Instituto Superior Tecnológico "Los Andes" ISTLA

CÓDIGO CIENTÍFICO REVISTA DE INVESTIGACIÓN

Vol. 6 – Núm. 1 / Enero – Junio – 2025

Desarrollo de un Modelo de Machine Learning para el Reconocimiento de Comportamientos Inusuales de Personas en Videos de Videovigilancia Comunitaria

Development of a Machine Learning Model for the Recognition of Unusual Behavior of People in Community Surveillance Videos

Desenvolvimento de um modelo de aprendizagem automática para o reconhecimento de comportamentos invulgares de pessoas em vídeos de videovigilância comunitaria

Muñoz Abad, Edgar Osvaldo
Escuela Politécnica del Chimborazo
osvaldo.munioz@espoch.edu.ec
https://orcid.org/0000-0003-4473-3620

Paguay Soxo, Paul Xavier
Escuela Politécnica del Chimborazo
paul.paguay@espoch.edu.ec
http://orcid.org/0000-0002-0262-9844

DOI / URL: https://doi.org/10.55813/gaea/ccri/v6/n1/912

Como citar:
Muñoz Abad, E. O., & Paguay Soxo, P. X. (2025). Desarrollo de un Modelo de Machine Learning para el Reconocimiento de Comportamientos Inusuales de Personas en Videos de Videovigilancia Comunitaria. *Código Científico Revista De Investigación*, *6*(1), 690–709. https://doi.org/10.55813/gaea/ccri/v6/n1/912

Recibido: *29/05/2025* ***Aceptado:*** *27/06/2025* ***Publicado:*** *30/06/2025*

Resumen
La videovigilancia comunitaria se está convirtiendo en una herramienta fundamental para fortalecer la seguridad en entornos urbanos. Sin embargo, el incremento exponencial en la cantidad de cámaras no ha sido acompañado por una mejora proporcional en la capacidad de monitoreo humano, lo que limita la detección oportuna de eventos anómalos. Este estudio presenta un modelo de aprendizaje profundo diseñado para el reconocimiento automático de comportamientos inusuales en videos de vigilancia. La arquitectura propuesta combina EfficientNet como extractor de características espaciales con una red ConvLSTM2D para modelar la dimensión temporal de los eventos. El conjunto de datos fue conformado por secuencias de imágenes etiquetadas correspondientes a diversas clases de eventos anómalos y normales. Los resultados experimentales demuestran que el modelo alcanza una puntuación AUC global de 0.8795, con valores superiores al 0.90 en categorías como “Fighting” y “Robbery”. La metodología propuesta muestra una alta capacidad de discriminación y generalización, validando su aplicabilidad en sistemas de videovigilancia inteligente. Los resultados obtenidos en este estudio demuestran que el uso de arquitecturas hibridas pueden mejorar la detección de comportamiento anómalos en videos de video vigilancia en tiempo no real.
Palabras clave: Videovigilancia, Detección de actividades anómalas, redes profundas, EfficientNet, ConvLSTM, visión por computadora.

Abstract
Community video surveillance is becoming a fundamental tool for strengthening security in urban environments. However, the exponential increase in the number of cameras has not been accompanied by a proportional improvement in human monitoring capacity, which limits the timely detection of anomalous events. This study presents a deep learning model designed for automatic recognition of unusual behavior in surveillance videos. The proposed architecture combines EfficientNet as a spatial feature extractor with a ConvLSTM2D network to model the temporal dimension of events. The dataset was comprised of labeled image sequences corresponding to various classes of anomalous and normal events. Experimental results show that the model achieves an overall AUC score of 0.8795, with values above 0.90 in categories such as “Fighting” and “Robbery”. The proposed methodology shows a high discrimination and generalization capacity, validating its applicability in intelligent video surveillance systems. The results obtained in this study demonstrate that the use of hybrid architectures can improve the detection of anomalous behavior in non-real-time video surveillance videos.
Keywords: Video surveillance, Anomalous activity detection, deep networks, EfficientNet, ConvLSTM, computer vision.

Resumo
A videovigilância comunitária está a tornar-se uma ferramenta essencial para reforçar a segurança em ambientes urbanos. No entanto, o aumento exponencial do número de câmaras não tem sido acompanhado por uma melhoria proporcional das capacidades de monitorização humana, limitando a deteção atempada de eventos anómalos. Este estudo apresenta um modelo de aprendizagem profunda concebido para o reconhecimento automático de comportamentos invulgares em vídeos de vigilância. A arquitetura proposta combina a EfficientNet como um extrator de caraterísticas espaciais com uma rede ConvLSTM2D para modelar a dimensão temporal dos eventos. O conjunto de dados consiste em sequências de imagens rotuladas correspondentes a várias classes de eventos anómalos e normais. Os resultados experimentais mostram que o modelo atinge uma pontuação AUC global de 0,8795, com valores acima de 0,90 em categorias como “Luta” e “Roubo”. A metodologia proposta apresenta uma elevada capacidade de discriminação e generalização, validando a sua aplicabilidade em sistemas de

videovigilância inteligentes. Os resultados obtidos neste estudo demonstram que a utilização de arquitecturas híbridas pode melhorar a deteção de comportamentos anómalos em vídeos de videovigilância em tempo não real.
Palavras-chave: Videovigilância, deteção de atividade anómala, redes profundas, EfficientNet, ConvLSTM, visão computacional.

Introducción

Los sistemas comunitarios de videovigilancia se han consolidado como un recurso útil y crucial para garantizar la seguridad en lugares públicos como ciudadelas, calles, entornos bancarios, centros comerciales, etc. La implementación de estas tecnologías permite monitorear zonas críticas y la detección de actividades anormales, tales como; crímenes, accidentes o comportamientos inusuales (Zahra et al., 2024). Sin embargo, el crecimiento en la cantidad de cámaras instaladas no ha coincidido con el aumento en la capacidad de humana para monitorearlas, generando grandes desafíos operativos para las entidades encargadas de la seguridad (Khan et al., 2020). Un aspecto que complica a un más la tarea de realizar un monitoreo manual es que los eventos anormales ocurren con menor frecuencia que las actividades normales, esto hace que la detección de estos eventos anómalos en grandes volúmenes de videos requiera un mayor esfuerzo humano (Myagmar-Ochir & Kim, 2023). Esta situación nos lleva a buscar soluciones más eficientes. Teniendo en cuenta los avances que se han tenido en el área de visión por computadora y aprendizaje profundo nos ha abierto la posibilidad de facilitar la detección automática de eventos anómalos en grandes volúmenes de video (Pham et al., 2022).

En estudios recientes hemos observado que los modelos de aprendizaje profundo tienen la capacidad de aprender e incrementar la precisión en las detecciones, al mismo tiempo que reducen los costos computacionales (Myagmar-Ochir & Kim, 2023). Adicionalmente, enfoques como la segmentación por áreas, han permitido mejorar la eficiencia en el procesamiento de videos, lo que facilita la incorporación de estos modelos en sistemas de videovigilancia comunitaria (Zahra et al., 2024). Ante esta realidad, creemos que resulta

urgente desarrollar algoritmos que permitan realizar la automatización en la detección de anomalías en videos, minimizando la necesidad de intervención humana, y así mejorar la eficiencia de las operaciones en los sistemas de seguridad comunitarias.

La visión por computadora es una tecnología en donde nos podemos apoyar para extraer la información relevante de imágenes y videos, y así utilizar estas características para el entrenamiento de redes neuronales profundas, como las Redes Neuronales Convolucionales (*CNN*) y las Redes Neuronales Recurrentes (*RNN*), que permiten analizar tanto el contenido visual como su evolución en el tiempo (Mohanapriya et al., 2024). Estudios recientes también destacan el valor de utilizar redes de tres dimensiones (3DCNN) en la detección de comportamientos anómalos vinculados con robos (Martínez-Mascorro et al., 2020), incluso en escenarios más complejos, donde existe una alta concurrencia de personas (Revathi & Kumar, 2017).

En el caso de Ecuador, muchas comunidades y barrios ya han instalado sistemas de videovigilancia, si bien proveen de visualización en tiempo real, carecen de capacidades avanzadas para detectar y clasificar actos vandálicos, robos y otras actividades delictivas de manera automática. Por esta razón, nuestra investigación busca desarrollar un modelo capaz de procesar estos videos y clasificar de manera automáticamente eventos anómalos, aportando así una solución tecnológica concreta para mejorar la seguridad comunitaria.

Con la finalidad de abordar la problemática que se ha expuesto, nuestra investigación se centra en la siguiente pregunta: ¿Cómo puede la visión por computadora y el aprendizaje automático ayudar a identificar y clasificar de manera eficiente eventos anómalos en videos de sistemas de video vigilancia comunitaria?

Para dar respuesta a esta interrogante, y considerando tanto el contexto del problema como los antecedentes revisados, se han definido los siguientes objetivos específicos:

1) Procesar los videos de videovigilancia seleccionados para entrenar el modelo, incluyendo ejemplos de comportamientos normales y también anómalos.
2) Extraer las características específicas de los datos, para mejorar su uso en el modelo de aprendizaje automático.
3) Entrenar el modelo de detección y clasificación mediante métodos de aprendizaje profundo.
4) Evaluar el desempeño del modelo utilizando indicadores como AUC (*Área Bajo la Curva ROC*), con el fin de asegurar su eficiencia en la detección de comportamientos anómalos.

Metodología

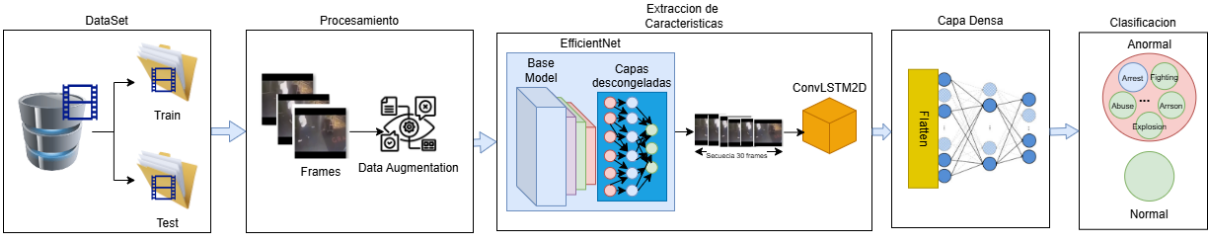
En la Figura 1 se presenta la arquitectura propuesta en el marco de esta investigación. Dentro de la arquitectura planteada, se emplea una combinación temporal y espacial para la extracción de características. Como extractor de características espaciales utilizamos el modelo base EfficientNet (Koonce, 2021) que es una arquitectura de red neuronal convolucional eficiente, diseñada para extraer características espaciales complejos con bajo costo computacional, en comparación con redes convolucionales tradicionales. Este modelo introduce un enfoque de escalamiento compuesto que optimiza simultáneamente la profundidad, el ancho y la resolución de entrada, lo cual permite mejorar la precisión sin incurrir en un incremento excesivo en los recursos requeridos (Tan & Le, 2019).

Por su parte, como modelador temporal se utiliza ConvLSTM2D que es una variante de las redes LSTM (Long Short-Term Memory) (Prakash et al., 2023) que incorpora convoluciones, permitiendo modelar la información espacial y temporal simultáneamente, lo cual es ideal para procesar secuencias de imágenes como videos. A diferencia de las LSTM convencionales, que tratan los datos como vectores unidimensionales, ConvLSTM2D conserva

la estructura espacial de las imágenes al integrar operaciones de convolución dentro de sus celdas de memoria (Shi et al., 2015).

Figura 1

Arquitectura del modelo propuesto



Nota: Autores (2025).

Conjunto de Datos

El conjunto de datos que fue utilizado para este estudio es el UFC Crime (Center for Research in Computer Vision, n.d.) diseñado específicamente para evaluar métodos de detección de anomalías en videos de vigilancia. Este conjunto de datos consta de 1900 videos de 13 tipos de evento anómalos (véase **¡Error! No se encuentra el origen de la referencia.**), tales como “Abuso”, “Ataque”, “Explosiones”, entre otras. Adicional, se incluye una categoría de Eventos Normales, que abarca videos sin incidentes anómalos, tanto en interiores como exteriores, y que varían en escenas diurnas y nocturnas.

Tabla 1

Descripción de las categorías del dataset UCF Crime

Categoría	Definición
Abuso	Este evento contiene videos que muestran comportamientos malos, cruel o violento contra niños, ancianos, a animales y mujeres
Arrestos	Policías arrestando personas
Incendios provocados	Personas prendiendo fuego deliberadamente a la propiedad
Ataques	Ataque físico o repentino violento contra alguien
Robo a Casas	Personas entrando a casas con intención de robar
Explosiones	Evento destructivo de algo que explota
Peleas	Dos o más personas atacándose

Normal	Videos sin delitos en interior y exterior
Accidente de vehículos	Accidentes de transito
Robo	Ladrones tomando dinero a la fuerza
Tiroteo	Disparos a alguien con arma
Robo en Tiendas	Personas robando en tiendas
Robo de objetos	Personas tomando objetos sin permiso
Vandalismo	Daño de propiedad publica

Nota: Center for Research in Computer Vision, n.d.

Las categorías que contiene más de videos son “Normal” y “Ataque”, en cambio “Explosiones” y “Tiroteo” contienen menos videos. El tiempo de duración promedio de cada video varía entre 1 y 5 minutos, dependiendo del tipo de evento. Esto posibilita registrar el comportamiento completo del evento en el video, desde su comienzo hasta su finalización.

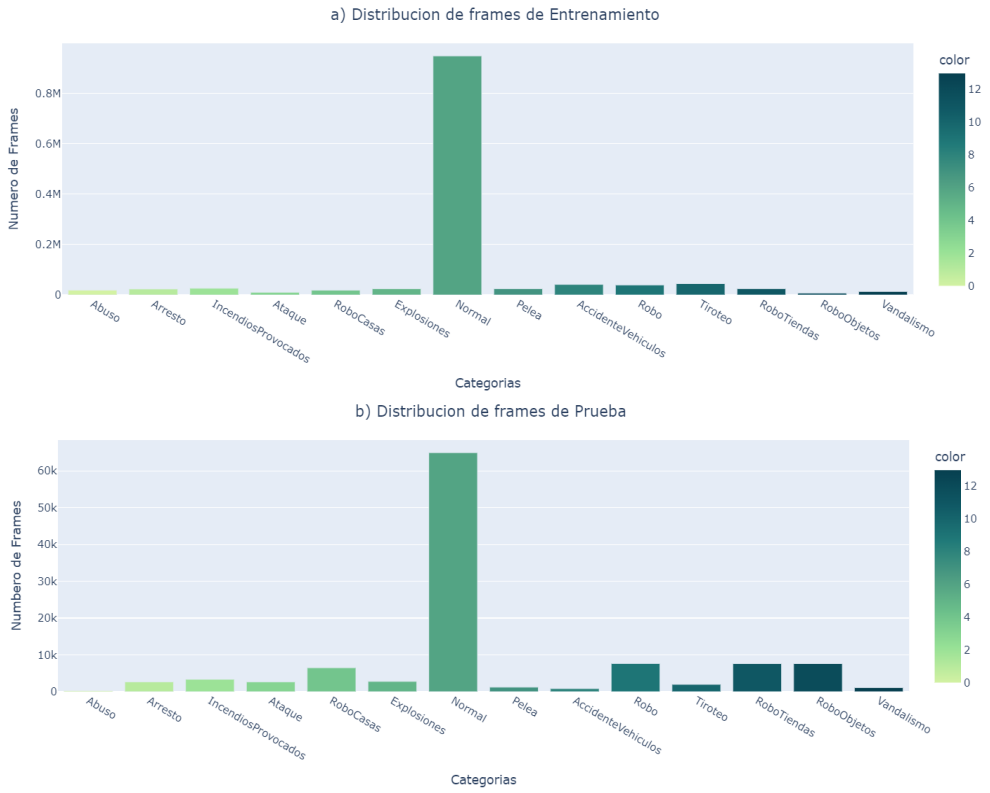
El conjunto de videos de entrenamiento y prueba de cada categoría corresponde al 80% y 20% respectivamente. Se crea un directorio para el conjunto de datos de entrenamiento y otro para prueba, en donde cada categoría corresponde a un subdirectorio. Los videos de cada categoría fueron segmentados en fotogramas de resolución 64x64 píxeles, etiquetados y almacenados en formato PNG. La etiqueta de cada fotograma guarda la información del video al que pertenece y la numero de la secuencia. La distribución de los fotogramas de cada categoría en los conjuntos de entrenamiento y prueba se muestra en la Figura 2.

Para mejorar la capacidad de generalización del modelo, se utilizaron técnicas de Aumento de datos (Maharana et al., 2022) utilizando la herramienta de Keras. Las transformaciones aplicadas al conjunto de entrenamiento fueron las siguientes: giros horizontales aleatorios, permiten aprender al modelo independientemente de la orientación del sujeto; desplazamiento horizontal y vertical, que simula cambios en la posición del objeto dentro del campo visual de la cámara; y reescalado de valores de píxeles, para normalizar las imágenes en el rango [0, 1].

Las transformaciones se aplicaron únicamente a los videos que se utilizaran en el conjunto de entrenamiento. Adicionalmente, las imágenes del conjunto de prueba se procesaron únicamente a través de un reescalado y una función de preprocesamiento, no se aplicaron modificaciones geométricas, con el objetivo de mantener la integridad del proceso de evaluación.

Figura 2

Distribución de Fotogramas de cada categoría del conjunto de: a) Entrenamiento; b) Prueba



Nota: Autores (2025).

Implementación de la etapa de extracción de características

Cada imagen es procesada mediante el modelo convolucional preentrenado EfficientNet, este modelo extrae las características de cada imagen con una alta eficiencia en la extracción de características, con un menor costo computacional a diferencia de otros modelos preentrenado (Ali et al., 2023).

Las capas finales del modelo EfficientNet fueron descongeladas para ajustarlas al conjunto de datos que tenemos. Nuestra salida que provee el modelo preentrenado es un vector

que recoge la información relevante de cada imagen. Al final estos vectores son organizados en un conjunto de secuencias para luego capturar la dinámica temporal del evento en la etapa siguiente de nuestra arquitectura.

El proceso de agrupación se lo realiza en una secuencia de 30 imágenes consecutivas, esto permite capturar la evolución del evento a lo largo del tiempo. Cada secuencia es representada por un tensor con la forma:

(n_secuencia,30,dim)

Donde dim representa la dimensión del vector de características extraído por EfficientNet.

Implementación de la etapa del modelo temporal

Una vez que se tiene las secuencias de imágenes, es necesario obtener la información temporal como espacial contenida en cada una de ellas. Esta etapa está conformada por los siguientes componentes:

- Capa ConvLSTM2D (Trinh et al., 2024): Procesa las secuencias de características mediante la combinación de las convoluciones 2D para capturar información espacial con una memoria temporal LSTM.
- BatchNormalization: Normaliza las activaciones para estabilizar el entrenamiento.

Implementación de la etapa de clasificación con capas densas

Una vez completado el modelado temporal, las salidas generadas son aplanadas para convertirlas en vectores unidemensionales (1D), los cuales permiten una transición eficiente hacia las capas densas del modelo. Estas capas densas se encargan de realizar la clasificación final, integrando las representaciones aprendidas tanto en la dimensión espacial como en la temporal. A través de este proceso, nuestro modelo es capaz de asignar una etiqueta a cada secuencia de video procesada.

La clasificación final no se restringe a una simple diferenciación entre “Anómalo” o “Normal”, sino que otorga una probabilidad de ser parte de una de las categorías de los eventos anómalos presentes en el conjunto de datos. Dentro de las categorías estudiadas se encuentran Abuso, Robo, Asalto, Vandalismo, entre otras, lo que transforma el modelo en un clasificador de múltiples categorías. En la capa de salida, se utiliza un softmax con 14 neuronas (una para cada categoría de evento), lo que facilita la clasificación de las secuencias en 14 posibles categorías.

Etapa de entrenamiento y evaluación

Nuestro modelo fue entrenado utilizando un optimizador Adam, el cual comúnmente es utilizado en escenarios de aprendizaje profundo por su capacidad para adaptar la tasa de aprendizaje durante el proceso de entrenamiento. La función categorical cross-entropy fue utilizada como función de pérdida, esta es adecuada para problemas de clasificación por multiclase.

La Tabla 2 muestra los hiperparámetros fundamentales que hemos utilizado en nuestra arquitectura durante el entrenamiento. Se empleó early stopping como herramienta para interrumpir el entrenamiento si la perdida de validación en nuestro entrenamiento no progresa.

Tabla 2
Hiperparámetros

Parámetro	Valor
Semilla Aleatoria	12
Tamaño de lote	64
Época	10
Tasa de aprendizaje	0.00003
Numero de Clases	14

Nota: Autores (2025).

Se optó por entrenar el modelo durante 10 épocas, ya que en las pruebas preliminares se observó que la curva de aprendizaje tendía a estabilizarse antes de la décima iteración. Esto

indicaba que el modelo era capaz de aprender las representaciones necesarias sin incurrir en sobreajuste, lo cual fue corroborado mediante la evaluación en el conjunto de validación.

En cuanto a la tasa de aprendizaje, se seleccionó un valor bajo (0.00003) con el objetivo de asegurar una convergencia estable, evitando grandes fluctuaciones en la función de pérdida durante el proceso de optimización. Dado que se trabaja con arquitecturas profundas y datos complejos como secuencias de video, un valor reducido permite realizar ajustes más finos en los pesos del modelo, lo que favorece una mejor generalización. Esta decisión también se basó en observaciones empíricas, donde tasas más altas resultaron en oscilaciones o incluso divergencias en el entrenamiento.

Para la evaluación del desempeño del modelo se empleó la técnica de división *holdout*. Esta estrategia permitió entrenar el modelo sobre una muestra amplia de los datos disponibles, validar su rendimiento durante el proceso de ajuste de hiperparámetros y, finalmente, evaluar su capacidad de generalización sobre un conjunto independiente. No se utilizó validación cruzada debido a las limitaciones computacionales y al elevado costo de procesamiento asociado al volumen de datos en formato de video.

Para el entrenamiento del modelo se utilizó una unidad de procesamiento gráfico (GPU) NVIDIA A100, proporcionada a través de la plataforma Google Colab. Esta configuración permitió acelerar significativamente el proceso de cálculo requerido por las redes neuronales profundas. El tiempo total de entrenamiento fue de aproximadamente cuatro horas, considerando el ajuste de los pesos a lo largo de múltiples épocas y la evaluación sobre los conjuntos de validación y prueba.

Resultados

Luego de completar el entrenamiento del modelo propuesto, se evaluó su rendimiento utilizando métricas tales como: AUC (*Área Bajo la Curva ROC*), que proporciona una medida

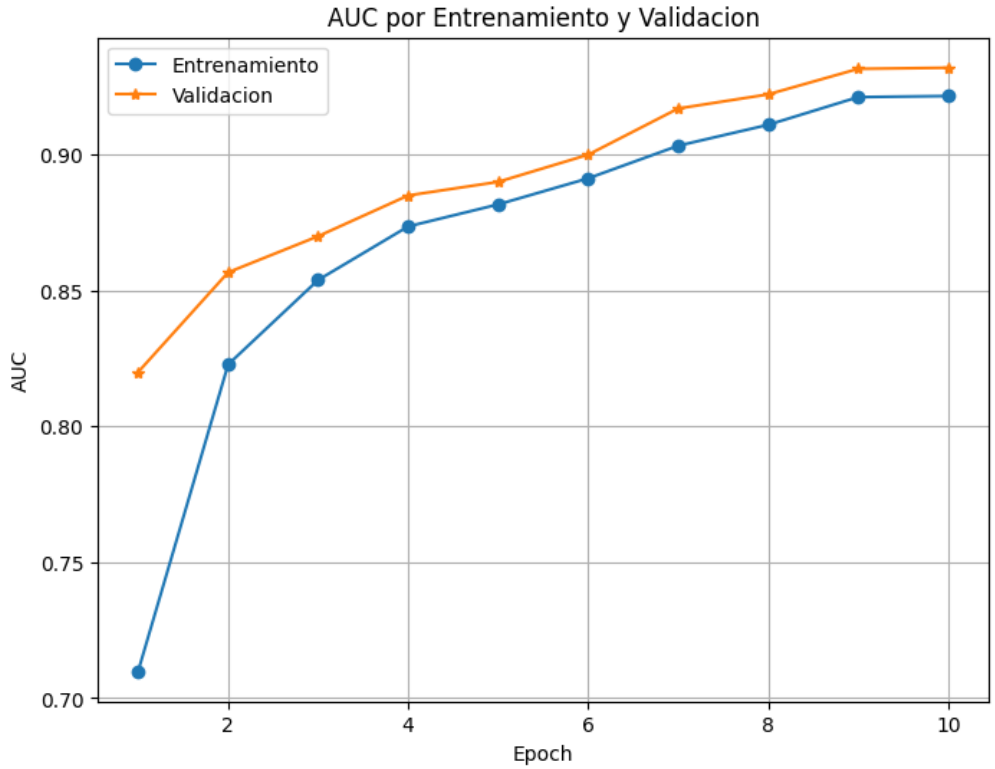
completa del rendimiento del modelo en problemas de clasificación multiclase; y curvas ROC (*Receiver Operating Characteristic*) las cuales permiten analizar la capacidad de clasificación del modelo en cada una de las clases del conjunto de datos. A continuación, se detallan los principales hallazgos.

1. **Análisis del AUC por cada época.**

La Figura 3 muestra los resultados del AUC que se han obtenido en proceso de entrenamiento y validación. Se puede observar un incremento constante en ambas curvas, lo que demuestra la capacidad que tiene el modelo para diferenciar el tipo de evento.

Figura 1

Evolución del AUC durante el entrenamiento y la validación.



Nota: Autores (2025).

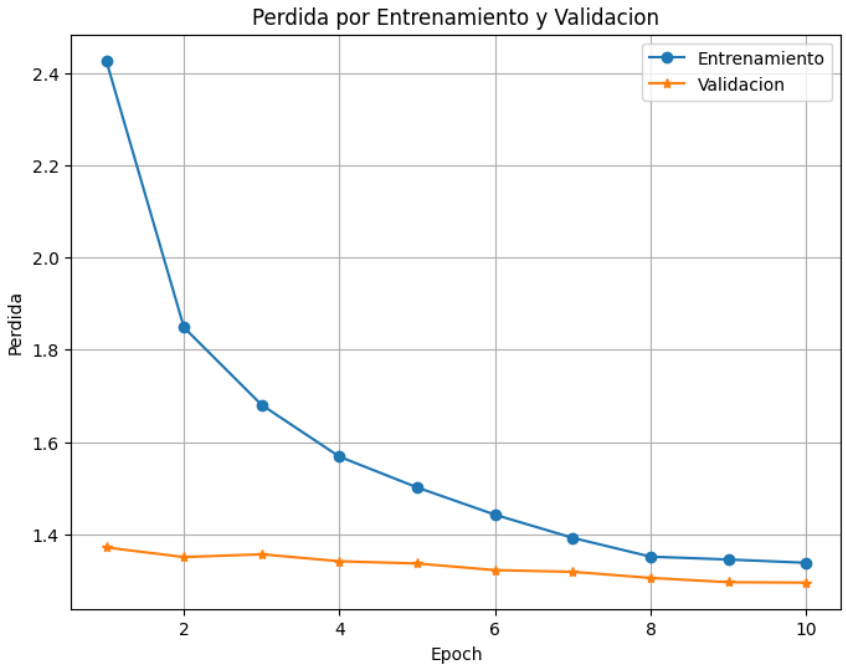
2. **Análisis de la función de perdida**

La Figura 4 muestra el progreso de la función de pérdida durante las 10 épocas de entrenamiento. Se observa una reducción constante en la pérdida en nuestro conjunto de entrenamiento, pasando de 2.42 a menos de 1.35, lo que indica que el modelo ha mejorado su

habilidad para ajustarse nuestros datos específicos. De igual forma, desde la segunda época la perdida de validación ha permanecido relativamente estable, registrando valores cercanos a 1.30.

Figura 2

Evolución de la pérdida durante el entrenamiento y la validación

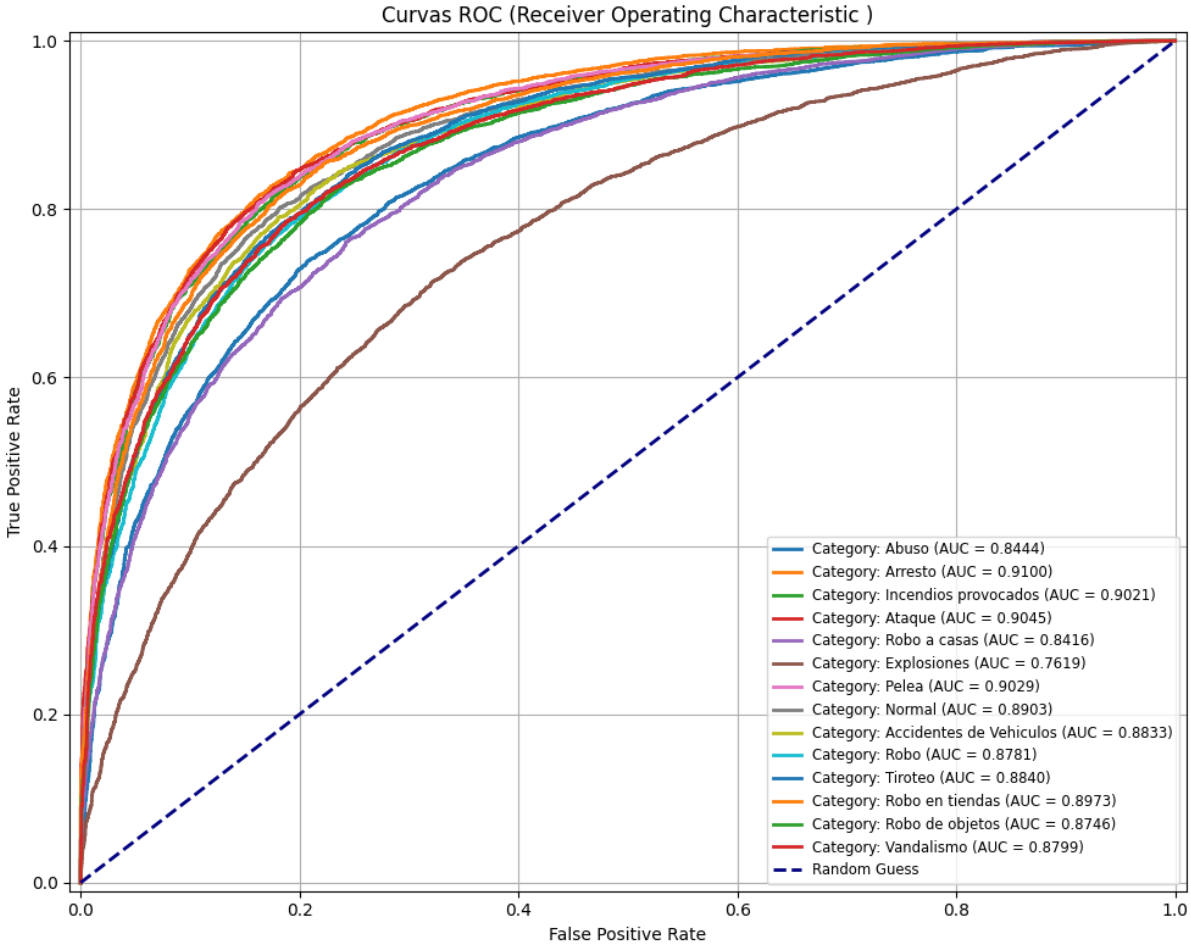


Nota: Autores (2025).

3. Análisis de la Curva ROC por categoría

La Figura 5 presenta las curvas ROC generadas para cada una de las categorías evaluadas por el modelo, en donde se observa un rendimiento destacado en clases como "Pelea" (AUC = 0.90), "Arresto" (AUC=0.91) y "Ataque" (AUC=0.90), lo que indica que el modelo logra identificar con alta precisión estos tipos de eventos. Adicionalmente, las categorías "Robo " (AUC = 0.87) y "Robo en objetos" (AUC=0.87) presentan un desempeño más moderado, posiblemente debido a la similitud visual entre clases o al desbalance en el número de muestras disponibles. De manera general, todas las curvas muestran un comportamiento ascendente pronunciado, lo que refleja una buena capacidad del modelo para diferenciar las clases.

Figura 3
Curvas ROC por categoría



Nota: Autores (2025).

4. Análisis de la evaluación Global

El valor final de AUC global sobre el conjunto de prueba fue de 0.8795, lo que corrobora que el modelo tiene una alta capacidad para diferenciar y clasificar eventos anómalos en secuencias de video.

Los resultados obtenidos en este estudio, evaluados mediante el AUC por categoría, muestran un rendimiento competitivo frente a modelos previamente propuestos en la literatura. En comparación con Sultani et al. (2018) que utiliza un enfoque débilmente supervisado con aprendizaje basado en *Multiple Instance Learning*, reportaron un AUC promedio de 0.75 para el conjunto de datos UCF-Crime. En contraste, nuestro modelo logra valores superiores en

categorías como “Pelea" (AUC = 0.90), "Arresto" (AUC=0.91) y "Ataque" (AUC=0.90), lo que refleja una mejora sustancial en la capacidad discriminativa.

La Tabla 3 describe los valores del AUC de cada clase que se han obtenido al evaluar nuestro modelo. Las métricas presentadas muestran que nuestro modelo tiene la capacidad de diferenciar correctamente los eventos de una categoría de los demás.

En general, el hecho de que todas las categorías superen un AUC de 0.80 confirma que el modelo mantiene un buen desempeño multicategórico, siendo capaz de adaptarse a las diferencias visuales y temporales entre distintos tipos de eventos.

Tabla 3
Resultados de evaluación global del modelo

Categoría	AUC
Abuso	0.84
Arrestos	0.91
Incendios provocados	0.90
Ataques	0.90
Robo a Casas	0.84
Explosiones	0.76
Peleas	0.90
Normal	0.89
Accidente de vehículos	0.88
Robo	0.87
Tiroteo	0.88
Robo en Tiendas	0.89
Robo de objetos	0.87
Vandalismo	0.87

Nota: Autores (2025).

Discusión

A pesar del rendimiento general positivo del modelo, es importante destacar varias limitaciones técnicas que condicionaron tanto su diseño como su desempeño.

En primer lugar, se identificaron restricciones computacionales significativas. Aunque se utilizó una GPU NVIDIA A100 a través de Google Colab, el entrenamiento de modelos que procesan secuencias de video es intensivo en memoria y tiempo. Esto impuso límites sobre la cantidad de datos que podían ser procesados simultáneamente, el tamaño de las secuencias, y la complejidad de la arquitectura. En un entorno de producción o con datasets aún más grandes, sería necesario disponer de infraestructura dedicada para garantizar escalabilidad y eficiencia.

Nuestro modelo alcanzó un AUC global de 0.8795, validando de esta forma su capacidad de generalización. Además, se obtuvieron valores superiores a 0.90 en categorías como “Arrestos”, “Peleas”, “Ataques” y “Incendios Provocados”, lo que refleja que la red logra identificar de forma precisa ciertos tipos de eventos con patrones espaciales y temporales bien definidos. No obstante, algunas clases presentaron un rendimiento inferior, como “Abuso” (AUC = 0.84) y “Explosiones” (AUC = 0.76) lo que indica que ciertos tipos de eventos aún representan un reto para el sistema. Estas limitaciones pueden atribuirse, en parte, a un desequilibrio en la distribución del número de ejemplos por clase, ya que pude llegar a favorecer a aquellas clases con más representaciones en el conjunto de entrenamiento. Asimismo, la similitud visual entre algunas categorías puede dificultar la discriminación precisa. Por ejemplo, las escenas etiquetadas como “Robo” pueden compartir patrones espaciales similares con “Robo de objetos” o “Robo en tiendas”, lo que dificulta la correcta diferenciación incluso para modelos entrenados.

Otro factor que se debe considerar es la alta variabilidad intra-clase. En el caso de la clase “Explosiones” variables como; la iluminación, el ángulo de la cámara o duración del evento pueden cambiar drásticamente de un fotograma a otro, afectando la consistencia de las representaciones aprendidas.

Para abordar estas limitaciones, se proponen las siguientes estrategias para trabajos futuros:

- Aplicar técnicas de balanceo de clases como SMOTE *(Synthetic Minority Over-sampling Technique)* (Chawla, Bowyer, Hall y Kegelmeyer, 2002), que mediante una variante adaptada puede operar sobre los vectores de características extraídos, generando ejemplos sintéticos para las clases subrepresentadas.
- Implementar mecanismos de atención como CBAM *(Convolutional Block Attention Module)* (Woo, Park, Lee y Kweon, 2018), dentro del extractor de características CNN para que el modelo pueda resaltar regiones relevantes del fotograma.
- Aumentar el tamaño de las secuencias o utilizar marcos intermedios que capturen mejor la progresión temporal del evento.
- Incorporar modelos híbridos que combinen CNN, LSTM, junto a mecanismos que mejoren la capacidad de diferenciar eventos similares.

Conclusión

En el presente estudio se desarrolló e implementó un modelo con una arquitectura hibrida conformada por un modelo extractor de características espaciales basado en EfficientNet, una capa ConvLSTM2D para el modelado de dependencias temporales, y capas densas que realizan la clasificación de eventos anómalos en secuencias de video. Esta combinación permitió capturar de forma eficiente tanto patrones espaciales como temporales, esenciales para el reconocimiento de comportamientos complejos en entornos dinámicos. El modelo demostró un rendimiento sólido, reflejado en los siguientes resultados:

- El AUC global de 0.8795, indicando una alta capacidad de discriminación general.
- AUC mayores a 0.90 en varias categorías.
- Una curva de pérdida convergente y estable a lo largo del proceso de entrenamiento, sin evidencia de sobreajuste.

El desempeño de nuestro modelo en las categorías más importantes, como “Arrestos”, “Incendios provocados”, “Ataques” y “Peleas”, demuestran la efectividad que tiene para ser aplicado en entornos de videovigilancia reales, donde la detección temprana de eventos críticos es fundamental para la seguridad en ambientes comunitarios. La alta precisión alcanzada valida la viabilidad de implementar modelo en sistemas automatizados de vigilancia, con el fin de disminuir la carga operativa en los entes de seguridad y mejorar así la capacidad de respuesta ante incidentes.

La arquitectura híbrida propuesta en este trabajo se presenta como una alternativa adecuada para sistemas de videovigilancia inteligentes, al ofrecer una clasificación robusta y eficiente de comportamientos anómalos. En conjunto, este desarrollo representa una contribución al campo de la visión por computadora orientada a la seguridad ciudadana, y sienta las bases para futuras investigaciones enfocadas en optimizar el monitoreo automatizado en contextos comunitarios.

Más allá del desempeño técnico, el modelo desarrollado posee un notable potencial de impacto social y comunitario. Su implementación en sistemas de videovigilancia automatizada podría fortalecer la seguridad ciudadana, optimizar la capacidad de respuesta ante eventos críticos y reducir la carga operativa en centros de monitoreo. Sin embargo, su adopción debe considerar aspectos éticos como la privacidad, la transparencia en las decisiones del modelo y la supervisión humana en entornos sensibles.

Como proyección futura, se plantea optimizar la arquitectura mediante modelos más ligeros como MobileNet (Sinha & El-Sharkawy, 2019) o EfficientNet-lite (Ab Wahab et al., 2021), que permitan su ejecución en dispositivos con recursos limitados. Asimismo, se propone adaptar el sistema para detección en tiempo real mediante técnicas de reducción de latencia, e incorporar mecanismos de interpretabilidad como Grad-CAM (Chen et al., 2020) que faciliten la comprensión del proceso de decisión. También se sugiere enriquecer el modelo con

información multimodal (como audio o sensores contextuales) y evaluar su desempeño en entornos reales y diversos, considerando variabilidad ambiental y social.

Referencias bibliográficas

Ab Wahab, M. N., Nazir, A., Ren, A. T. Z., Noor, M. H. M., Akbar, M. F., & Mohamed, A. S. A. (2021). EfficientNet-Lite and hybrid CNN-KNN implementation for facial expression recognition on Raspberry Pi. *IEEE Access, 9*, 134065–134080. https://doi.org/10.1109/ACCESS.2021.3113337

Ali, M. S., Hassan, A., Rahim, A., Ashraf, M. H., Rahim, A., & Saghir, S. (2023). Motor imagery EEG classification using fine-tuned deep convolutional EfficientNetB0 model. In *2023 3rd International Conference on Artificial Intelligence (ICAI)* (pp. 1–6). IEEE.

Center for Research in Computer Vision. (n.d.). *UCF-Crime dataset*. University of Central Florida. Retrieved March 18, 2025, from https://www.crcv.ucf.edu/research/real-world-anomaly-detection-in-surveillance-videos/

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, L., Chen, J., Hajimirsadeghi, H., & Mori, G. (2020, March). Adapting Grad-CAM for embedding networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Freire-Obregón, D., Barra, P., Castrillón-Santana, M., & De Marsico, M. (2021). Inflated 3D ConvNet context analysis for violence detection. *Machine Vision and Applications, 33*(1), 15. https://doi.org/10.1007/s00138-021-01264-9

Khan, L. U., Yaqoob, I., Tran, N. H., Kazmi, S. M. A., Dang, T. N., & Hong, C. S. (2020). Edge-computing-enabled smart cities: A comprehensive survey. *IEEE Internet of Things Journal, 7*(10), 10200–10232. https://doi.org/10.1109/JIOT.2020.2987070

Koonce, B. (2021). EfficientNet. In *Convolutional neural networks with Swift for TensorFlow: Image recognition and dataset categorization* (pp. 109–123). Springer.

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings, 3*(1), 91–99.

Martínez-Mascorro, G. A., Abreu-Pederzini, J. R., Ortiz-Bayliss, J. C., & Terashima-Marín, H. (2020). Suspicious behavior detection on shoplifting cases for crime prevention by using 3D convolutional neural networks. *arXiv preprint* arXiv:2005.02142.

Mohanapriya, S., Saranya, S. M., Dinesh, K., Jawaharsrinivas, S., Lintheshwar, S., & Logeshwaran, A. (2024). Anomaly detection in video surveillance. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1–5). https://doi.org/10.1109/ICCCNT61001.2024.10725557

Myagmar-Ochir, Y., & Kim, W. (2023). A survey of video surveillance systems in smart city. *Electronics, 12*(17), Article 3567. https://doi.org/10.3390/electronics12173567

Pham, H. H., Khoudour, L., Crouzil, A., Zegers, P., & Velastin, S. A. (2022). Video-based human action recognition using deep learning: A review. *arXiv preprint* arXiv:2208.03775.

Prakash, S., Jalal, A. S., & Pathak, P. (2023). Forecasting COVID-19 pandemic using Prophet, LSTM, hybrid GRU-LSTM, CNN-LSTM, Bi-LSTM and Stacked-LSTM for India. In *2023 6th International Conference on Information Systems and Computer Networks (ISCON)* (pp. 1–6). https://doi.org/10.1109/ISCON57294.2023.10112065

Revathi, A. R., & Kumar, D. (2017). An efficient system for anomaly detection using deep learning classifier. *Signal, Image and Video Processing, 11*(2), 291–299. https://doi.org/10.1007/s11760-016-0935-0

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., & Woo, W. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *arXiv preprint* arXiv:1506.04214. http://arxiv.org/abs/1506.04214

Sinha, D., & El-Sharkawy, M. (2019, October). Thin MobileNet: An enhanced MobileNet architecture. In *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 280–285). IEEE. https://doi.org/10.1109/UEMCON47517.2019.8993089

Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6479–6488).

Tan, M., & Le, Q. (2019, May). EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning* (pp. 6105–6114). PMLR.

Trinh, T.-D., Vu-Ngoc, T.-S., Le-Nhi, L.-T., Le, D.-D., Nguyen, T.-B., & Pham, T.-B. (2024). Violence detection in videos based on CNN feature for ConvLSTM2D. In *Proceedings of the 5th ACM Workshop on Intelligent Cross-Data Analysis and Retrieval* (pp. 33–36).

Varela-Tapia, E. A., Acosta-Guzmán, I. L., Fajardo-Romero, I. J., & Oviedo-Peñafiel, J. A. (2024). *Inteligencia Artificial Aplicada con técnicas de Procesamiento de Lenguaje Natural y Machine Learning en el campo de la salud.* Editorial Grupo AEA. https://doi.org/10.55813/egaea.l.83

Wang, Z., Yang, Y., Liu, Z., & Zheng, Y. (2023). Deep neural networks in video human action recognition: A review. *arXiv preprint* arXiv:2305.15692.

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3–19). https://doi.org/10.1007/978-3-030-01234-2_1

Zahra, A., Ghafoor, M., Munir, K., Ullah, A., & Ul Abideen, Z. (2024). Application of region-based video surveillance in smart cities using deep learning. *Multimedia Tools and Applications, 83*(5), 15313–15338. https://doi.org/10.1007/s11042-021-11468-w