan la importancia de la elección de métricas apropiadas para evaluar el rendimiento de los modelos. Además, es indispensable considerar el contexto y las limitaciones específicas de cada conjunto de datos, ya que esto puede influir en la interpretación de los resultados. La comparación entre modelos, también aporta información valioso sobre las eficacias relativas de los diferentes modelos y extractores de características. Esté análisis constituye un paso fundamental para comprender la eficacia y las limitaciones de los modelos en la detección de eventos anómalos en contextos de videovigilancia.

Es importante señalar que la interpretación de algunos resultados puede estar influenciada por posibles sesgos presentes en los conjuntos de datos utilizados en estos estudios. La diversidad y representatividad de las muestras son factores determinantes para la validez de los modelos desarrollados. Algunos de los conjuntos de datos pueden estar sesgados hacia ciertos escenarios o patrones, lo que podría limitar la capacidad de generalización de los modelos. Este sesgo podría introducir desafíos significativos, especialmente cuando se trata de eventos anómalos que pueden manifestarse en diversas situaciones del mundo real.

Además, es indispensable considerar la posible falta de muestras en algunos conjuntos de datos para eventos específicos. La carencia de diversidad en la cantidad de ejemplos puede afectar negativamente la capacidad de los modelos para aprender y generalizar patrones más amplios. Es esencial reconocer estas limitaciones al interpretar los resultados y al extrapolar el rendimiento de los modelos a situaciones fuera del ámbito de entrenamiento. Es por esto que la búsqueda de conjuntos de datos más equilibrados y representativos es un paso clave

para futuras investigaciones, garantizando así una evaluación más rigurosa y confiable de los modelos.

1.8. Metodología

- Esta tesis se enmarca en el proyecto "Administración inteligente de problemas de seguridad ciudadana a través de modelos y herramientas generados a partir de plataformas para territorios inteligentes apoyadas por estrategias de participación ciudadana en la ciudad de Medellín", colaboración entre la Universidad de Antioquia y la Universidad Nacional de Colombia. El enfoque de esta investigación deriva directamente de la línea de investigación del proyecto. Considerando que el objetivo principal del proyecto es la seguridad ciudadana, esta tesis tiene como propósito fundamental aplicar técnicas de aprendizaje de máquina para la detección de delitos mediante el análisis de elementos multimedia. Por tanto, es importante identificar en primera instancia las técnicas más relevantes utilizadas en años recientes para abordar este desafío.

- Se establecen como criterios de selección para las técnicas y modelos aquellos que presenten estructuras flexibles, permitiendo una eficiente incorporación de nuevas variables. La metodología de investigación adoptada es de corte experimental, lo que resalta la necesidad de determinar parámetros que ofrezcan los resultados óptimos. Por lo que se propone dirigir la búsqueda hacia modelos fundamentados en el aprendizaje de máquina. Estos sistemas ofrecen la versatilidad de integrar nuevos datos y parámetros para el análisis de información. Adicionalmente, se contemplan los algoritmos basados en esquemas espacio-temporales debido a su amplio uso en análisis de video, especialmente en aplicaciones que requieren considerar el contexto de los eventos.

- Para la implementación de las técnicas seleccionadas, es esencial contar con conjuntos de datos extensos para calibrar los modelos y técnicas. Sin embargo, debido a las limitaciones temporales inherentes a la duración de la maestría y la ejecución del proyecto, la etapa de recolección de datos no se considera viable. La generación de un conjunto de datos completo y diverso, abarcando una variedad de tipos de eventos y un gran volumen de repeticiones, demandaría una infraestructura instalada y un considerable período de tiempo.

- En consecuencia, la investigación se fundamenta en simulaciones que se nutren de la información extraída de conjuntos de datos públicos disponibles en línea. Estos conjuntos han sido creados por la comunidad académica y se orientan hacia la resolución del desafío de detección de delitos. La elección de los conjuntos de datos se basa en las métricas de rendimiento de trabajos anteriores realizados por diversos investigadores, siempre enfocándose en aquellos que exhiban resultados satisfactorios.

- A su vez, se debe determinar al menos un tipo de delito con base en las anomalías presentes en los conjuntos de datos. Esta decisión es crucial debido a la necesidad de

contar con un número considerable de repeticiones para el desarrollo eficaz del modelo de predicción.

- Con base en las técnicas y conjuntos de datos elegidos, surge la tarea de diseñar un modelo de aprendizaje de máquina que sea adaptable a situaciones reales, específicamente dirigido a la detección de delitos. Durante esta etapa, el enfoque radica en llevar a cabo un análisis exhaustivo de las técnicas más relevantes. El propósito es identificar oportunidades de mejora, explorar opciones de integración con otras metodologías ó aplicar métodos de selección de parámetros, entre otros factores que se relacionen con los criterios de rendimiento y eficiencia necesarios para su aplicación en un caso de estudio real. En esencia, esta fase busca perfeccionar el modelo de tal manera que no solo cumpla con los estándares de rendimiento, sino que también asegure una velocidad de procesamiento adecuada para su implementación en un entorno de uso real.

- Finalmente, se debe someter el modelo a una evaluación rigurosa utilizando los diversos conjuntos de datos previamente seleccionados. Esta evaluación tiene como objetivo principal validar su rendimiento. Los resultados se proponen medir a través de métricas y parámetros fundamentales de desempeño, exactitud y precisión, ya que estas métricas son las más ampliamente empleadas en esta categoría de investigaciones.

2 Desarrollo

En este capítulo se abordan los temas fundamentales relacionados con los conjuntos de datos que serán considerados para el desarrollo del modelo. La selección de estos conjuntos se basa primordialmente en la naturaleza de su contenido, se da prioridad a aquellos que estén directamente vinculados con la detección de delitos o violencia en videos. Además, se lleva a cabo un análisis de las técnicas utilizadas en el ámbito de la detección de anomalías en video, especialmente las que están relacionadas con los enfoques de aprendizaje débilmente supervisado y aprendizaje supervisado. Tras la selección de los conjuntos de datos y el análisis de las técnicas existentes, el siguiente paso consiste en proponer una arquitectura basada en dicho análisis y establecer una metodología para su evaluación. Este enfoque busca establecer la base para el desarrollo y la evaluación del modelo en harás de lograr una solución efectiva para la detección de eventos delictivos en entornos de videovigilancia.

2.1. Conjuntos de datos

La calidad y relevancia de los conjuntos de datos se ha convertido en un factor crítico para el éxito de cualquier modelo de aprendizaje automático. Detrás de cada sistema inteligente que reconoce rostros, sugiere productos o predice enfermedades, yace un conjunto de datos sólido y diverso que alimenta la capacidad de estas máquinas para generalizar e identificar patrones en los datos.

El entrenamiento de un modelo de aprendizaje automático es comparable a enseñar a un niño. Así como un niño requiere de exposición a una amplia variedad de situaciones y ejemplos para comprender y aprender a reconocer objetos y conceptos, un modelo de aprendizaje automático requiere una rica fuente de datos para aprender las relaciones y patrones entre diferentes atributos. En este contexto, los conjuntos de datos son la "experiencia" que un modelo recopila para comprender el mundo que lo rodea.

El papel de los conjuntos de datos va más allá de simplemente nutrir el aprendizaje del modelo. La calidad de los datos utilizados para el entrenamiento directamente impacta en la capacidad del modelo de generalizar y adaptarse a nuevas situaciones. Un conjunto de datos bien construido debe reflejar la diversidad y complejidad del problema que el modelo intentará resolver en el mundo real. Si se presentan datos insuficientes o sesgados, el modelo

puede enfrentar dificultades para tomar decisiones precisas y relevantes en situaciones reales.

En el ámbito de la inteligencia artificial, los datos son el cimiento sobre el cual se construye toda la estructura del aprendizaje automático. Los modelos avanzados y las técnicas innovadoras pueden tener un impacto significativo, pero sin una base sólida de datos, su potencial queda limitado. Por lo tanto, la selección, preparación y limpieza de los conjuntos de datos de alta calidad se ha convertido en una disciplina crucial para aquellos que buscan desarrollar sistemas inteligentes confiables y eficaces.

Las técnicas empleadas para la detección de violencia en videos se fundamentan en el análisis de características visuales. Han surgido varios modelos de aprendizaje profundo, como las redes neuronales convolucionales 3D y las redes neuronales recurrentes (RNN), destinados a la identificación de violencia en videos. Estos modelos tienen la capacidad de identificar patrones de movimiento gracias a la componente espacial que incorporan. Sin embargo, persisten desafíos en la detección de violencia en videos debido a diversas limitaciones. Uno de estos obstáculos es la carencia de datos etiquetados. Los conjuntos de datos con etiquetas son esenciales para entrenar y validar los modelos de aprendizaje profundo en aplicaciones como por ejemplo la detección de delitos en videos. Adicionalmente, detectar este tipo de situaciones en tiempo real demanda considerables recursos de procesamiento y memoria.

Otra dificultad radica en detectar delitos en contextos y situaciones diversas. Un delito puede acontecer en diferentes escenarios, como calles, hogares o escuelas. Cada contexto puede presentar características visuales únicas que los sistemas de detección de violencia deben reconocer. El incremento en el uso de tecnologías de videovigilancia ha generado la necesidad de desarrollar técnicas específicas para la detección de delitos y anomalías en videos. Con el objetivo de abordar esta problemática, se han elaborado numerosos conjuntos de datos que permiten entrenar modelos de aprendizaje automático. Estos modelos son empleados para identificar acciones sospechosas, como robos, peleas o incidentes de violencia. Los conjuntos de datos son sumamente valiosos debido a la cantidad de ejemplos que proporcionan, así como a la inclusión de información etiquetada, como la ubicación de las personas, el tipo de evento y, en algunos casos, anotaciones a nivel de cuadro y en algunos casos hasta a nivel de píxel.

Uno de los conjuntos de datos más ampliamente utilizados para la detección de delitos o violencia en videos es UCFCrime [Sultani et al., 2019], desarrollado por la Universidad de Florida Central. Este conjunto de datos abarca categorías realistas que incluyen peleas, vandalismo, robos, entre otros. A partir de UCFCrime se han llevado a cabo numerosos estudios con el propósito de desarrollar técnicas para la detección de eventos anómalos. También se han creado subconjuntos, como UCF-Crime2Local [Landi et al., 2019] y HR-Crime [Boekhoudt et al., 2021], que se enfocan en anomalías relacionadas con humanos y

excluyen videos con problemas de enfoque, entre otros.

Además, se han presentado enfoques que combinan video y audio, como es el caso del dataset XD-Violence, que ofrece 217 horas de video y abarca 6 tipos de violencia. Este conjunto de datos se considera un referente en la detección de violencia, junto con otros conjuntos como LAD (Large-scale Anomaly Detection) [Wan et al., 2021] y RWF2000 (Real-World Fighting) [Cheng et al., 2021]. LAD, en particular, se creó con etiquetas a nivel de fotograma con el objetivo de facilitar desarrollos que involucren métodos de aprendizaje supervisado.

Otros conjuntos de datos enfocados en la detección de violencia o que abarcan anomalías relacionadas con peleas, robos y situaciones similares son Ubi Fights [Degardin and Proença, 2021] y CCTV-Fights [Perez et al., 2019], entre otros. En términos generales, estos conjuntos de datos tienen un valor significativo para el desarrollo de tecnologías de videovigilancia que puedan mejorar la seguridad pública y prevenir delitos. Para conocer más detalles, se pueden consultar otros conjuntos de datos en la Tabla **2-1**.

Tabla 2-1: Conjuntos de datos enfocados en delitos

Nombre	**Muestras**	**Categorías**	**Etiquetado**	**Año**
UCFCrime [Sultani et al., 2019]	1900	13	Vídeo	2018
RWF2000 [Cheng et al., 2021]	2000	2	Vídeo	2019
CCTV-Fights [Perez et al., 2019]	1000	2	Cuadro	2019
RLVS [Soliman et al., 2019]	2000	2	Vídeo	2019
XD-Violence [Wu et al., 2020]	4754	6	Vídeo	2020
Ubi-Fights [Degardin and Proença, 2021]	1000	2	Cuadro	2020
LAD [Wan et al., 2021]	2000	14	Cuadro	2021

Considerando los conjuntos de datos previamente mencionados, se plantea la utilización específica de las categorías centradas en delitos dirigidos hacia personas, dado que constituyen uno de los problemas principales relacionados con la seguridad ciudadana en Colombia. Estos conjuntos de datos, que incluyen categorías realistas como peleas, vandalismo y robos en casos como UCFCrime, XD-Violence, LAD, y otros, no solo ofrecen la diversidad necesaria, sino también las etiquetas cruciales para el proceso de entrenamiento. La abundancia de información presente en estos conjuntos se convierte en la materia prima esencial que permite a los modelos de aprendizaje automático aprender a identificar patrones y realizar predicciones precisas en tiempo real.

2.2. Modelo de aprendizaje de máquina

Para la elaboración del modelo, se deben tener en cuenta diversos aspectos. Dado que el objetivo es construir un modelo de predicción orientado a la detección de delitos hacia personas en videos de videovigilancia, y que sea aplicable en escenarios de la vida cotidiana. Se plantea la necesidad de que la aplicación resultante sea desplegable en cualquier plataforma. En este sentido, se optó por la implementación de contenedores Docker, garantizando así la portabilidad y eficiencia en la ejecución del modelo en diferentes entornos.

Un elemento fundamental es la consideración de que el modelo se entrenará con secuencias de fotogramas, lo cual implica un notable costo computacional durante la fase de entrenamiento. Para abordar este desafío, se ha desarrollado un algoritmo generador de datos de video. Este algoritmo tiene la capacidad de enviar pequeños lotes de videos para entrenar el modelo de manera eficiente, evitando posibles problemas de saturación de memoria. Por último, se ha decidido incorporar en el desarrollo del modelo enfoques basados en LSTM y CNN3D, que han demostrado obtener los mejores resultados en aplicaciones similares. La métrica de desempeño seleccionada es la curva ROC con AUC, considerada la más apropiada para abordar eficazmente la problemática específica que se enfrenta en la detección de delitos en videos de videovigilancia.

2.2.1. Contenedor de Docker

Se ha trabajado en la configuración de una imagen de Docker que incorpora las versiones específicas de TensorFlow y demás librerías utilizadas en este trabajo de tesis. Este enfoque se ha adoptado con el propósito de asegurar una ejecución consistente y portátil de la aplicación. Docker es reconocido por su capacidad de virtualización y la habilidad que tiene para lograr el empaquetado de una aplicación junto con sus dependencias en un contenedor. Este contenedor puede desplegarse sin inconvenientes en sistemas operativos y arquitecturas de hardware diferentes, eliminando preocupaciones de compatibilidad.

En esta etapa del proyecto se debe considerar la necesidad de desarrollar una aplicación que pueda ser ejecutada en una variedad de dispositivos, cada uno con sus propias prestaciones, características e incluso sistemas operativos. Los modelos de este tipo suelen tener requisitos específicos en cuanto a librerías y paquetes, los cuales podrían generar incompatibilidades con la plataforma de ejecución. La elección de Docker como contenedor para el modelo es importante en términos de eficiencia y portabilidad de la aplicación. Al aprovechar Docker, se consigue encapsular todas las dependencias y configuraciones esenciales para la ejecución del modelo en un entorno aislado. Esta estrategia garantiza que el modelo funcione de manera confiable en distintos entornos de implementación, simplificando enormemente el proceso de despliegue.

En la Figura **2-1** se presenta la estructura del archivo Dockerfile diseñado para la generación de una imagen que integra las librerías y versiones cruciales para el despliegue de la aplicación. Esta configuración se basa principalmente en una imagen de TensorFlow versión 2.2.3, que permite la explotación de unidades de procesamiento gráfico (GPU). La inclusión de una GPU es esencial para dar la capacidad a los sistemas encargados de procesar volúmenes sustanciales de datos gráficos en intervalos de tiempo reducidos. Además, es necesario incorporar librerías fundamentales como Pandas, destinada al procesamiento y análisis de datos, Scikit-learn, una biblioteca que proporciona herramientas esenciales para el aprendizaje automático, y Keras, que facilita la implementación de redes neuronales, entre otras librerías. Estos componentes contribuyen significativamente a tareas diversas, como el procesamiento de imágenes, la visualización de datos y el manejo de herramientas matemáticas.

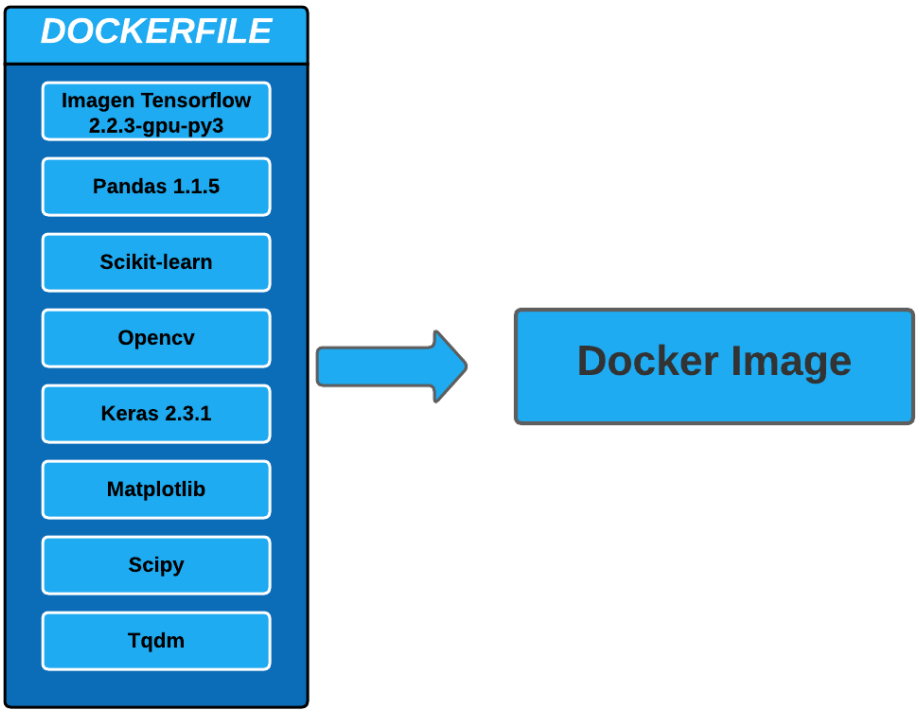


Figura 2-1: DockerFile.

La implementación de esta imagen de Docker simplifica el proceso de despliegue de la aplicación y brinda flexibilidad en términos de ejecución. Gracias a esto, la aplicación puede ser ejecutada eficientemente en una variedad de dispositivos que cumplan con los requisitos mínimos de Docker y dispongan de una GPU. Este elemento representa un gran avance para la fase de experimentación y pruebas, ya que garantiza la portabilidad del modelo desarrollado. Al tener la capacidad de desplegar la aplicación en diferentes equipos de cómputo, se

facilita la validación de su rendimiento en entornos diversos, lo que es crucial para evaluar su robustez.

2.2.2. Generador de datos de video

La fase de carga de datos durante el entrenamiento de un modelo de aprendizaje automático con videos es crítica. Es necesario considerar que al cargar la totalidad del conjunto de datos en la memoria, se enfrentan restricciones relacionadas con la capacidad de almacenamiento en la memoria RAM del equipo de cómputo. En términos prácticos, aproximadamente cada minuto de grabación en calidad 720p a 30 FPS consume alrededor de 60 MB. Extrapolando esta información a una hora, se obtiene un consumo de memoria de aproximadamente 3.6 GB. Tomando como ejemplo un conjunto de datos significativo como UCFCrime, que abarca 1900 muestras representativas de 128 horas de video, se requerirían unos considerables 460 GB. Este hecho se convierte en un problema sustancial a medida que el tamaño del conjunto de datos aumenta. Por esta razón, se ha dedicado un esfuerzo al desarrollo de un generador de datos de video, una iniciativa indispensable para optimizar y acelerar el proceso de entrenamiento de modelos de aprendizaje profundo que trabajan con secuencias de video.

En el esquema del pseudocódigo 1, se evidencia la estructura del generador de datos de video desarrollado. Este generador se compone principalmente de seis funciones, siendo una de ellas el constructor encargado de la inicialización del objeto generador de datos. Las demás funciones desempeñan diversas tareas esenciales, como el envío aleatorio de lotes de datos, la carga de muestras, la ejecución de operaciones de preprocesamiento en las imágenes, y el método principal que se ocupa de enviar los paquetes de datos según el parámetro “batch size” Este diseño modular garantiza una organización clara y eficiente del generador, permitiendo un control preciso sobre las diferentes etapas del proceso de carga y preprocesamiento de datos.

Algorithm 1 Clase generador de datos de video

function Generador de datos de video (ruta de datos raíz, paso temporal, longitud temporal, redimensionar)
function Generador de archivos(directorio de datos, archivos de datos)
function Cargar muestras(Categoría de datos)
function Mezclar datos(muestras)
function Preprocesamiento de imagen(imagen)
function Generador de datos(datos, tamaño de lote, mezclar)

Mediante el método ”Generador de archivos”del pseudocódigo 2, se pretende explorar los directorios que contienen los archivos de los fotogramas de las secuencias de video. El objetivo principal es extraer una cantidad específica de muestras de cada video, según el número

predefinido de fotogramas. Por ejemplo, si un video consta de aproximadamente 200 fotogramas y se establece una longitud temporal de 50, el código determinará que se deben extraer 4 muestras individuales para su posterior uso en el entrenamiento del modelo. Este método devuelve la cantidad de muestras junto con sus correspondientes etiquetas, proporcionando así un mecanismo eficiente para estructurar el conjunto de datos.

Algorithm 2 Generador de archivos

```
function Generador de archivos(directorio de datos, archivos de datos)
    for each archivo f in archivo de datos do
        Leer datos de archivo csv f
        Extraer etiquetas e imágenes de los datos
        if total de imágenes ≥ Longitud temporal then
            Generar muestras con longitud temporal y paso temporal
            yield muestras, etiquetas
        end if
    end for
end function
```

El método de carga de muestras, visualizado en el pseudocódigo 3, recibe como parámetro el nombre de la categoría de datos, que puede ser ya sea datos de entrenamiento, validación o evaluación. Su función principal consiste en recuperar y proporcionar una nueva muestra perteneciente a la categoría especificada.

Algorithm 3 Cargar muestras

```
function Cargar muestras(Categoría de datos)
    Ruta de datos ← join(ruta de datos raíz, Categoría de datos)
    archivo csv ← lista de archivos en directorio de datos
    Inicializa el generador de archivos
    Iterate sobre el generador de archivos y almacena las muestras en la lista de datos
    return lista de datos
end function
```

En lo que respecta a la fase de envío de muestras de manera aleatoria y al preprocesamiento de imágenes, estas funciones fundamentales se pueden apreciar en el código presentado en el pseudocódigo 4. En estas funciones, la imagen se redimensiona y normaliza, operaciones esenciales para llevar a cabo experimentos con entradas de datos de diversas resoluciones. Este enfoque experimental tiene el propósito de analizar cómo influye el tamaño de la imagen en la velocidad de procesamiento y en la eficacia de las predicciones del modelo.
Finalmente, se encuentra el método generador de datos del pseudocódigo 5. Este método es central, ya que en él se define el tamaño del lote y si los datos deben organizarse de manera

Algorithm 4 Datos aleatorios y preprocesar imagen

```
function MEZCLAR DATOS(muestras)
    mezclar datos
    return muestras
end function
function PREPROCESAMIENTO DE IMAGEN(imagen)
    redimensionar imagen a (x, y)
    normalizar imagen a valor
    return imagen
end function
```

ordenada o aleatoria. Su función principal es proporcionar un paquete que contenga muestras y etiquetas del tamaño del lote, listo para ser procesado por el modelo.
La idea de esta clase para generar datos de video es que en lugar de cargar todo el conjunto de datos en la memoria se logre que el generador ejecute una carga dinámica de datos según sea necesario, lo que resulta en un uso más eficiente de los recursos y una aceleración significativa del proceso de entrenamiento.

Además de la eficiencia en la gestión de la memoria, este generador de datos también ofrece la flexibilidad de generar conjuntos de datos de diversos tamaños y formatos, lo que resulta fundamental para evaluar el rendimiento del modelo en una amplia gama de situaciones y escenarios para la tarea detección de delitos en video.

2.2.3. Arquitectura del modelo

Con base en la información recopilada sobre las técnicas utilizadas para la detección de anomalías en videos, se destaca la consistencia de los enfoques que involucran redes neuronales convolucionales 3D y modelos LSTM. Por esta razón, se toma la decisión de desarrollar una arquitectura que integre ambos componentes, aprovechando su capacidad para analizar datos tridimensionales, una característica indispensable para el desarrollo de sistemas que buscan analizar archivos de video.

Hasta ahora, muchas de las propuestas presentadas por la comunidad académica han arrojado resultados en un rango de aproximadamente 80 % a 95 % de AUC, indicando así un buen rendimiento en términos de capacidad de discriminación. Sin embargo, es crucial tener en cuenta que varias de estas propuestas se han desarrollado con conjuntos de datos que carecen de un número suficiente de muestras para lograr una generalización efectiva en la predicción. Este problema se debe en parte a que hasta ahora están surgiendo nuevos conjuntos de datos especializados en la detección de delitos o violencia, un área que está experimentando un crecimiento reciente. Es esencial continuar contribuyendo a este campo, trabajando en la

Algorithm 5 Generador de datos

```
function Generador de datos(datos, tamaño de lote, mezclar)
    numero de muestras ← longitud de datos
    if mezclar then
        Mezclar datos(muestras)
    end if
    while True do
        for i in tamaño de lote do
            muestras del lote ← datos[i]
            inicializar array de muestras y etiquetas
            for each muestra del lote in muestras del lote do
                data ← muestra del lote[0]
                etiqueta ← muestra del lote[1]
                inicializar lista de datos temporales
                for each imagen in muestra do
                    try:
                    imagen ← cv2.imread(imagen)
                    imagen ← Preprocesar imagen(imagen)
                    lista de datos temporal.append(imagen)
                end for
                array muestra.append(lista de datos temporal)
                array etiqueta.append(etiqueta)
            end for
            array muestra ← numpy array of array muestra
            array etiqueta ← numpy array of array etiqueta
            yield array muestra, array etiqueta
        end for
    end while
end function
```

creación de conjuntos de datos robustos que cuenten con una amplia variedad de muestras para cada clase, etiquetas precisas que permitan el trabajo con distintos niveles de aprendizaje supervisado, y muestras que aborden la diversidad de escenarios, ambientes y situaciones. Estos elementos son cruciales para entrenar modelos capaces de reconocer situaciones del mundo real de manera efectiva.

En la Figura **2-2** se observa la propuesta de arquitectura empleada para la ejecución de experimentos. El objetivo es lograr un modelo que a partir de capas de convolución 3D y capas LSTM, alcance un nivel de desempeño importante para la detección de delitos. En este contexto, se consideraron cuidadosamente los elementos listados en la Tabla **2-2** donde se describe su principal utilidad dentro de la arquitectura planteada.

Tabla 2-2: Elementos que componen el modelo de aprendizaje de máquina

Nombre	**Descripción**
Batch Normalization	Esta capa de normalización por lotes busca acelerar el proceso de entrenamiento al tiempo que mejora la generalización del modelo.
Convoluciones 3D y MaxPooling 3D	Estas capas desempeñan un papel importante al analizar los datos tanto en dimensiones espaciales como temporales, buscando patrones en la información a lo largo del tiempo.
Dropout	La inclusión de capas de dropout es esencial para prevenir el sobreajuste durante el entrenamiento, mejorando así la capacidad del modelo para generalizar a datos no vistos.
Filtros capas convolucionales	Estos filtros contribuyen al proceso de convolución a lo largo de las tres dimensiones de los datos de entrada. Actúan como mecanismo de reducción de dimensionalidad, permitiendo la creación de representaciones más compactas.
Capas Dense y Sigmoidal	Estas capas se configuran para llevar a cabo una clasificación binaria, ya que el objetivo primordial es detectar la presencia o ausencia de situaciones delictivas en las secuencias de video analizadas.
Capas LSTM	Las capas LSTM desempeñan un papel fundamental en el análisis y la detección de patrones en la dimensión temporal debido a su capacidad para analizar secuencias de datos y memorizar información a largo plazo.

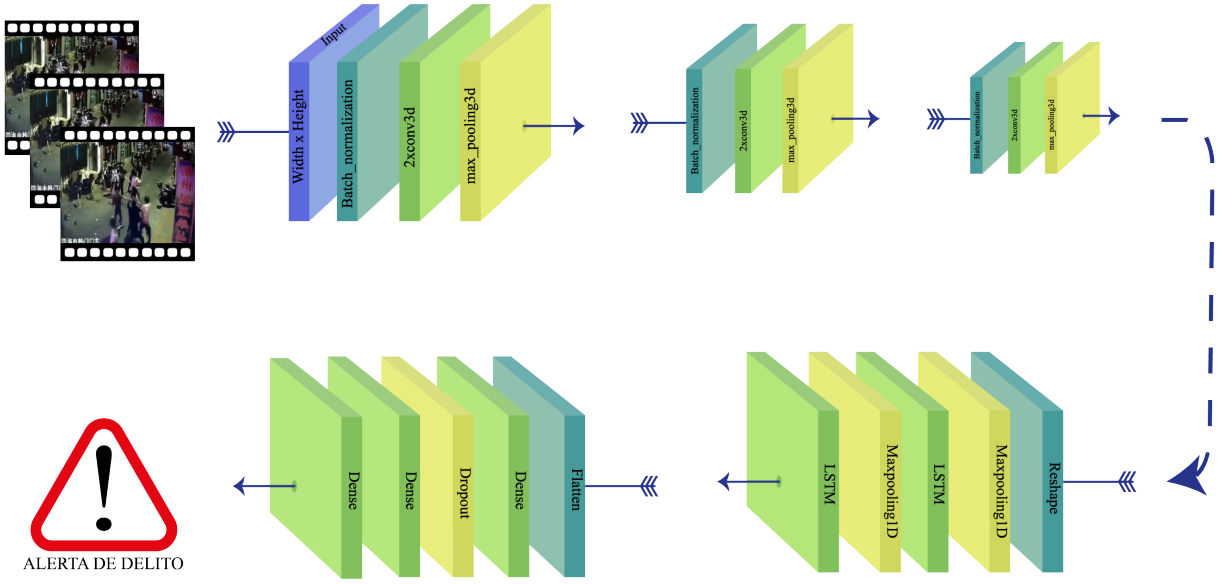


Figura 2-2: Arquitectura.

2.2.4. Metodología evaluación del modelo

La determinación de cuándo un modelo está preparado para ser implementado en producción es un proceso influenciado por varios factores. Uno de los aspectos más fundamentales es el objetivo final del proyecto y el alcance que se pretende lograr. Sin embargo, la calidad y cantidad de los datos disponibles para el entrenamiento del modelo representan un desafío que puede obstaculizar este proceso. La idoneidad de un modelo para la producción no solo depende de su rendimiento en la fase de entrenamiento, sino también de su capacidad para generalizar y adaptarse a situaciones del mundo real, un logro que se facilita significativamente con conjuntos de datos sólidos y diversos.
Un proceso iterativo que se puede llevar a cabo para el desarrollo y validación del modelo es el siguiente:

1. **Preparación de datos:** Asegurarse de que los datos sean adecuados y estén correctamente preparados para el entrenamiento y la evaluación del modelo.

2. **Diseño y entrenamiento del modelo:** Desarrollar y entrenar el modelo utilizando los datos disponibles.

3. **Validación:** Evaluar el rendimiento del modelo utilizando técnicas de validación para estimar su capacidad de generalización.

4. **Ajuste y optimización:** Realizar ajustes en los hiper-parámetros del modelo para mejorar su rendimiento.

5. **Datos de prueba independientes:** Evaluar el modelo en datos que no se utilizaron durante el entrenamiento ni la validación.

6. **Interpretación y explicabilidad:** Comprender cómo toma decisiones el modelo y si estas son interpretables y coherentes con el dominio del problema.

7. **Validación en producción:** Implementar el modelo en un entorno de producción controlado y supervisar su rendimiento en el mundo real.

8. **Monitoreo continuo:** Continuar monitoreando el rendimiento del modelo después de su despliegue y realizar ajustes según sea necesario.

3 Resultados

En esta sección, se presentan y analizan los resultados derivados del trabajo centrado en la aplicación de técnicas de aprendizaje de máquina en el contexto de la seguridad ciudadana. El objetivo principal de este trabajo es desarrollar una aplicación capaz de detectar delitos mediante el análisis de elementos multimedia. Para alcanzar este propósito, se llevaron a cabo diversas etapas, que incluyeron la revisión y selección de técnicas, así como la elección de conjuntos de datos para el entrenamiento del modelo y, finalmente, el desarrollo del modelo.

3.1. Revisión y selección del conjunto de datos

Uno de los aspectos críticos al iniciar el desarrollo de una aplicación relacionada con técnicas de aprendizaje de máquina es la disponibilidad de datos, estas técnicas requieren grandes cantidades de datos para aprender a distinguir entre las clases o categorías que se desean analizar. Para abordar este paso, esencialmente se pueden seguir dos enfoques. El primero implica la generación de un conjunto de datos propio, que conlleva a la tarea de adquirir, almacenar y etiquetar los datos. En este caso, la calidad, cantidad y diversidad de las muestras son consideraciones fundamentales, y es indispensable cumplir con regulaciones y consideraciones éticas relacionadas con la privacidad de la información. El segundo enfoque se centra en el uso de datos públicos, donde los conjuntos de datos están disponibles para la comunidad académica, proporcionados por instituciones académicas, empresas o incluso el gobierno. Independientemente del camino elegido, es esencial evaluar si los datos utilizados en el desarrollo del modelo de aprendizaje de máquina cumplen con ciertas características.

Evaluar un conjunto de datos implica analizar diversos aspectos para determinar su calidad, utilidad y aplicabilidad en contextos específicos. Por ejemplo, en [Oh, 2011], los autores proponen un método de evaluación centrado en la superposición de categorías. Destacan que la calidad del conjunto de datos impacta significativamente en el proceso de clasificación, estableciendo la importancia de la necesidad de un método que evalúe esta calidad. En este caso, y basándose en las características extraídas de cada muestra ubicadas en un espacio bidimensional, el método se fundamenta en la proporción de área superpuesta entre categorías, estableciendo que a mayor área, menor precisión.

Por otro lado, según [Althnian et al., 2021], la tarea de clasificación se torna especialmente desafiante en conjuntos de datos pequeños. La principal razón de este desafío radica en que

con un tamaño limitado de datos, existe el riesgo de sesgar la clasificación. En este estudio, se emplearon seis modelos de clasificación distintos y seis métricas diversas, junto con conjuntos de datos de distintos tamaños, con el objetivo de evaluar el impacto del tamaño del conjunto de datos en el rendimiento. La conclusión alcanzada en esta investigación sugiere que construir un conjunto de datos suficientemente representativo puede ser incluso más crucial que la elección del clasificador.

Con base en lo anterior se plantean algunos criterios clave que se pueden utilizar para calificar un conjunto de datos.

- **Representatividad:** ¿El conjunto de datos logra representar a la población o el fenómeno que se esta estudiando?
- **Diversidad:** Se debe considerar la diversidad en las instancias del conjunto de datos. Un conjunto de datos diverso abarca diferentes situaciones, contextos y condiciones, lo que mejora la generalización del modelo.
- **Tamaño:** Se debe evaluar el tamaño del conjunto de datos en términos de la cantidad de instancias y la cobertura temporal. Un conjunto de datos más grande y variado puede ser beneficioso, pero también es importante equilibrar el tamaño con la calidad y la representatividad.
- **Equilibrio de Clases:** En problemas de clasificación es importante verificar si hay un equilibrio adecuado entre las clases. Un conjunto de datos desequilibrado puede sesgar los modelos hacia clases dominantes.
- **Etiquetado:** Examinar la calidad del etiquetado. Los datos etiquetados de manera precisa y consistente son esenciales para el entrenamiento de modelos de aprendizaje supervisado.
- **Consistencia Temporal:** Si el conjunto de datos es temporal, verificar si hay consistencia temporal en las etiquetas y las características. Esto es especialmente relevante si se están modelando eventos que cambian con el tiempo.
- **Disponibilidad y Accesibilidad:** La disponibilidad y accesibilidad del conjunto de datos es esencial. Debe ser fácilmente accesible para otros investigadores o profesionales que deseen utilizarlo para investigaciones adicionales o aplicaciones prácticas.
- **Origen y Metodología de Recopilación:** Comprende el origen del conjunto de datos y la metodología utilizada para su recopilación. La transparencia en estos aspectos es fundamental para entender posibles sesgos o limitaciones que se puedan presentar.

- **Licencia y Ética:** Verificar la licencia del conjunto de datos y asegurarse de que su uso sea ético y legal. Es importante cumplir con las regulaciones y considerar la privacidad de las personas involucradas.
- **Impacto en la Comunidad:** Considerar el impacto potencial del conjunto de datos en la comunidad académica y profesional. Un conjunto de datos que cumpla con todos los requerimientos puede contribuir significativamente al avance del conocimiento o a la solución de problemas del mundo real.
- **Comparación con Conjuntos de Datos Existentes:** Realizar comparaciones con conjuntos de datos existentes en el mismo dominio permitirá identificar las fortalezas y debilidades del conjunto de datos.

A partir de estos criterios y con el objetivo de establecer con que datos se realiza el entrenamiento y desarrollo del modelo, se plantea analizar algunos de los conjuntos de datos enfocados en detección de delitos que se encuentran disponibles con el objetivo de que cumplan con los criterios descritos en la Tabla **3-1**

Tabla 3-1: Criterios de evaluación para un conjunto de datos

Criterio	Descripción
Representatividad	Contener muestras de delitos hacia personas, en diversos ambientes y escenarios.
Diversidad	Incluir eventos como peleas, robos, asaltos de modo que abarque los delitos que se pueden llegar a presentar en un ambiente real.
Tamaño	Incluir suficientes muestras de modo que a partir de esto logre generar una buena diversidad y representatividad.
Equilibrio de Clases	Que exista una proporción similar entre las muestras de eventos delictivos y no delictivos.
Etiquetado	Los datos deben estar etiquetados para ser usados en técnicas de aprendizaje supervisado. Para este caso por ser video, debe ser a nivel de fragmento.
Consistencia Temporal	La consistencia temporal es fundamental en este caso debido a que las técnicas están basadas en análisis espacio temporal.
Disponibilidad y Accesibilidad	Se busca que los datos sean abiertos y permitan el uso en la aplicación planteada.

3.1.1. Comparativa conjuntos de datos

A partir de los criterios mencionados previamente se planeta realizar una comparativa con los conjuntos de datos disponibles enfocados en detección de delitos o violencia que puedan

aportar al entrenamiento del modelo planteado.

1. **Representatividad:**
   - **RWF2000:** Contiene una variedad de situaciones de lucha en entornos del mundo real.
   - **LAD2000:** Se enfoca en anomalías en videos, aunque no se centra exclusivamente en violencia.
   - **Real Life Violence Situations:** Representación específica de situaciones violentas en la vida real.
   - **Ubi-Fights:** Centrado en peleas urbanas, brindando un contexto específico.
   - **UCF-Crime:** Se enfoca en actividades criminales, abordando diversas formas de delitos.
2. **Diversidad:**
   - **RWF2000:** Diversidad en escenarios y tipos de luchas.
   - **LAD2000:** Enfocado en anomalías, podría tener diversidad en eventos anómalos. Sin embargo, no todos enfocados en delitos hacia personas.
   - **Real Life Violence Situations:** Diversidad en situaciones de violencia cotidiana.
   - **Ubi-Fights:** Se centra en peleas urbanas, lo que podría limitar la diversidad de contextos.
   - **UCF-Crime:** Diversidad en actividades criminales. No todas las categorías se enfocan en delitos hacia personas.
3. **Tamaño:**
   - **RWF2000:** 2000 clips de video.
   - **LAD2000:** 2000 clips de video.
   - **Real Life Violence Situations:** 2000 clips de video.
   - **Ubi-Fights:** 1000 clips de video.
   - **UCF-Crime:** 1900 clips de video.
4. **Equilibrio de Clases:**
   - **RWF2000:** Presenta un balance adecuado en las clases.
   - **LAD2000:** Debido al numero de clases las muestras por categoría son pocas.
   - **Real Life Violence Situations:** Presenta un balance adecuado en las clases.

- **Ubi-Fights:** Se evidencia un desequilibrio en las categorías con un 21.6 % de eventos anómalos.
- **UCF-Crime:** Debido al número de clases las muestras por categoría son pocas.

5. **Etiquetado:**
   - **RWF2000:** Etiquetado para luchas a nivel de video.
   - **LAD2000:** Etiquetado para anomalías a nivel de cuadro.
   - **Real Life Violence Situations:** Etiquetado para situaciones de violencia en la vida real a nivel de video.
   - **Ubi-Fights:** Etiquetado para peleas urbanas a nivel de cuadro.
   - **UCF-Crime:** Etiquetado para actividades criminales a nivel de video.
6. **Consistencia Temporal:**
   - **RWF2000:** Se basa en anomalías temporales.
   - **LAD2000:** Se basa en anomalías temporales.
   - **Real Life Violence Situations:** Se basa en anomalías temporales.
   - **Ubi-Fights:** Se basa en anomalías temporales.
   - **UCF-Crime:** Se basa en anomalías temporales.
7. **Disponibilidad y Accesibilidad:**
   - **RWF2000:** Accesible.
   - **LAD2000:** Accesible.
   - **Real Life Violence Situations:** Accesible.
   - **Ubi-Fights:** Accesible.
   - **UCF-Crime:** Accesible.

Con base en esta comparativa se logra establecer que a pesar de que se tienen disponibles varios conjuntos de datos enfocados en delitos, se presentan limitaciones en cuanto al tema especifico que se desea clasificar, los delitos cometidos hacia persona en entornos reales. En algunos de los casos no se cumplen con los criterios de representatividad, diversidad y tamaño, debido a la distribución que se genera a causa de las clases. Por otra parte se evidencia dificultades en cuanto al equilibrio de las clases y nivel de etiquetado. Por esta razón, se busca que a partir de los datos disponibles se pueda generar un nuevo conjunto de datos que mejore las limitaciones mencionadas.

3.1.2. Desarrollo conjunto de datos

El desarrollo efectivo del modelo depende en gran medida de la disponibilidad de conjuntos de datos de alta calidad y debidamente etiquetados, ya que esto incide directamente en su rendimiento. Por lo tanto, se plantea la creación de un nuevo conjunto de datos que combina secciones de diversas fuentes de datos centradas en la detección de delitos hacia personas en entornos reales. Este proceso es fundamental, ya que los datos etiquetados son indispensables para el entrenamiento y la validación de modelos de aprendizaje profundo destinados a la detección de delitos en video. Los modelos de aprendizaje profundo demandan una considerable cantidad de datos etiquetados para poder aprender y reconocer patrones en los conjuntos de datos. Además, la elaboración de un conjunto de datos específico para la detección de delitos en videos mediante la selección de datos debidamente etiquetados proporciona un gran aporte para el proceso de entrenamiento y validación.

La creación del conjunto de datos permite a los investigadores obtener una comprensión más profunda de las características visuales inherentes a varios tipos de delitos. La detección de delitos presenta un desafío significativo debido a la variabilidad existente en los tipos de delitos y las características visuales que representan cada uno de ellos. Por consiguiente, el propósito principal de este nuevo conjunto de datos es mejorar la capacidad de los sistemas para detectar delitos en distintos contextos y situaciones, contribuyendo así al avance de la eficacia de los sistemas de detección de delitos.

El conjunto de datos se centra exclusivamente en situaciones relacionadas con delitos hacia personas, excluyendo videos con situaciones enfocadas en daño al patrimonio y segmentos de video de baja calidad o problemas de enfoque. Además, cada muestra se etiqueta cuidadosamente para contribuir al entrenamiento de modelos de aprendizaje supervisado. La metodología utilizada implica la extracción de fragmentos de video de alrededor de 5 segundos que contienen situaciones de delitos, como robos, agresiones, peleas, entre otros. Esta metodología se evidencia en la Figura **3-1**. Las principales fuentes de datos para este desarrollo incluyen extracciones de segmentos de video de conjuntos de datos relacionados como:

- **RWF2000** (Situaciones de peleas en entornos del mundo real)
- **LAD2000** (Conjunto de Datos Benchmark de Anomalías en Video)
- **Real Life Violence Situations** (Situaciones de Violencia en la Vida Real)
- **Ubi-Fights** (Conjunto de Datos de Peleas Urbana)
- **UCF-Crime** (Conjunto de Datos de Actividades Criminales de la Universidad de Florida Central)

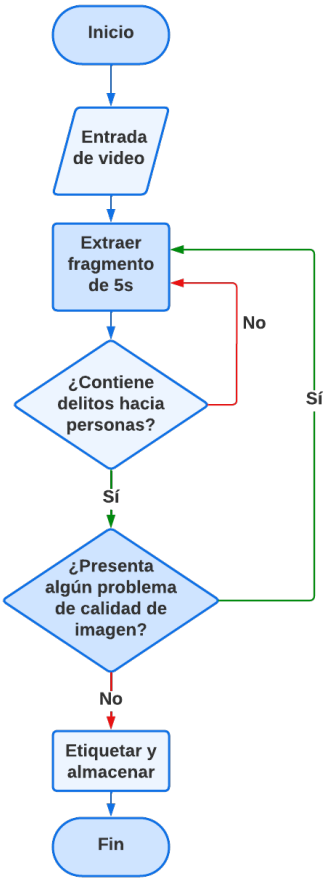


Figura 3-1: Diagrama de flujo conjunto de datos

En la metodología empleada para la creación del conjunto de datos, las muestras tienen una duración aproximada de 5 segundos, y la mayoría de los videos presentan una tasa de fotogramas por segundo (FPS) que oscila entre los 25 y 30 FPS, resultando en muestras de alrededor de 125 fotogramas. Se realiza un minucioso trabajo para asegurar la consistencia temporal, garantizando que cada muestra capture de manera efectiva el evento etiquetado a lo largo de cada uno de los fotogramas.

Conjunto de datos CrimeDetectionDataset

La descripción del conjunto de datos desarrollado se evidencia en la Tabla **3-2**. En conjunto con un análisis de los criterios establecidos para la evaluación de un conjunto de datos en la Tabla **3-3**.

Tabla 3-2: CrimeDetectionDataset

Nombre	Muestras	Categorías	Etiquetado	Delitos	Año
CrimeDetection	9000	2	Cuadro	4874	2024

Tabla 3-3: Criterios para un conjunto de datos

Criterio	Descripción
Representatividad	Contiene muestras de delitos hacia personas, de diversos conjuntos de datos.
Diversidad	Incluye eventos como peleas, robos, hurtos, abuso de modo que abarca una diversidad de eventos que pueden ocurrir en un ambiente real.
Tamaño	Tamaño total de 9000 muestras distribuidas en 2 clases para la detección de eventos delictivos.
Equilibrio de Clases	Compuesto por 4874 eventos positivos y 4126 negativos, con una proporción del 54.16 % .
Etiquetado	Etiquetado a nivel de cuadro o fragmento.
Consistencia Temporal	Se garantiza la consistencia temporal en cada uno de los eventos desde el primer fotograma hasta el último.
Disponibilidad y Accesibilidad	Los datos se hacen públicos en una plataforma de carácter científico (https://zenodo.org/).

3.2. Experimentos

El proceso de entrenamiento del modelo propuesto se inicia con la arquitectura previamente diseñada, que incluye un esquema de capas convolucionales 3D y capas LSTM. Los experimentos se llevan a cabo considerando las siguientes configuraciones:

1. Arquitectura con solo convoluciones 3D
2. Arquitectura con convoluciones 3D y una capa LSTM
3. Arquitectura con convoluciones 3D y dos capa LSTM

3.2.1. Experimentos conjunto de datos RWF2000 y Real Life Violence Situations (RLVS) arquitectura 1

En un primer enfoque, se realizaron pruebas con los conjuntos de datos RWF2000 y Real Life Violence Situations utilizando una arquitectura de red neuronal compuesta principalmente por capas de convolución 3D. Para este caso inicial, el objetivo es analizar la influencia de la

longitud de las muestras en el rendimiento del modelo. Se comenzó tomando 64 fotogramas por cada muestra, estableciendo una frecuencia de muestreo cada 2 fotogramas. Luego, se llevaron a cabo pruebas dividiendo esta muestra en conjuntos de 20 y 10 fotogramas para aumentar el número de instancias, pero disminuyendo la riqueza de la información en cada muestra. Por esta razón, los experimentos se organizaron según la Tabla **3-4**.

Tabla 3-4: Experimentos RWF2000 y RLVS a 256 px de resolución

Experimento	**Conjunto de datos**	**Fotogramas**	**Muestras entrenamiento**	**Muestras validación**	**Muestras evaluación**
1	RWF2000	64	1600	200	200
2	RWF2000	20	4800	600	600
3	RWF2000	10	9600	1200	1200
4	RLVS	64	862	112	110
5	RLVS	20	4319	555	547
6	RLVS	10	8914	1142	1130

Los procesos de entrenamiento se ejecutaron a lo largo de 30 épocas, asignando el 80 % de los datos para el entrenamiento y un 10 % respectivamente para la validación y la evaluación. En términos generales, esta etapa muestra un comportamiento similar al patrón observado en la gráfica de la Figura **3-2**, la cual corresponde al entrenamiento del experimento 5.

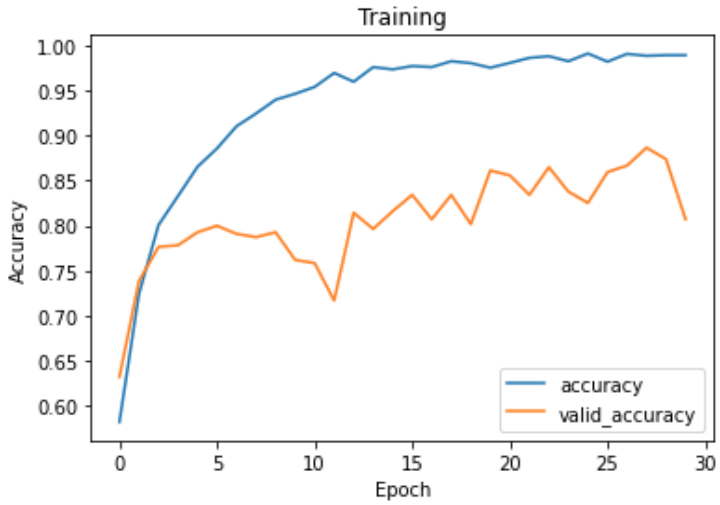
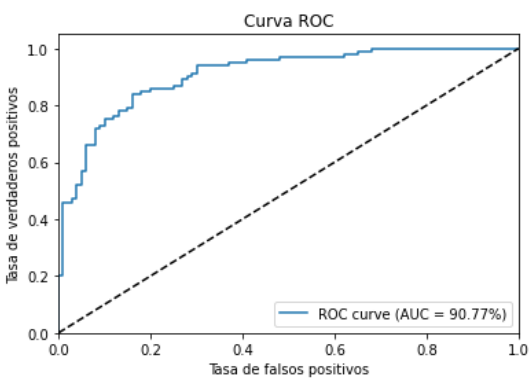
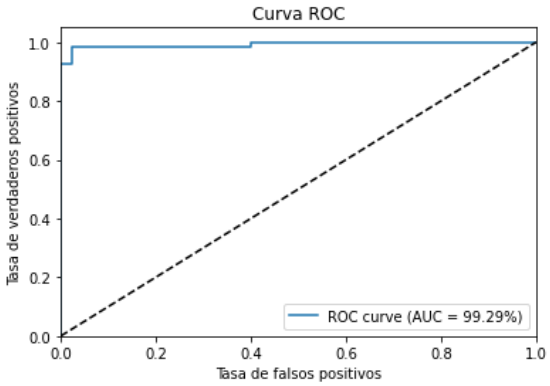


Figura 3-2: Entrenamiento experimentos con RWF2000 y RLVS variando la longitud de la muestra.

Sin embargo, como se puede observar en las figuras (**3-3**, **3-4** y **3-5**), a medida que se reduce

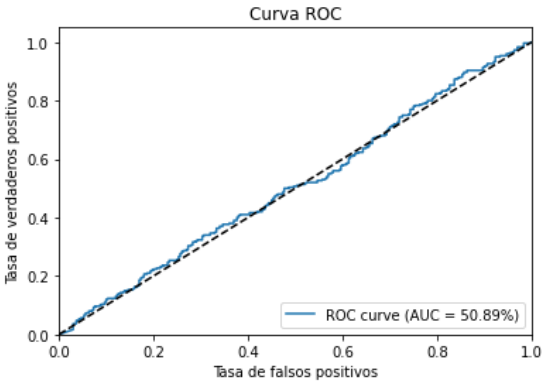


(a) Curva ROC-AUC Experimento 1

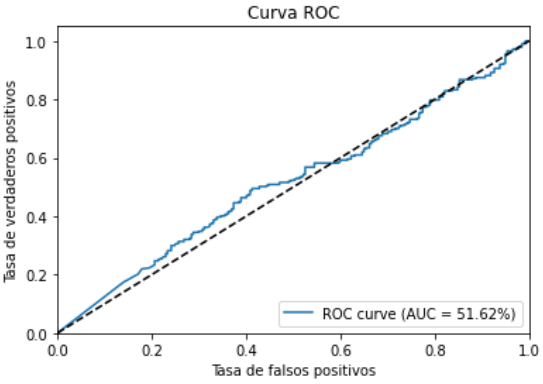


(b) Curva ROC-AUC Experimento 4

Figura 3-3: Curva ROC-AUC a 64 fotogramas.



(a) Curva ROC-AUC Experimento 2

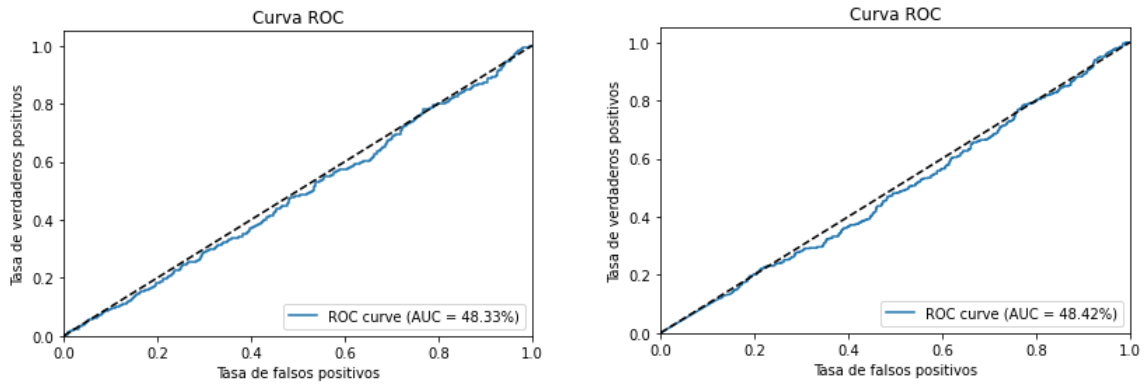


(b) Curva ROC-AUC Experimento 5

Figura 3-4: Curva ROC-AUC a 20 fotogramas.

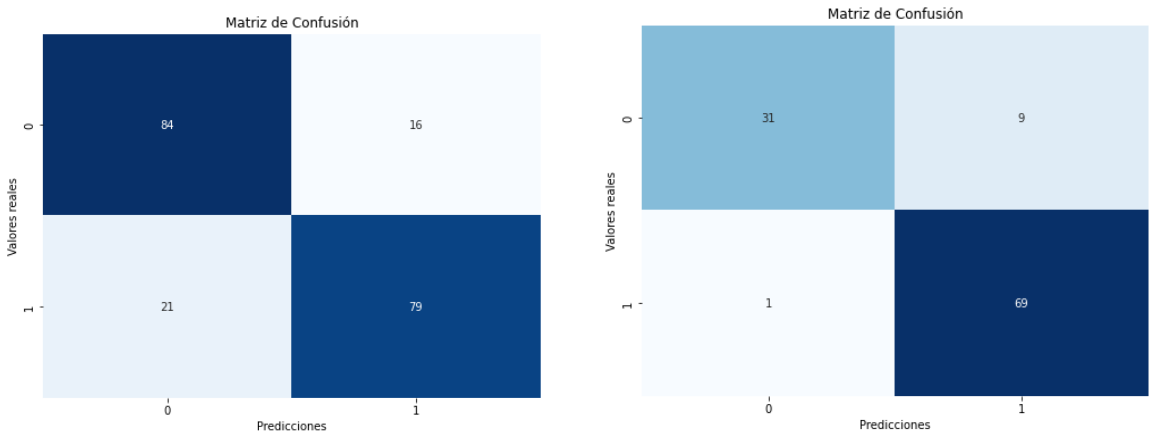
la longitud de la muestra, el modelo experimenta una pérdida en su capacidad para discriminar los eventos. Por esta razón, se ha fijado una cantidad de muestras de 64 fotogramas para los experimentos posteriores.

Para obtener una visión más clara de los resultados obtenidos con 64 fotogramas en los experimentos 1 y 4, que presentan valores de AUC del 90.77 % y 99.29 %, respectivamente, se presenta la matriz de confusión en la Figura **3-6**. En esta figura, se puede apreciar que el modelo tiene la capacidad de clasificar en cierta medida entre un evento delictivo y uno normal. No obstante, es crucial realizar experimentos con conjuntos de datos más extensos para determinar si estos resultados pueden generalizarse en diferentes situaciones de un entorno real.



(a) Curva ROC-AUC Experimento 3 (b) Curva ROC-AUC Experimento 6

Figura 3-5: Curva ROC-AUC a 10 fotogramas.



(a) Matriz de confusión Experimento 1 (b) Matriz de confusión Experimento 4

Figura 3-6: Matriz de confusión entrenamiento a 64 fotogramas.

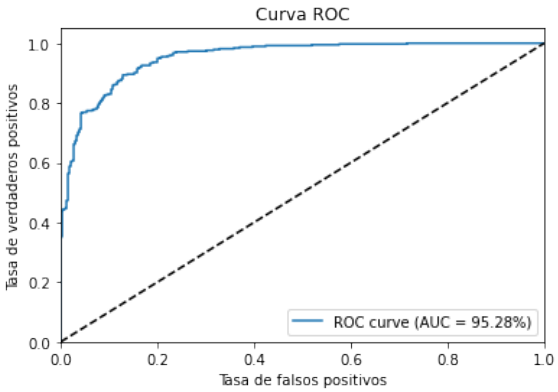
3.2.2. Experimentos conjunto de datos CrimeDetectionDataset arquitectura 1

Los experimentos se focalizaron en analizar el desempeño del modelo con el conjunto de datos CrimeDetectionDataset y evaluar cómo las variaciones en la resolución de los videos impactan en sus resultados. Las alteraciones en la resolución se detallan en la Tabla **3-5**, comenzando con 256x256 píxeles y concluyendo con 64x64 píxeles.

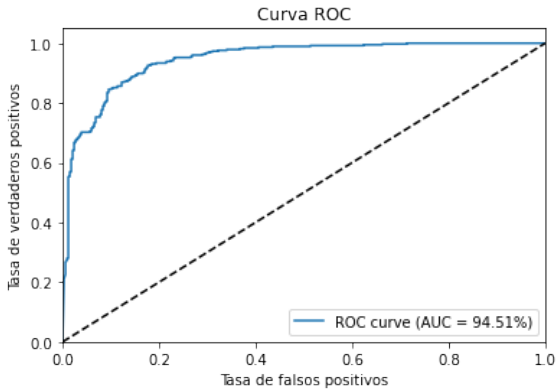
Con base en los resultados de los experimentos 7, 8 y 9, ilustrados en la Figura **3-7**, se observa que la capacidad del modelo para discernir entre eventos positivos y negativos no se ve considerablemente afectada por la reducción en la resolución de los datos de entrada.

Tabla 3-5: Experimentos CrimeDetectionDataset

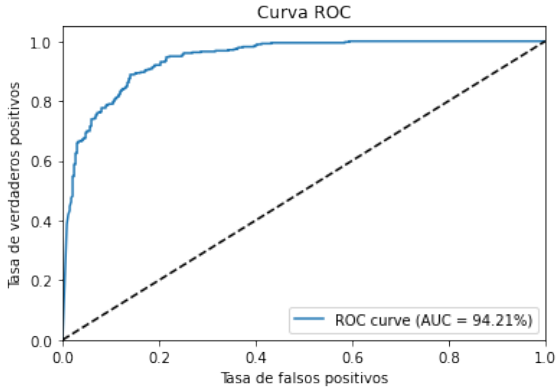
Experimento	Resolución	Muestras entrenamiento	Muestras validación	Muestras evaluación	Precisión
7	256	5452	663	711	87.90 %
8	128	5452	663	711	87.06 %
9	64	5452	663	711	86.78 %



(a) Curva ROC-AUC Experimento 7



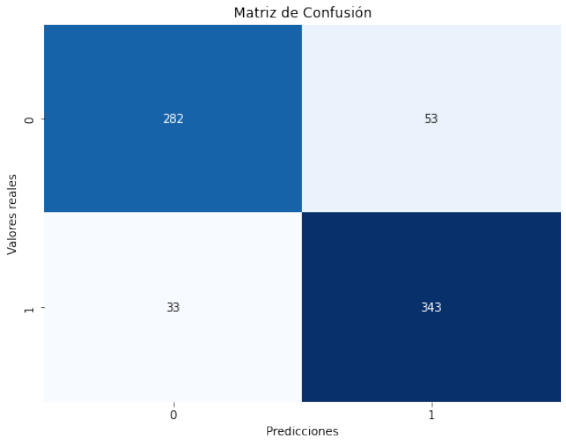
(b) Curva ROC-AUC Experimento 8



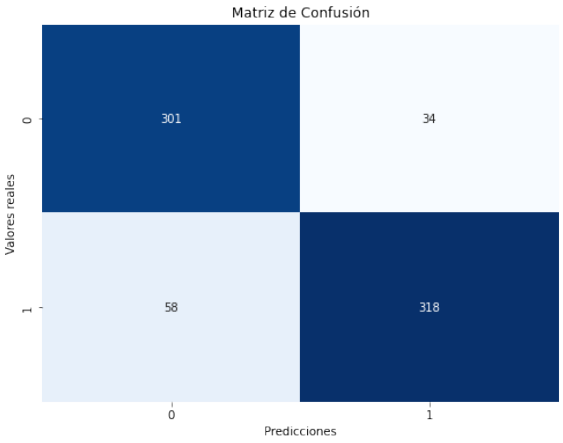
(c) Curva ROC-AUC Experimento 9

Figura 3-7: Curva ROC-AUC experimentos con CrimeDetectionDataset.

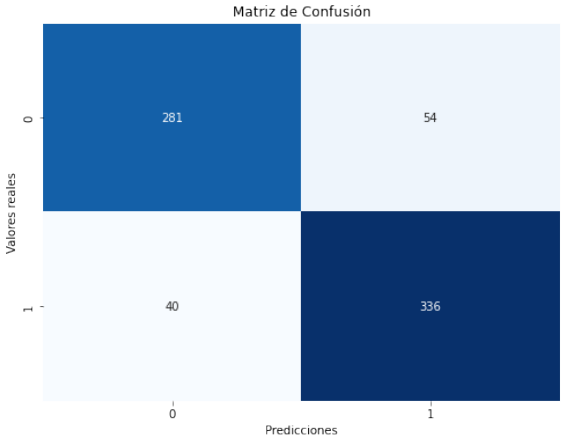
Sin embargo, hay un beneficio significativo en el tiempo de entrenamiento, que disminuye de 8.13 a 2.20 y 1.87 horas respectivamente en cada experimento. Es importante señalar que el tiempo de entrenamiento no es una métrica crítica en este contexto, ya que es una tarea que se realiza solo durante la puesta a punto del modelo. No obstante, es un parámetro relevante al planificar nuevos experimentos, ya que los valores de AUC y Precisión logran mantenerse consistentes.



(a) Matriz de confusión Experimento 7



(b) Matriz de confusión Experimento 8



(c) Matriz de confusión Experimento 9

Figura 3-8: Matriz de confusión experimentos con CrimeDetectionDataset.

Los resultados de estos experimentos presentados en la Figura **3-8** revelan un desempeño satisfactorio del modelo en la tarea de clasificación. Se observa coherencia entre las predicciones del modelo y las clases, reflejando capacidad para distinguir eventos delictivos de situaciones reales. La matriz de confusión, al mostrar una baja cantidad de falsos positivos y falsos negativos, respalda la eficiencia del modelo en la tarea de detección de eventos delictivos. Estos resultados son prometedores para una posterior implementación del modelo en escenarios del mundo real.

3.2.3. Experimentos conjunto de datos CrimeDetectionDataset arquitecturas 1, 2 y 3

En la Tabla **3-6**, se presentan los resultados obtenidos mediante la utilización del conjunto de datos CrimeDetectionDataset con videos de 64x64 píxeles. En esta fase, se lleva a cabo una comparación entre las tres arquitecturas propuestas con el objetivo de determinar cuál presenta un mejor rendimiento en términos de predicción. Se planteó la adición de capas LSTM con la finalidad de mejorar la capacidad del modelo en el análisis de la dimensión temporal, aprovechando su habilidad para capturar patrones y relaciones temporales a largo plazo. Al comparar los resultados de los experimentos 9, 10 y 11, se evidencia que el modelo mantiene su eficacia en la tarea de clasificación propuesta, incluso con la incorporación de las capas LSTM.

Tabla 3-6: Comparación arquitecturas usando CrimeDetectionDataset

Experimento	**Arquitectura**	**Precisión**	**AUC-ROC**	**AUC- Precision, Recall**
9	1 (Conv 3D)	86.78 %	94.21 %	94.44 %
10	2 (Conv 3D + 1LSTM)	83.68 %	92.94 %	93.44 %
11	3 (Conv 3D + 2LSTM)	87.06 %	93.89 %	94.17 %

3.2.4. Especificaciones técnicas hardware

La elección y configuración del hardware para los experimentos desempeña un papel muy importante en la ejecución y el rendimiento de los modelos enfocados en análisis de archivos de vídeos. Se optó por un sistema con especificaciones técnicas robustas que garantizan la capacidad de procesamiento y almacenamiento necesaria para manejar grandes volúmenes de datos de vídeo. Con una selección adecuada del hardware se busca lograr una ejecución eficiente de los algoritmos y modelos de aprendizaje máquina. Además, es necesario incorporar una unidad de procesamiento gráfico (GPU) de alto rendimiento para acelerar el procesamiento paralelo y la inferencia de los modelos. Las especificaciones técnicas del equipo donde fueron ejecutados los experimentos se muestran en la Tabla **3-7**.

Tabla 3-7: Especificaciones técnicas hardware

Dispositivo	**Características**
CPU	Intel Core i7 9700k
RAM	32GB - 3200MHz
GPU	3060TI 8GB

Durante la implementación y evaluación de los modelos, se logra alcanzar un tiempo de predicción promedio de alrededor de 90 milisegundos, lo cual indica una capacidad significativa para su implementación en entornos del mundo real. Este resultado es especialmente relevante para aplicaciones de seguridad ciudadana y videovigilancia, donde la rapidez de detección y respuesta a eventos anómalos es fundamental. Este indicador clave confirma la viabilidad de estos sistemas para su despliegue en entornos operativos, con el propósito de contribuir a la mejora de la seguridad ciudadana y el bienestar de las comunidades.

4 Conclusiones y recomendaciones

4.1. Conclusiones

- Durante el desarrollo de esta investigación, se exploraron diversos enfoques y modelos para la detección de delitos en archivos de video mediante técnicas de aprendizaje de máquina. En comparación con otros enfoques, los modelos basados en redes neuronales convolucionales 3D y modelos LSTM demostraron una notable eficacia en la identificación de patrones temporales y la detección de anomalías en secuencias de video. La capacidad de estas arquitecturas para capturar relaciones temporales a largo plazo y comprender el contexto dinámico de las escenas de sistemas de videovigilancia las posiciona como herramientas indispensables para mejorar la seguridad ciudadana y contribuir en la tarea de detección de delitos. Sin embargo, se destaca la importancia de continuar investigando y refinando estos enfoques, así como de explorar nuevas técnicas que permitan mejorar la precisión y la eficiencia de los modelos.

- En el contexto de las técnicas de aprendizaje de máquina enfocadas en detección de delitos en video, la gestión de conjuntos de datos extensos es un desafío determinante en el proceso de desarrollo de los modelos. La optimización de recursos computacionales es una etapa esencial para reducir la complejidad y la duración de las fases de entrenamiento e inferencia de los modelos. Este proceso es fundamental para facilitar una implementación eficaz de este tipo de sistemas en entornos reales de videovigilancia, donde la capacidad de respuesta y la eficiencia son características criticas para la integración y el apoyo hacia las entidades encargadas de dar soporte a los eventos delictivos de cada región.

- En este trabajo de tesis se identificó que la selección adecuada de métricas para evaluar el desempeño de un modelo es un aspecto primordial en la investigación y desarrollo de soluciones en el campo de la detección de anomalías en video. Las métricas de AUC-ROC y AUC-PR son indispensables en este contexto, pues ofrecen una evaluación clara de la capacidad del modelo para discriminar entre clases, incluso cuando no existe equilibrio de clases en el conjunto de datos. Se debe resaltar que la implementación de métricas como la "Precisión"puede resultar engañosa en este tipo de problemas, ya que podría reflejar altos valores simplemente debido a la distribución desigual de

las clases, sin necesariamente reflejar la efectividad del modelo en la detección de anomalías. Por el contrario, las métricas de AUC-ROC y AUC-PR ofrecen una visión más integral y precisa del rendimiento del modelo pues consideran la representatividad del conjunto de datos utilizado en el proceso de entrenamiento y evaluación del modelo, lo que resulta fundamental para el desarrollo y la validación efectiva de soluciones en la detección de anomalías en videos. Lo anterior, promueve la transparencia y la fiabilidad de los resultados, lo que a su vez fortalece la confianza en las soluciones propuestas y su capacidad para abordar los desafíos en aplicaciones de seguridad ciudadana en un entorno real.

- Una etapa esencial para el desarrollo de un modelo que logre generalizar adecuadamente en el proceso de detección de delitos en video recae en la importancia crítica de evaluar de manera adecuada los conjuntos de datos destinados al entrenamiento y la preparación del modelo. Factores como el balance de clases, la representatividad, el tamaño, el etiquetado y la consistencia temporal de los datos ejercen una influencia directa en la respuesta del modelo. La selección cuidadosa y la evaluación exhaustiva de los conjuntos de datos son fundamentales para garantizar que los modelos resultantes sean robustos, precisos y generalizables. La calidad de los datos utilizados en el proceso de entrenamiento no solo afecta la capacidad de predicción del modelo, sino que también puede influir en su desempeño en situaciones del mundo real. Por lo tanto, es crucial invertir tiempo y esfuerzo en la construcción y validación de conjuntos de datos de alta calidad para lograr resultados confiables y efectivos en la detección de anomalías en diversos contextos y aplicaciones.
- Durante la última etapa del proyecto, se identificó la necesidad de contar con contenedores y sistemas que faciliten el despliegue de la aplicación en diferentes entornos. Estos sistemas desempeñan un papel crucial en el proceso de puesta a punto en entornos reales, pues garantizan la portabilidad y la consistencia del entorno de ejecución. Asimismo, permiten una gestión eficiente de los recursos y facilitan la escalabilidad de la aplicación, aspectos fundamentales para su despliegue y operación en escenarios prácticos de seguridad y videovigilancia.

4.2. Recomendaciones

Estas recomendaciones están dirigidas a motivar a la comunidad académica a continuar explorando nuevas técnicas y enfoques en el campo de la detección de delitos y anomalías en video. El objetivo es contribuir significativamente a la mejora de la seguridad ciudadana y al bienestar de las comunidades. El avance en este campo requiere un compromiso continuo con la investigación y el desarrollo de soluciones que aborden los desafíos emergentes en la prevención y detección del crimen. Algunos aspectos que se pueden tener en cuenta para contribuir al tema son los siguientes:

- Es esencial impulsar el desarrollo y la diversificación de los conjuntos de datos. Explorar y crear conjuntos de datos más diversos y representativos que abarquen una amplia gama de escenarios y condiciones cruciales para mejorar la robustez y la capacidad de generalización de los modelos. Además, se debe prestar especial atención al etiquetado de los datos para garantizar su compatibilidad con los diversos enfoques de aprendizaje supervisado que se pueden aplicar en este campo.

- La gestión eficiente de los recursos computacionales es indispensable para el despliegue efectivo de sistemas de detección de delitos en vídeos. En este sentido, resulta fundamental explorar y aplicar técnicas de optimización en los modelos desarrollados. La reducción significativa de los tiempos de inferencia constituye un objetivo primordial, ya que permite ofrecer un apoyo inmediato y preciso a las autoridades encargadas del orden público. En un contexto donde la videovigilancia y la respuesta en tiempo real son fundamentales, la agilidad y la eficacia de los sistemas de detección se convierten en aspectos críticos para garantizar la seguridad y la tranquilidad de la comunidad. Es por ello que la investigación continua en métodos de optimización de modelos es una estrategia clave para mejorar la capacidad de respuesta de estas herramientas tecnológicas y su utilidad en escenarios de aplicación del mundo real.

- Enriquecer la investigación mediante procesos de validación experimental en entornos reales de videovigilancia es una etapa que podría proporcionar una perspectiva sobre la efectividad y la escalabilidad de los modelos desarrollados en situaciones cotidianas, reflejando con mayor fidelidad los desafíos y condiciones a los que se enfrentarán en la implementación práctica. Al someter los modelos a entornos de videovigilancia reales, se podría obtener una retroalimentación invaluable sobre el desempeño en escenarios dinámicos y variados, donde pueden surgir situaciones imprevistas que impacten en el proceso de detección de delitos.

- La validación experimental en entornos reales también permite identificar posibles limitaciones y áreas de mejora en los modelos, destacando aspectos que pueden requerir ajustes o refinamientos para optimizar su eficacia y precisión. Asimismo, brinda la oportunidad de evaluar la integración de los modelos con los sistemas de videovigilancia existentes, asegurando una implementación fluida y efectiva en aplicaciones de seguridad ciudadana. Este proceso de validación aporta una perspectiva que podría fortalecer la confianza en la capacidad de los modelos para abordar los desafíos del mundo real en la detección de delitos.

- Establecer colaboraciones interdisciplinarias que reúnan a expertos en seguridad ciudadana y profesionales encargados de formular políticas públicas. Esta sinergia podría contribuir a abordar de manera integral los desafíos relacionados con la detección de delitos y la promoción de entornos urbanos seguros. La participación de especialistas en seguridad ciudadana aporta un profundo entendimiento de las dinámicas delictivas y de las necesidades específicas de las comunidades, lo que permite adaptar los modelos y las estrategias de detección a contextos específicos. La colaboración interdisciplinaria promueve un enfoque orientado a resultados, donde la combinación de conocimientos técnicos y experiencia práctica buscan maximizar el impacto positivo en la sociedad.

- El intercambio de conocimientos entre instituciones académicas, organismos gubernamentales y organizaciones encargadas del orden público son fundamentales para impulsar el desarrollo de soluciones integrales y sostenibles en materia de seguridad ciudadana. Al trabajar en conjunto, se pueden generar avances significativos que contribuyan a construir entornos urbanos más seguros para todos los ciudadanos.

Bibliografía

[Acsintoae et al., 2021] Acsintoae, A., Florescu, A., Georgescu, M.-I., Mare, T., Sumedrea, P., Ionescu, R. T., Khan, F. S., and Shah, M. (2021). Ubnormal: New benchmark for supervised open-set video anomaly detection. *Computer Vision and Pattern Recognition*.

[Adam et al., 2008] Adam, A., Rivlin, E., Shimshoni, I., and Reinitz, D. (2008). Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):555–560.

[Althnian et al., 2021] Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A. B., Alzakari, N., Abou Elwafa, A., and Kurdi, H. (2021). Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Applied Sciences*, 11(2).

[Anthopoulos, 2015] Anthopoulos, L. G. (2015). *Understanding the Smart City Domain: A Literature Review*, pages 9–21. Springer International Publishing, Cham.

[Boekhoudt et al., 2021] Boekhoudt, K., Matei, A., Aghaei, M., and Talavera, E. (2021). Hr-crime: Human-related anomaly detection in surveillance videos. *CoRR*, abs/2108.00246.

[Carreira and Zisserman, 2017] Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. pages 4724–4733.

[Catlett et al., 2019] Catlett, C., Cesario, E., Talia, D., and Vinci, A. (2019). Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments. *Pervasive and Mobile Computing*, 53.

[Cheng et al., 2021] Cheng, M., Cai, K., and Li, M. (2021). Rwf-2000: An open large scale video database for violence detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4183–4190.

[Degardin and Proença, 2021] Degardin, B. and Proença, H. (2021). Iterative weak/self-supervised classification framework for abnormal events detection. *Pattern Recognition Letters*, 145:50–57.

[Dubey et al., 2019] Dubey, S., Boragule, A., and Jeon, M. (2019). 3d resnet with ranking loss function for abnormal activity detection in videos. In *2019 International Conference on Control, Automation and Information Sciences (ICCAIS)*, pages 1–6.

[Farnebäck, 2003] Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. volume 2749, pages 363–370.

[Feng et al., 2021] Feng, J.-C., Hong, F.-T., and Zheng, W.-S. (2021). Mist: Multiple instance self-training framework for video anomaly detection. pages 14004–14013.

[Gemmeke et al., 2017] Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

[Gorr et al., 2003] Gorr, W., Olligschlaeger, A., and Thompson, Y. (2003). Short-term forecasting of crime. *International Journal of Forecasting*, 19.

[Hasan et al., 2016] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., and Davis, L. S. (2016). Learning temporal regularity in video sequences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–742.

[Hershey et al., 2016] Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. W. (2016). CNN architectures for large-scale audio classification. *CoRR*, abs/1609.09430.

[Isafiade and Bagula, 2020] Isafiade, O. E. and Bagula, A. B. (2020). Series mining for public safety advancement in emerging smart cities. *Future Generation Computer Systems*, 108.

[Kamoona et al., 2023] Kamoona, A. M., Gostar, A. K., Bab-Hadiashar, A., and Hoseinnezhad, R. (2023). Multiple instance-based video anomaly detection using deep temporal encoding–decoding. *Expert Systems with Applications*, 214:119079.

[Karpathy et al., 2014] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *CVPR*.

[Kliper-Gross et al., 2012] Kliper-Gross, O., Hassner, T., and Wolf, L. (2012). The action similarity labeling challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:615–621.

[Landi et al., 2019] Landi, F., Snoek, C. G. M., and Cucchiara, R. (2019). Anomaly locality in video surveillance. *ArXiv*, abs/1901.10364.

[Lu et al., 2013] Lu, C., Shi, J., and Jia, J. (2013). Abnormal event detection at 150 fps in matlab. In *2013 IEEE International Conference on Computer Vision*, pages 2720–2727.

[Luo et al., 2017] Luo, W., Liu, W., and Gao, S. (2017). A revisit of sparse coding based anomaly detection in stacked rnn framework. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 341–349.

[Lv et al., 2021a] Lv, H., Chen, C., Cui, Z., Xu, C., Li, Y., and Yang, J. (2021a). Learning normal dynamics in videos with meta prototype network. *Computer Vision and Pattern Recognition*.

[Lv et al., 2021b] Lv, H., Zhou, C., Cui, Z., Xu, C., Li, Y., and Yang, J. (2021b). Localizing anomalies from weakly-labeled videos. *Computer Vision and Pattern Recognition*.

[Mahadevan et al., 2010] Mahadevan, V., Li, W., Bhalodia, V., and Vasconcelos, N. (2010). Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981.

[Majhi et al., 2021] Majhi, S., Das, S., Bremond, F., Dash, R., and Sa, P. (2021). Weakly-supervised joint anomaly detection and classification. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–7, Los Alamitos, CA, USA. IEEE Computer Society.

[Maqsood et al., 2021] Maqsood, R., Bajwa, U., Saleem, G., Raza, R., and Anwar, M. (2021). Anomaly recognition from surveillance videos using 3d convolutional neural networks.

[Medapati et al., 2019] Medapati, P. K., Murthy, P. H. S. T., and Sridhar, K. P. (2019). Lamstar: For iot-based face recognition system to manage the safety factor in smart cities. *Transactions on Emerging Telecommunications Technologies*.

[Mehran et al., 2009] Mehran, R., Oyama, A., and Shah, M. (2009). Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942.

[OECD, 2021] OECD (2021). OECD, "Better life index, security". Accedido: 2022-03-17.

[Oh, 2011] Oh, S. (2011). A new dataset evaluation method based on category overlap. *Computers in Biology and Medicine*, 41(2):115–122.

[Pang et al., 2019] Pang, Y., Zhang, L., Ding, H., Fang, Y., and Chen, S. (2019). Spath: Finding the safest walking path in smart cities. *IEEE Transactions on Vehicular Technology*, 68.

[Perez et al., 2019] Perez, M., Kot, A. C., and Rocha, A. (2019). Detection of real-world fights in surveillance videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2662–2666.

[Ramzan et al., 2019] Ramzan, M., Abid, A., Khan, H. U., Awan, S. M., Ismail, A., Ahmed, M., Ilyas, M., and Mahmood, A. (2019). A review on state-of-the-art violence detection techniques. *IEEE Access*, 7:107560–107575.

[Rathore et al., 2016] Rathore, M., Ahmad, A., Paul, A., and Rho, S. (2016). Urban planning and building smart cities based on the internet of things using big data analytics. *Computer Networks*, 101:63–80.

[Rathore et al., 2018] Rathore, M., Paul, A., Ahmad, A., Chilamkurti, N., Hong, W.-H., and Seo, H. (2018). Real-time secure communication for smart city in high-speed big data environment. *Future Generation Computer Systems*, 83:638–652.

[Sandler et al., 2018] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

[SIEDCO, 2021] SIEDCO (2021). Estadística delictiva. Accedido: 2022-03-23.

[Simić et al., 2020] Simić, M., Perić, M., Popadić, I., Perić, D., Pavlović, M., Vučetić, M., and Stanković, M. (2020). Big data and development of smart city: System architecture and practical public safety example. *Serbian Journal of Electrical Engineering*, 17:337–355.

[Soliman et al., 2019] Soliman, M. M., Kamal, M. H., El-Massih Nashed, M. A., Mostafa, Y. M., Chawky, B. S., and Khattab, D. (2019). Violence recognition from videos using deep learning techniques. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 80–85.

[Soomro et al., 2012] Soomro, K., Zamir, A. R., and Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402.

[Sultani et al., 2019] Sultani, W., Chen, C., and Shah, M. (2019). Real-world anomaly detection in surveillance videos. *Computer Vision and Pattern Recognition*.

[Tian et al., 2021] Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J. W., and Carneiro, G. (2021). Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. *Computer Vision and Pattern Recognition*.

[Tran et al., 2015] Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. pages 4489–4497.

[Ullah and Petrosino, 2017] Ullah, I. and Petrosino, A. (2017). A spatiotemporal feature learning approach for dynamic scene recognition.

[Ullah et al., 2021a] Ullah, W., Ullah, A., Haq, I. U., Muhammad, K., Sajjad, M., and Baik, S. W. (2021a). Cnn features with bi-directional lstm for real-time anomaly detection in surveillance networks. *Multimedia Tools and Applications*.

[Ullah et al., 2021b] Ullah, W., Ullah, A., Hussain, T., Khan, A., and Baik, S. W. (2021b). An efficient anomaly recognition framework using an attention residual lstm in surveillance videos. *Sensors*.

[Vahdani and Tian, 2023] Vahdani, E. and Tian, Y. (2023). Deep learning-based action detection in untrimmed videos: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4302–4320.

[Wan et al., 2021] Wan, B., Jiang, W., Fang, Y., Luo, Z., and Ding, G. (2021). Anomaly detection in video sequences: A benchmark and computational model. *CoRR*, abs/2106.08570.

[Wu et al., 2021] Wu, J., Zhang, W., Li, G., Wu, W., Tan, X., Li, Y., Ding, E., and Lin, L. (2021). Weakly-supervised spatio-temporal anomaly detection in surveillance video.

[Wu et al., 2020] Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., and Yang, Z. (2020). Not only look, but also listen: Learning multimodal violence detection under weak supervision. *Computer Vision and Pattern Recognition*.

[Xu et al., 2022] Xu, Y., Huang, C., Nan, Y., and Lian, S. (2022). Tad: A large-scale benchmark for traffic accidents detection from video surveillance.

[Zhang and Yu, 2018] Zhang, S. and Yu, H. (2018). Person re-identification by multi-camera networks for internet of things in smart cities. *IEEE Access*, 6.

[Zhong et al., 2019] Zhong, J.-X., Li, N., Kong, W., Liu, S., Li, T. H., and Li, G. (2019). Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1237–1246.

[Zhu and Yang, 2018] Zhu, C. and Yang, Y. (2018). Face detection and recognition based on deep learning in the monitoring environment. *Communications in Computer and Information Science*, pages 698–705.