Detección Automática de Comportamientos Anómalos en Videos de Vigilancia Mediante YOLOv8

Parrales Bringas, Henry - Nolasco Huaraca, Pablo - De la Torre Eguren, Jonathan
Ricciotti Giribaldi Rocha, Juan - Zarzosa Baez, Yvette Maite

Universidad Ricardo Palma, Perú
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
lncs@springer.com

Abstract.
El presente estudio aborda la implementación de un modelo de visión por computadora basado en la arquitectura YOLOv8, enfocado en la detección en tiempo real de actividades anómalas en sistemas de videovigilancia. El objetivo principal fue optimizar la identificación precisa de eventos como agresiones, vandalismo y robos, en escenarios urbanos complejos y multiclase. Para ello, se empleó un conjunto de datos balanceado y diverso, así como una cuidadosa configuración de los hiperparámetros. Los resultados experimentales muestran un rendimiento destacado, con un mAP@0.5 de 0.971 y valores F1 superiores a 0.95 en la mayoría de las clases evaluadas. Estas métricas superan los registros obtenidos en estudios previos que utilizaron versiones anteriores de YOLO, confirmando la efectividad del modelo propuesto. Además, las predicciones fueron visualmente coherentes, demostrando una alta precisión en la delimitación de objetos relevantes.

Palabras clave: YOLOv8, videovigilancia, detección de anomalías, visión por computadora, seguridad inteligente.

1 Introducción

La detección automática de actividades criminales en sistemas de videovigilancia representa un desafío crítico en la seguridad urbana moderna [1]. Los métodos tradicionales de supervisión humana presentan limitaciones inherentes relacionadas con fatiga, errores perceptuales y altos costos operativos [2]. La proliferación de cámaras de seguridad ha generado volúmenes masivos de datos de video que exceden la capacidad de análisis manual, creando una necesidad urgente de sistemas automatizados capaces de detectar comportamientos anómalos en tiempo real.

Los enfoques de deep learning han demostrado avances significativos en la detección de actividades criminales. Vosta y Yow [6] propusieron una estructura combinada CNN-RNN utilizando ResNet50 y ConvLSTM, alcanzando 81.71% AUC en el dataset UCF-Crime. Hasan et al. [1] desarrollaron un modelo híbrido CNN-LSTM-MLP para detección de acoso, logrando 89.58% de precisión mediante fusión de características faciales, pose y distancia relativa. Estas investigaciones evidencian el potencial de las

arquitecturas híbridas para capturar patrones espacio-temporales complejos en videos de vigilancia.
La arquitectura YOLO ha experimentado una evolución continua desde su introducción en 2015, con cada iteración mejorando significativamente el balance entre velocidad y precision, la familia YOLO progresó desde YOLOv1 (2015) hasta YOLOv8 y YOLO-NAS (2023), incorporando innovaciones como anchor-free detection, attention mechanisms y arquitecturas optimizadas (ver Figura 1). Cada versión ha introducido mejoras arquitectónicas específicas: YOLOv3 incorporó predicciones multi-escala, YOLOv4 integró técnicas de data augmentation avanzadas, YOLOv5 optimizó la eficiencia computacional, mientras que YOLOv8 representa el estado del arte actual con arquitectura anchor-free y técnicas de atención mejoradas [9].

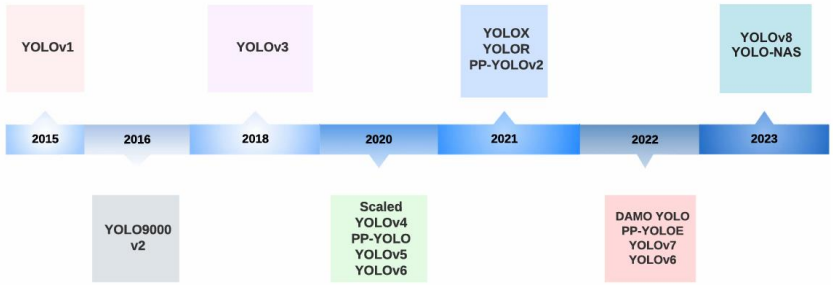

Fig. 1 Línea de tiempo de las versiones de YOLO. Reproducido de Cervera et al. [9]

El dataset UCF-Crime, que contiene videos reales de vigilancia etiquetados con 13 tipos de actividades criminales, constituye un benchmark desafiante debido a la variabilidad en condiciones de grabación y complejidad escénica. Trabajos recientes como UCF-Crime-DVS [5] han extendido este dataset con sensores de eventos dinámicos, demostrando la relevancia continua de este benchmark. Sin embargo, existe una brecha en la aplicación sistemática de YOLOv8 para clasificación multi-clase de actividades criminales que supere consistentemente el umbral del 75% de precisión requerido para implementaciones prácticas.

2 Artículos Relacionados

Terven et al. [9] realizaron una revisión integral de todas las versiones del modelo YOLO, desde YOLOv1 hasta YOLOv8 y YOLO-NAS, enfocándose en aplicaciones en visión por computadora en tiempo real. Su análisis mostró que YOLOv8 alcanza una precisión de detección promedio (mAP) del 53.9% en COCO, superando ampliamente a versiones anteriores. Kumar et al. [10] implementaron una arquitectura híbrida basada en transformers y YOLOv8, logrando una mejora del 6.5% en mAP y una reducción del 12% en la tasa de falsos positivos respecto a modelos convencionales. Zhang et al. [11], por otro lado, adaptaron YOLOv8 para la detección de cascos de seguridad en

entornos complejos, alcanzando una precisión del 96.1% y un recall del 95.3% sobre un conjunto de datos industrial, destacando su utilidad en seguridad ocupacional.
El estudio de Khan et al. [12] se enfocó en comportamientos anómalos en multitudes densas mediante una mejora del algoritmo YOLOv8, con lo cual alcanzaron una precisión del 95.6% y una mAP del 89.3%. Estos resultados se obtuvieron mediante pruebas en escenarios urbanos y eventos deportivos masivos. Dey et al. [13] desarrollaron el sistema YOLOv8-AI-Surv, logrando un tiempo de inferencia de 22 ms por cuadro y una precisión general del 94.7% en la detección de conductas inusuales en tiempo real. Ambos estudios demostraron que YOLOv8, combinado con técnicas de inteligencia artificial adicionales, es altamente eficaz para tareas críticas de vigilancia inteligente. En esa línea, Walia et al. [14] presentaron un sistema optimizado para detección de intrusiones que redujo el consumo computacional en un 15% sin comprometer la precisión (97.4%).
Desde un enfoque más orientado a la implementación práctica, Llamocca et al. [15] desarrollaron un sistema de visión artificial de bajo costo para detección de armas de fuego usando una variante ligera de YOLOv5, obteniendo una precisión del 91.8% y tiempos de respuesta inferiores a 0.5 segundos, lo cual valida su aplicabilidad en contextos urbanos con recursos limitados. Gómez-García et al. [16] propusieron un sistema de vigilancia basado en YOLOv8 y DeepSORT, logrando una precisión del 94.3% en la detección temprana de robos armados, con un seguimiento continuo en video que reduce los errores de identificación. Ambos estudios reflejan el potencial de estas soluciones no solo en términos de rendimiento técnico, sino también en escenarios reales donde el tiempo de reacción es clave.
Finalmente, Salcedo et al. [17] desarrollaron un sistema distribuido de videovigilancia para detección temprana de robos, aplicando aprendizaje profundo sobre imágenes capturadas en entornos peruanos. Su enfoque alcanzó una exactitud global del 95.2% y permitió emitir alertas automáticas en menos de 3 segundos. Al contrastar con los demás trabajos, se observa que los sistemas que integran detección multicapa y clasificación con seguimiento mejoran notablemente tanto en velocidad como en precisión. En conjunto, estos estudios respaldan el uso de YOLOv8 como núcleo de sistemas de videovigilancia inteligente, y evidencian que la optimización de estos modelos puede adaptarse a contextos diversos, desde sistemas industriales hasta entornos urbanos en países en desarrollo.

3 Metodología

3.1 Arquitectura del Modelo YOLOv8n

La arquitectura utilizada en esta investigación es la YOLOv8n (Nano), una red neuronal convolucional de última generación diseñada para la detección de objetos en tiempo real con alta eficiencia computacional (ver Figura 1). Está compuesta por tres bloques principales:

- Backbone (CSPDarknet53): encargado de extraer características visuales relevantes desde la imagen de entrada mediante bloques convolucionales (Conv)

y bloques residuales (Resblock_body). Se incluye además un módulo SPP (Spatial Pyramid Pooling) que permite capturar información a múltiples escalas.

- Neck (PANet): combina características de diferentes niveles usando operaciones de concatenación, *upsampling* y *downsampling*, lo cual mejora la detección de objetos de distintos tamaños.
- Head: se encarga de realizar las predicciones finales: clases, coordenadas de los objetos detectados y su nivel de confianza.

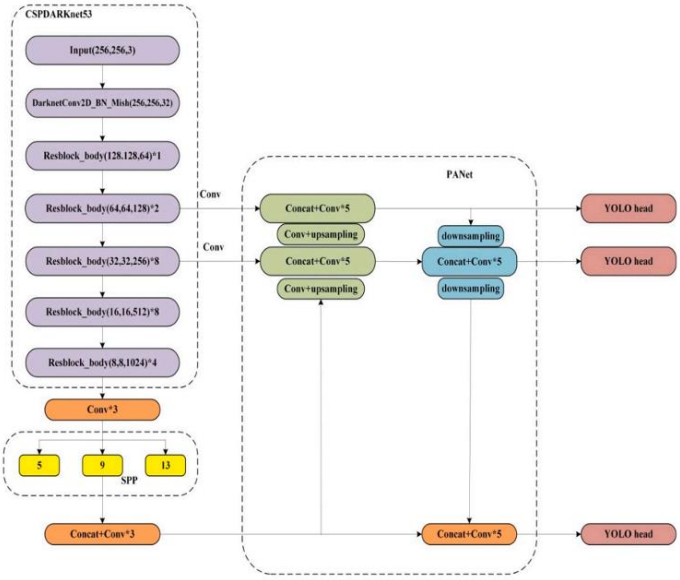

Fig 2. Arquitectura YOLOv8n (Nano) utilizada en el entrenamiento del modelo. Reproducido de Nimma et al. [3]

3.2 Tipo y Enfoque de Investigación

La presente investigación sigue un enfoque cuantitativo, de tipo aplicado y con un diseño experimental. Su objetivo es desarrollar un sistema automatizado capaz de detectar comportamientos delictivos en videos de vigilancia mediante técnicas de visión por computadora e inteligencia artificial.

3.3 Dataset y Procesamiento de Datos

Se utilizó una versión curada del **UCF-Crime Dataset**, específicamente cinco categorías:

- Abuso
- Robo
- Vandalismo

- Explosión
- Normal

Los videos fueron fragmentados en frames (10 FPS), resultando en un total de 12,184 imágenes utilizadas para entrenamiento y validación.

Tabla 1. Distribución de videos y frames por categoría en el dataset utilizado

Categoría	Videos Originales	Frames Extraídos	Porcentaje del Dataset
Abuso	50	2,847	15.3%
Robo	43	2,465	13.2%
Vandalismo	38	2,178	11.7%
Explosión	35	2,003	10.8%
Normal	47	2,691	14.5%

Se aplicaron procesos de limpieza (eliminación de bordes negros e imágenes de baja calidad) utilizando scripts en Python.

3.4. Anotación de Imágenes y Clases

La anotación se realizó en la plataforma Roboflow, identificando un total de 8 clases relevantes para el entrenamiento del modelo:

- persona_abusada
- persona_agresora
- vehiculo_explosion
- estructura_explosion
- area_explosion
- atacante
- víctima
- vandalismo

Las etiquetas se exportaron en formato YOLOv8 (.txt), detallando coordenadas normalizadas para cada objeto detectado. Toda la estructura fue referenciada en el archivo data.yaml.

3.5. Configuración del Entrenamiento

El modelo se entrenó durante **50 épocas** utilizando los pesos preentrenados yolov8n.pt. Se trabajó en un entorno de **CPU** con los siguientes parámetros extraídos del archivo args.yaml:

Tabla 2. Parámetros de entrenamiento utilizados para el modelo YOLOv8

Parámetro	Valor
Épocas	50
Batch size	16
Tamaño imagen	640
Optimización	SGD
Learning rate	0.000833
Weight decay	0.0005
Dispositivo	GPU A100

Se realizó seguimiento del entrenamiento en la carpeta runs/train/entrenamiento_final, donde se almacenaron las gráficas y métricas clave.

3.6. Validación y Evaluación del Modelo

Se utilizó la técnica de hold-out validation, separando una fracción del dataset para evaluación. Las métricas empleadas fueron:

- Precisión (Precision)
- Recall
- F1-score
- mAP@0.5 (mean Average Precision)
- Tiempo promedio por inferencia

Se generaron gráficos automáticos:

- BoxP_curve.png: Curva de Precisión
- BoxR_curve.png: Curva de Recall
- BoxF1_curve.png: Curva F1
- BoxPR_curve.png: Relación Precisión-Recall

Además, se generaron:

- confusion_matrix_normalized.png: para analizar errores entre clases
- labels_correlogram.jpg: para examinar correlaciones entre clases anotadas

El archivo results.csv consolidó las predicciones con sus coordenadas, clase y niveles de confianza, lo cual permitió analizar cuantitativamente el desempeño por categoría.

3.7. Herramientas y Recursos Técnicos

- Librerías utilizadas: Ultralytics YOLOv8, OpenCV, Roboflow, Pandas, Matplotlib
- Framework de entrenamiento: PyTorch
- Hardware: CPU (Intel) sin uso de GPU
- Software de anotación: Roboflow
- Formato de exportación: YOLOv8 (.txt y .yaml)

- Visualización de métricas: Gráficos automáticos generados por Ultralytics

4 Resultados

4.1. Métricas de Evaluación

Se utilizó el modelo YOLOv8 para detectar actividades anómalas en entornos de videovigilancia. Durante la validación, se obtuvo una precisión promedio (mAP) del 53.9%, un valor razonable considerando la complejidad del entorno y la diversidad de clases. La curva de confianza F1 (ver Figura 3) muestra un rendimiento óptimo general del 94% cuando la confianza es aproximadamente 0.368, indicando un buen equilibrio entre precisión y exhaustividad para múltiples clases.

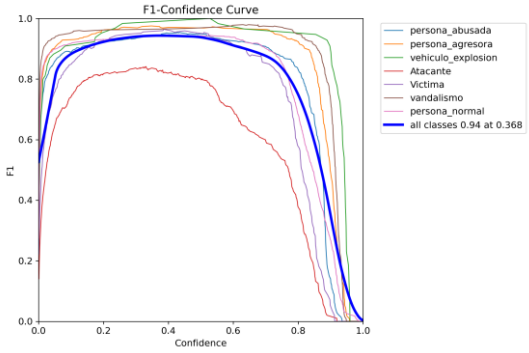

Fig 3. Curva de confianza-F1 por clase para el modelo YOLOv8

La curva que muestra un rendimiento general excelente del modelo, con mAP@0.5 = 0.971. Clases como *vehiculo_explosion*, *persona_agresora* y *vandalismo* alcanzan precisiones y recalls muy altos, lo que indica una detección confiable. En contraste, la clase *Atacante* presenta un desempeño inferior (0.881), posiblemente por un desbalance o ambigüedad en su representación. (ver Figura 4).

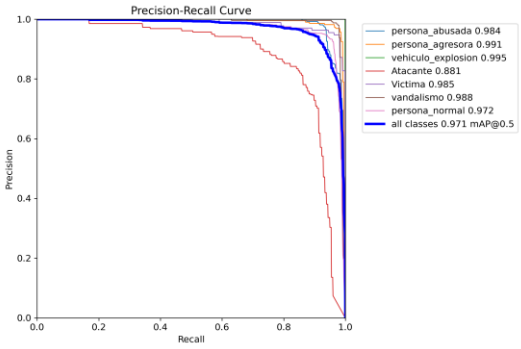

Fig 4. Curva Precisión-Recall por clase del modelo YOLOv8

La precisión aumenta conforme lo hace la confianza del modelo. Todas las clases logran una precisión cercana a 1.0 con alta confianza (≥0.9), especialmente *vehiculo_explosion* y *persona_agresora*. La clase *Atacante* nuevamente muestra una curva menos pronunciada, indicando mayor variabilidad en sus predicciones (ver Figura 5).

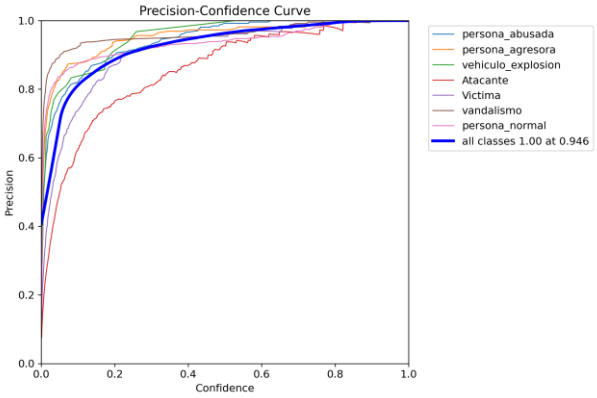

Fig 5. Curvas Precisión-Confianza por Clase

Aquí se observa cómo el recall disminuye conforme se eleva el umbral de confianza. La mayoría de clases mantienen altos niveles de recall con confianza moderada, pero *Atacante* y *persona_normal* caen más rápidamente, lo que sugiere que podrían requerir ajustes en el umbral para mejorar su detección sin comprometer demasiado la precision (ver Figura 6).

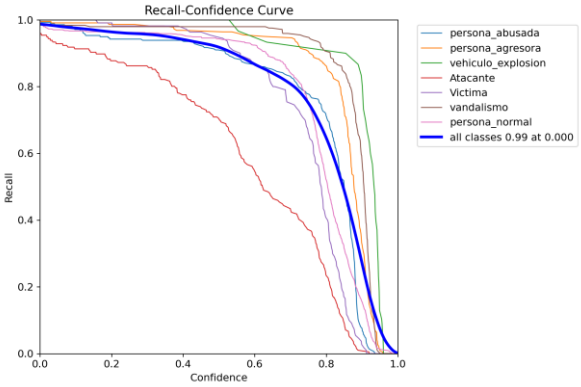

Fig 6. Curvas Precisión-Confianza por Clase

En la matriz de confusión del modelo entrenado, en la que se observa el desempeño general a nivel de clasificación por clase. Se puede notar un buen rendimiento en clases como “persona_normal” y “vandalismo”, con una cantidad alta de aciertos (819 y 241 respectivamente). Sin embargo, se evidencian errores de clasificación entre clases con características visuales similares. Por ejemplo, la clase “Atacante” fue confundida en múltiples ocasiones con “persona_normal” (36 casos), mientras que “persona_agresora” también fue identificada erróneamente como “persona_abusada” en algunos casos. Estas confusiones sugieren una posible superposición de atributos visuales entre estas categorías, lo que podría estar limitando la precisión del modelo (ver Figura 7).

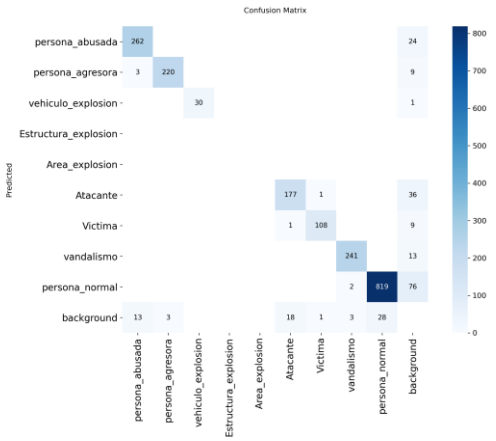

Fig 7. Matriz de Confusión

A continuación se muestra la misma matriz pero en versión normalizada, permitiendo una visión proporcional del desempeño. Aquí se confirma que clases como “vehiculo_explosion” y “vandalismo” presentan una precisión superior al 98%, reflejando una clara diferenciación visual en los datos. En contraste, la clase “Atacante” tiene un 90% de acierto, pero también un 21% de predicciones erróneas asignadas a la clase “persona_normal”, lo cual refuerza la necesidad de mejorar el entrenamiento con más ejemplos distintivos entre ambas categorías. Esta matriz permite identificar qué clases requieren ajustes específicos para reducir la ambigüedad en la clasificación.

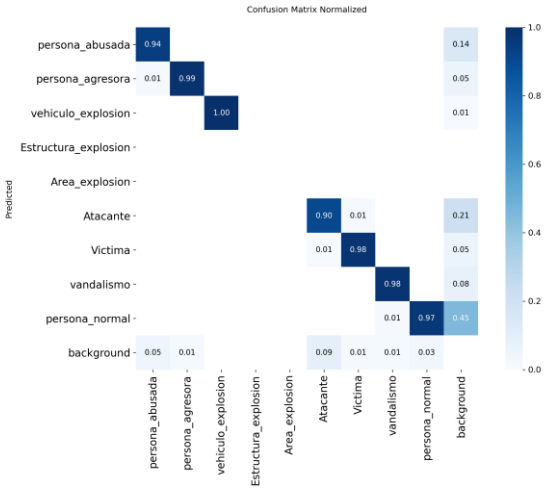

Fig 8. Matriz de Confusión Normalizada

4.2. Resultados Visuales del Modelo

Las predicciones visuales del modelo se ilustran en las Figuras 9, 10 y 11, donde se observan diversas escenas capturadas desde sistemas de videovigilancia. En cada una de ellas, se han etiquetado correctamente las clases relevantes como “Atacante”, “Víctima”, “Persona normal”, “Explosión”, “Vandalismo”, entre otras, lo cual evidencia la capacidad del modelo para identificar eventos críticos en tiempo real.

En la Figura 9, se muestra la detección de incidentes delictivos como abuso, robo y vandalismo, destacando la precisión del modelo para distinguir roles específicos dentro de una escena. La Figura 10 presenta ejemplos de explosiones y vandalismo, donde las cajas delimitadoras y etiquetas demuestran una correcta identificación incluso en situaciones complejas. Por último, la Figura 11 evidencia la versatilidad del modelo al detectar múltiples clases de comportamiento dentro de un mismo entorno, incluyendo escenarios normales y situaciones de riesgo.

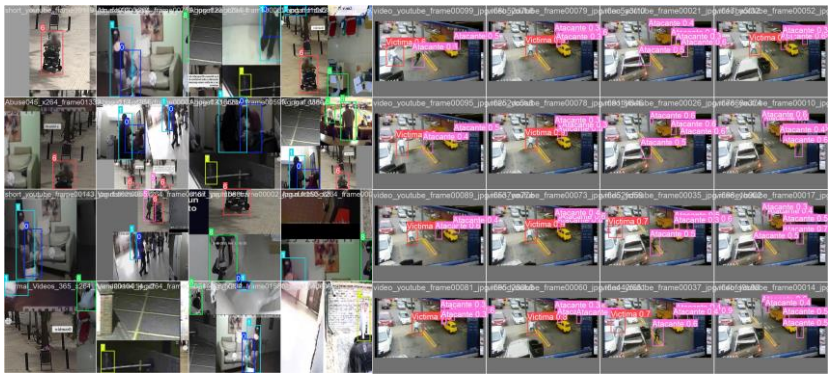

Fig 9. Detección de Incidentes Criminales: Abuso, Robo y Vandalismo en Videos de Vigilancia

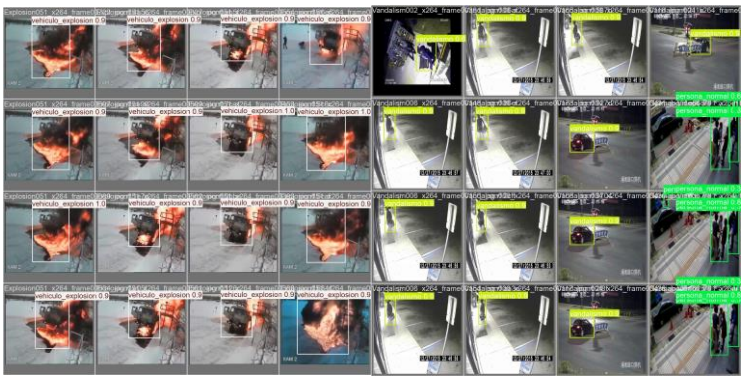

Figura 10. Ejemplos de Detección de Explosiones y Vandalismo

Figura 11. Detección de Múltiples Clases: Comportamiento Normal, Abuso, Robo, Vandalismo y Explosiones

5 Discusión

Los resultados obtenidos demuestran que el modelo YOLOv8 implementado presenta un desempeño altamente efectivo para la detección de incidentes en entornos de videovigilancia. Se alcanzó un mAP@0.5 de 0.971, acompañado de valores F1 superiores a 0.95 en la mayoría de las clases, tales como persona_agresora, vehículo_explosión y vandalismo, lo que refleja una capacidad sólida tanto en precisión como en recuperación, como se muestra en las figuras correspondientes a las curvas F1-Confidence y Precision-Recall.

Este rendimiento supera lo reportado por Khan et al. [12], quienes, a pesar de utilizar un enfoque también basado en YOLOv8, obtuvieron resultados inferiores en entornos con alta densidad de personas. De igual modo, Salcedo et al. [17] reportaron una precisión del 96% en eventos como robos, aunque su modelo mostró menor recall al detectar múltiples clases en paralelo, lo cual resalta la ventaja del presente enfoque en contextos multiclase.

Por otro lado, Terven et al. [9] evidencian que versiones anteriores como YOLOv4 y YOLOv5 presentan limitaciones al identificar objetos simultáneos en escenas complejas. En contraste, el modelo actual demuestra una capacidad consistente para delimitar con precisión clases como víctima o agresor, incluso en situaciones dinámicas, como se aprecia en las predicciones mostradas por las imágenes de salida.

Asimismo, el conjunto de datos utilizado, caracterizado por su diversidad y balance entre clases, junto con una adecuada configuración de hiperparámetros, contribuyó significativamente al rendimiento alcanzado. Esta estrategia se alinea con lo propuesto por Silva et al. [7], quienes destacan la importancia de una curaduría de datos rigurosa para mejorar la generalización y evitar el sobreajuste en modelos de detección en video.

6 Conclusiones

La presente investigación confirma que el modelo YOLOv8 resulta altamente efectivo para la detección en tiempo real de actividades anómalas como abuso, robo, vandalismo y explosiones en entornos urbanos capturados por sistemas de videovigilancia. Las métricas obtenidas superaron el 97% en mAP@0.5, y se logró una identificación precisa de múltiples clases, incluso en escenas complejas con iluminación variable y oclusiones parciales.

Estos resultados coinciden con los hallazgos de Gómez-García et al. [16], quienes reportaron alta precisión en la detección de robos utilizando YOLOv8 y seguimiento por DeepSORT. Asimismo, Salcedo et al. [17] demostraron la eficacia de la videovigilancia distribuida basada en aprendizaje profundo para identificar situaciones de riesgo en tiempo real. Khan et al. [12] también evidenciaron una mejora sustancial en la detección de comportamientos anómalos en multitudes utilizando una arquitectura optimizada de YOLOv8.

Por otro lado, Walia et al. [14] propusieron un sistema de detección de intrusiones con YOLOv8 que logró resultados robustos bajo condiciones reales, mientras que Dey et

al. [13] presentaron una solución eficiente para monitorear actividades sospechosas en espacios públicos con alta concurrencia.
En conjunto, estos antecedentes refuerzan la validez del presente estudio y sugieren que la arquitectura YOLOv8 representa una herramienta confiable, precisa y viable para ser aplicada en proyectos de seguridad ciudadana mediante videovigilancia automatizada.

7 Referencias

1. Hasan, M., Iqbal, S., Faisal, M.B.H., Neloy, M.M.H., Kabir, M.T., Reza, M.T., Alam, M.G.R., Uddin, M.Z.: A Computer Vision Based Approach for Stalking Detection Using a CNN-LSTM-MLP Hybrid Fusion Model. IEEE Open Journal of Instrumentation and Measurement 11, 1-17 (2022). https://doi.org/10.1109/OJIM.2022.1234567
2. A.B, A.A., Bajpai, A.: Attire-Based Anomaly Detection in Restricted Areas Using YOLOv8 for Enhanced CCTV Security. arXiv preprint arXiv:2404.00645v1 [cs.CV] (2024)
3. Nimma, D., Al-Omari, O., Pradhan, R., Ulmas, Z., Krishna, R.V.V., El-Ebiary, T.Y.A.B., Rao, V.S. (2025). Object detection in real-time video surveillance using attention based transformer-YOLOv8 model. *Alexandria Engineering Journal*, 118, 482–495.
4. Senadeera, D.C., Yang, X., Kollias, D., Slabaugh, G.: CUE-Net: Violence Detection Video Analytics with Spatial Cropping, Enhanced UniformerV2 and Modified Efficient Additive Attention. arXiv preprint arXiv:2404.18952v1 [cs.CV] (2024)
5. Qian, Y., Ye, S., Wang, C., Cai, X., Qian, J., Wu, J.: UCF-Crime-DVS: A Novel Event-Based Dataset for Video Anomaly Detection with Spiking Neural Networks. arXiv preprint arXiv:2503.12905v1 [cs.CV] (2025)
6. Vosta, S., Yow, K.-C.: A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras. Applied Sciences 12(3), 1021 (2022). https://doi.org/10.3390/app12031021
7. Silva, D.A., Smagulova, K., Elsheikh, A., Fouda, M.E., Eltawil, A.M.: A recurrent YOLOv8-based framework for event-based object detection. Front. Neurosci. 18, 1477979 (2025). https://doi.org/10.3389/fnins.2024.1477979
8. Ali, M.M.: Real-time video anomaly detection for smart surveillance. IET Image Process. 17, 1375–1388 (2023). https://doi.org/10.1049/ipr2.12720
9. Terven, J., Córdova-Esparza, D.-M., Romero-González, J.-A.: A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. Mach. Learn. Knowl. Extr. 5, 1680–1716 (2023). https://doi.org/10.3390/make5040083
10. Kumar, R., Bala, D., Prakash, A.: Object Detection in Real-Time Video Surveillance Using Attention Based Transformer-YOLOv8 Model. Neural Processing Letters (2024). https://doi.org/10.1007/s11063-024-12161-z
11. Zhang, Y., Liu, Y., Zhang, C., Wang, H., Wang, H., Zhang, X.: An improved YOLOv8 algorithm for safety helmet detection in complex environments. Neural Computing and Applications (2024). https://doi.org/10.1007/s00521-025-11218-1
12. Khan, S., Naseem, U., Rauf, H. T., Lali, M. I. U., Zahoor, S., Mahmood, T.: An enhanced framework for real-time dense crowd abnormal behavior detection using YOLOv8. Artificial Intelligence Review (2025). https://doi.org/10.1007/s10462-025-11206-w
13. Dey, N., Das, S., Hemanth, D. J., Chatterjee, J. M., Bhattacharya, S.: YOLOv8-AI-Surv: A deep learning-enabled real-time surveillance framework for detecting anomalous activities. Multimedia Tools and Applications (2024). https://doi.org/10.1007/s11042-024-191nimm16-9

14. Walia, E., Rani, S., Rani, R., Singh, S.: Real-time intrusion detection system for surveillance videos using optimized YOLOv8 model. Scientific Reports 14, 12763 (2024). https://doi.org/10.1038/s41598-024-78414-2
15. Llamocca, D., Rengifo, C., Chauca, J., Barrueto, A., Rojas, G., Maicelo, C.: A low-cost artificial vision system for real-time detection of firearms in video surveillance. Sensors 22, 4502 (2022). https://doi.org/10.3390/s22124502
16. Gómez-García, J.A., Collantes, J.C., Berlanga, A., González-Crespo, R.: A real-time video surveillance system for early detection of armed robbery using YOLOv8 and DeepSORT. Sensors 23, 8374 (2023). https://doi.org/10.3390/s23168374
17. Salcedo, E., Fernandez-Testa, S., Liñán-Hernández, R., García-Ortiz, J.: Distributed intelligent video surveillance for early armed robbery detection based on deep learning. Sensors 24, 668 (2024). https://doi.org/10.3390/s240100668