



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado

Facultad de Ingeniería de Sistemas e Informática

Unidad de Posgrado

**Modelo de visión computacional para la detección de
actividad sospechosa en bodegas de Trujillo usando
redes neuronales**

TESIS

Para optar el Grado Académico de Magíster en Ingeniería de
Sistemas e Informática con mención en Ingeniería de Software

AUTOR

Zandor Yandari SÁNCHEZ AGREDA

ASESOR

Mg. Jorge Leoncio GUERRA GUERRA

Lima, Perú

2026



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Sánchez, Z. (2026). *Modelo de visión computacional para la detección de actividad sospechosa en bodegas de Trujillo usando redes neuronales*. [Tesis de maestría, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Unidad de Posgrado]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Zandor Yandari Sánchez Agreda
Tipo de documento de identidad	DNI
Número de documento de identidad	47543567
URL de ORCID	https://orcid.org/0009-0005-1174-638X
Datos de asesor	
Nombres y apellidos	Jorge Leoncio Guerra Guerra
Tipo de documento de identidad	DNI
Número de documento de identidad	08473333
URL de ORCID	https://orcid.org/0000-0003-2392-7821
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Giovana Melva Valverde Ayala
Tipo de documento	DNI
Número de documento de identidad	08536719
Miembro del jurado 1	
Nombres y apellidos	William Martin Enriquez Maguiña
Tipo de documento	DNI
Número de documento de identidad	06179457
Miembro del jurado 2	
Nombres y apellidos	Herminio Paucar Curasma
Tipo de documento	DNI
Número de documento de identidad	32920084

Datos de investigación	
Línea de investigación	C.0.5.8. Inteligencia artificial
Grupo de investigación	No aplica.
Agencia de financiamiento	Sin financiamiento.
Ubicación geográfica de la investigación	Edificio: Universidad Nacional Mayor de San Marcos País: Perú Departamento: Lima Provincia: Lima Distrito: Cercado de Lima Latitud: -12.055926 Longitud: -77.083469
Año o rango de años en que se realizó la investigación	2024 - 2025
URL de disciplinas OCDE	Ingeniería de sistemas y comunicaciones https://purl.org/pe-repo/ocde/ford#2.02.04



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Vicedecanato de Investigación y Posgrado
Unidad de Posgrado

**ACTA DE SUSTENTACIÓN DE TESIS PARA OPTAR EL GRADO ACADÉMICO DE
MAGÍSTER EN INGENIERÍA DE SISTEMAS E INFORMÁTICA CON MENCIÓN EN
INGENIERÍA DE SOFTWARE**

A los cuatro (4) días del mes de febrero de 2026, siendo las 12:00 pm., se reunieron en el Auditorio Alfredo Celso Alva Bravo, el Jurado de Tesis conformado por los siguientes docentes:

Dra. Giovana Melva Valverde Ayala (Presidente)
Mg. William Martin Enriquez Maguiña (Miembro)
Mg. Herminio Paucar Curasma (Miembro)
Mg. Jorge Leoncio Guerra Guerra (Miembro Asesor)

Se inició la Sustentación invitando al candidato a Magíster **ZANDOR YANDARI SÁNCHEZ AGREDA**, para que realice la exposición oral de la tesis para optar el Grado Académico de Magíster en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software, siendo la Tesis intitulada:

**"MODELO DE VISIÓN COMPUTACIONAL PARA LA DETECCIÓN DE ACTIVIDAD
SOSPECHOSA EN BODEGAS DE TRUJILLO USANDO REDES NEURONALES"**

Concluida la exposición, los miembros del Jurado de Tesis procedieron a formular sus preguntas que fueron absueltas por el graduando; acto seguido se procedió a la evaluación correspondiente, habiendo obtenido la siguiente calificación:

..... Diecisiete (17) Muy bueno

Por tanto, el presidente del Jurado, de acuerdo con el Reglamento General de Estudios de Posgrado, otorga al Bachiller **ZANDOR YANDARI SÁNCHEZ AGREDA** el Grado Académico de Magíster en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software.

Siendo las 13:45 horas, el presidente del Jurado de Tesis, da por concluido el acto académico de Sustentación de Tesis.


Dra. Giovana Melva Valverde Ayala
(Presidente)


Mg. Herminio Paucar Curasma
(Miembro)

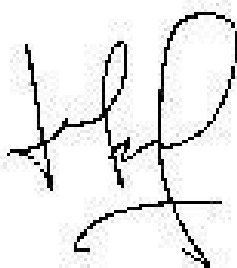

Mg. William Martin Enriquez Maguiña
(Miembro)


Mg. Jorge Leoncio Guerra Guerra
(Miembro Asesor)

CERTIFICADO DE SIMILITUD

Yo, Jorge Leoncio Guerra Guerra en mi condición de asesor acreditado con Dictamen N 000340-2024-UPG-VDIP-FISI/UNMSM, de la tesis, cuyo título es **Modelo de Visión Computacional para la Detección de Actividad Sospechosa en Bodegas de Trujillo usando Redes Neuronales** presentado por el bachiller **Zandor Yandari Sánchez Agreda**, para optar el grado académico de Magister en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software, CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante la plataforma PIDS, el documento evaluado cuenta con el 4 % de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional.**

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del grado académico de Magister en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software.



Lima, 11 de noviembre de 2025

Firma del Asesor

Nombres y Apellidos del asesor: Jorge Leoncio Guerra Guerra

DNI 08473333

DEDICATORIA

A Dios por la fortaleza que me brinda a través de la Fe.

A mis madres, María Valentina Agreda Ulloa y Lucía
Ysabel Agreda Ulloa, por apoyarme en cada decisión
y demostrarme su amor día a día.

AGRADECIMIENTO

No tengo palabras suficientes para agradecer a mis madres, María Valentina Agreda Ulloa y Lucía Ysabel Agreda Ulloa, por su amor incondicional, apoyo constante y paciencia desde el inicio hasta la culminación de la presente tesis.

Manifiesto mi especial gratitud al catedrático Mg. Jorge Leoncio Guerra Guerra, asesor de esta tesis, por su valiosa orientación y los aportes académicos que enriquecieron significativamente este trabajo de investigación.

Agradezco de manera particular a la empresa que me brindó acceso a los datos de videovigilancia necesarios para realizar este estudio, permitiendo que esta investigación tuviera un sustento empírico sólido.

A los docentes de la Universidad Nacional Mayor de San Marcos que forman parte de la Maestría en Ingeniería de Sistemas e Informática, por todo su aporte académico, transferencia de conocimientos y valiosos consejos científicos que contribuyeron a mi formación profesional y al desarrollo de esta tesis.

Finalmente, expreso mi gratitud a la Universidad Nacional Mayor de San Marcos por contribuir a mi crecimiento académico al aceptarme en esta prestigiosa institución.

ÍNDICE GENERAL

DEDICATORIA	2
AGRADECIMIENTO	3
LISTA DE FIGURAS	8
LISTA DE TABLAS	9
RESUMEN	10
ABSTRACT	12
CAPÍTULO 1: INTRODUCCIÓN	14
1.1 Situación Problemática	14
1.2 Formulación del Problema	17
1.2.1 Problema General	17
1.2.2 Problemas Específicos	17
1.3 Justificación Teórica	18
1.4 Justificación Práctica	18
1.5 Objetivos de la Investigación	19
1.5.1 Objetivo General	19
1.5.2 Objetivos Específicos	19
CAPÍTULO 2: MARCO TEÓRICO	20
2.1 Marco filosófico o epistemológico de la investigación	20
2.2 Antecedentes del Problema	21
2.3 Bases Teóricas	61
2.3.1 Visión Computacional	61
2.3.1.1 Procesamiento Digital de Imágenes	61
2.3.1.2 Red Neuronal Convolucional	62
2.3.1.3 Detección y Seguimiento de Objetos	62
2.3.1.3.1 Detección de Personas	63
2.3.1.3.2 Seguimiento de Objetos	64
2.3.1.3.3 Características Espacio-Temporales	65
2.3.2 Análisis Espacio-Temporal	65
2.3.2.1 Representación de Trayectorias	66
2.3.2.2 Normalización de Datos	66
2.3.3 Datos Secuenciales	67
2.3.3.1 Redes Neuronales Recurrentes	67
2.3.3.2 Redes de Memoria a Largo y Corto Plazo	68
2.3.3.2.1 Componentes	69
2.3.3.2.2 Datos de Entrenamiento, Validación y Pruebas	70
2.3.3.3 Optimización	71
2.3.3.3.1 Algoritmo ADAM	71
2.3.3.3.2 Error Cuadrático Medio	72
2.3.4. Aprendizaje No Supervisado	73
2.3.4.1 Modelado de Patrones Normales	73

2.3.4.2 Anomalías	74
2.3.4.3 Detección de Anomalías	74
2.3.5 Actividades Sospechosas	75
2.3.5.1 Clasificación	76
2.3.5.2 Merodeo	77
2.4 Marcos Conceptuales	77
CAPÍTULO 3: METODOLOGÍA	81
3.1 Tipo y Diseño de Investigación	81
3.1.1 Identificación de Variables	81
3.1.2 Enfoque de Investigación	81
3.1.3 Tipo de Investigación	81
3.1.4 Nivel o Alcance de Investigación	82
3.1.5 Diseño de Investigación	82
3.2 Unidad de Análisis	83
3.3 Población de Estudio	84
3.4 Muestra	84
3.5 Técnicas de Recolección de Datos	85
3.6 Descripción del Dataset	85
3.6.1 Naturaleza de los datos y tipo de aprendizaje	87
3.6.2 Anonimización del dataset	87
3.7 Diagrama del Pipeline	87
CAPÍTULO 4: RESULTADOS Y DISCUSIÓN	93
4.1 Análisis, Interpretación y Discusión de Resultados	93
4.1.1 Análisis de Resultados	93
4.1.2 Interpretación de Resultados	96
4.1.3 Discusión de Resultados	99
4.1.3.1 Cumplimiento del Objetivo General	100
4.1.3.2 Cumplimiento de los Objetivos Específicos	102
4.1.3.2.1 Cumplimiento del Primer Objetivo Específico	102
4.1.3.2.2 Cumplimiento del Segundo Objetivo Específico	104
4.1.3.2.3 Cumplimiento del Tercer Objetivo Específico	105
4.2 Pruebas de Hipótesis	107
4.2.1 Hipótesis General	107
4.2.2 Hipótesis Específicas	108
4.2.3 Prueba para la Hipótesis General	108
4.2.4 Prueba para la Primera Hipótesis Específica	109
4.2.5 Prueba para la Segunda Hipótesis Específica	111
4.2.6 Prueba para la Tercera Hipótesis Específica	113
4.3 Presentación de Resultados	115
4.3.1 Resultados del Objetivo General	116
4.3.2 Resultados de los Objetivos Específicos	116

4.3.3 Validación Integral del Modelo	117
CONCLUSIONES	118
RECOMENDACIONES	120
REFERENCIAS BIBLIOGRÁFICAS	121
ANEXOS	140
ANEXO A: Clasificación de papers según fuentes	140
ANEXO B: Pseudocódigo del seguimiento de objetos con YOLO	140
ANEXO C: Parámetros de entrenamiento del modelo LSTM	141
ANEXO D: Ejemplo de ejecución del módulo de detección de trayectorias sospechosas	142

LISTA DE FIGURAS

Figura 1 Arquitectura CNN Generalizada para Detección de Actividades Sospechosas	32
Figura 2 Arquitectura GAN Generalizada para Detección de Actividades Sospechosas	33
Figura 3 Arquitectura CNN-LSTM Generalizada para Detección de Actividades Sospechosas	39
Figura 4 Esquema de la propuesta CNN-LSTM para la Detección de Actividad Sospechosa	58
Figura 5 Representaciones de objetos. (a) Centroide, (b) diversos puntos, (c) parches rectangulares, (d) parche elíptico, (e) diversos parches basados en partes, (f) armazón del objeto, (g) borde entero del objeto, (h) marcadores en el borde del objeto, (i) silueta del objeto	62
Figura 6 Proceso de preparación del dataset de aprendizaje para modelos de Visión Computacional	69
Figura 7 Selección representativa de clips de video del dataset utilizado en el estudio.....	85
Figura 8 Esquema del proceso de extracción de trayectorias en la Fase 1	87
Figura 9 Ejemplos de detección de trayectorias y tabla de datos extraídos	87
Figura 10 Consolidación Secuencial de archivos CSV y Filtrado de Trayectorias	88
Figura 11 Integración y resultado final del archivo CSV consolidado	89
Figura 12 Generación de secuencias temporales mediante ventanas deslizantes	89
Figura 13 Pseudocódigo para la generación de ventanas deslizantes para el entrenamiento	90
Figura 14 Esquema del Entrenamiento del Modelo LSTM	91
Figura 15 Distribución de frecuencia de las longitudes de las trayectorias	95
Figura 16 Visualización del sistema de seguimiento para la Trayectoria 12	100
Figura 17 Detección y trazado de trayectorias humanas mediante YOLO11 en 42 secuencias de video	102
Figura 18 Procesamiento llevado a cabo para el Cumplimiento del Segundo Objetivo Específico	103
Figura 19 Procedimiento llevado a cabo para el Cumplimiento del Tercer Objetivo Específico	105

LISTA DE TABLAS

Tabla 1 Criterios de Inclusión y Exclusión de Artículos Científicos	21
Tabla 2 Agrupación de artículos según diversos tipos de Arquitectura	22
Tabla 3 Agrupación de artículos según diversos tipos de Técnicas	24
Tabla 4 Agrupación de artículos según tipos de Métricas	27
Tabla 5 Arquitecturas de Visión Computacional para detectar Actividad Sospechosa en video	28
Tabla 6 Técnicas de Visión Computacional para detectar Actividad Sospechosa en video...	40
Tabla 7 Métricas para evaluar las técnicas de detección de Actividad Sospechosa en video	50
Tabla 8 Comparación de modelos para detección sospechosa versus propuesta YOLO11-LSTM.....	59
Tabla 9 Especificaciones técnicas del dataset de trayectorias normales	84
Tabla 10 Estadísticas descriptivas de las trayectorias	92
Tabla 11 Distribución de trayectorias según el número de puntos	94
Tabla 12 Resultados de detección de anomalías en trayectorias de desplazamiento sospechosas	99

RESUMEN

En el norte del Perú, la inseguridad ha alcanzado niveles críticos, afectando significativamente a las bodegas y otros comercios de escala reducida. Datos provistos por la Asociación de Bodegueros del Perú indican que aproximadamente 250000 establecimientos son víctimas de delitos anualmente, generando pérdidas económicas que superan los 500000 soles diarios. Hasta julio de 2025, más de 2000 bodegas han cerrado de forma indefinida debido a la extorsión y los robos. En respuesta, los propietarios han incrementado la instalación de cámaras de seguridad como medida disuasiva; sin embargo, carecen de la capacidad para detectar actividades sospechosas a través del análisis de video.

Frente a este escenario, la visión computacional constituye una alternativa tecnológica para resolver esta dificultad, dado que emplea las redes neuronales artificiales para replicar determinadas capacidades de la visión humana. De manera especial, las redes neuronales convolucionales (CNN) muestran ventajas significativas al momento de localizar y etiquetar objetos en imágenes y videos. No obstante, su limitación radica en la falta de memoria temporal, lo que dificulta el análisis de secuencias dinámicas en grabaciones de videovigilancia. Para superar esta restricción, las redes de memoria a largo y corto plazo (LSTM) proporcionan un complemento ideal, debido al potencial que manejan para procesar datos secuenciales y capturar patrones temporales con precisión.

Esta investigación propone el desarrollo de una arquitectura de visión computacional que integra una red neuronal convolucional con unidades de memoria a largo y corto plazo, cuyo propósito es detectar actividades sospechosas de merodeo en entornos de bodegas. El enfoque metodológico consta de dos etapas principales: primero, la detección de objetos tipo persona en videos, incluyendo el trazado de las trayectorias; segundo, el escrutinio de dichas trayectorias para identificar patrones asociados a comportamientos anómalos. Los resultados muestran que la propuesta logra un índice de detección del 100 % en la identificación de trayectorias relacionadas con el merodeo. Además, la red neuronal convolucional reportó una precisión del 98.33 % al asignar categorías a objetos,

mientras que la red neuronal recurrente modificada exhibe consistencia al predecir comportamientos normales, manteniendo consistencia en la detección de anomalías. Esta propuesta representa un avance significativo en la aplicación de modelos de visión computacional para fortalecer la seguridad en bodegas, ofreciendo una solución innovadora y efectiva frente a los desafíos de la inseguridad en el contexto peruano.

ABSTRACT

Northern Peru faces unprecedented security challenges that have severely impacted small-scale retail businesses, particularly neighborhood convenience stores. Statistical evidence from the Peruvian Grocery Store Owners Association reveals that approximately 250,000 commercial establishments fall victim to criminal activities annually, resulting in economic losses exceeding 500,000 soles per day. By July 2025, over 2,000 convenience stores had permanently ceased operations due to systematic extortion and theft. Consequently, business owners have increasingly invested in surveillance camera systems as a deterrent mechanism; however, these installations lack the capability to automatically identify suspicious activities through video analysis.

Within this framework, computer vision emerges as a pivotal technology to address these security concerns, leveraging artificial neural networks to replicate specific aspects of human visual perception. Convolutional Neural Networks (CNNs) demonstrate particular proficiency in object detection and classification tasks within images and video sequences. Nevertheless, their primary limitation lies in the absence of temporal memory capabilities, which constrains their effectiveness in analyzing dynamic sequences captured by surveillance systems. To overcome this constraint, Long Short-Term Memory (LSTM) networks present themselves as an optimal complement, given their capacity to process sequential data and accurately capture temporal patterns.

This research presents the creation of a computer vision model that integrates a Convolutional Neural Network with a Long Short-Term Memory network, specifically designed to detect suspicious loitering activities within convenience store environments. The methodological approach encompasses two primary phases: initially, person detection in video footage and trajectory mapping; subsequently, trajectory analysis to identify patterns associated with anomalous behaviors. Performance assessment reveals that the proposed model attains a 100% detection rate in identifying trajectories related to loitering behaviors. Furthermore, the convolutional neural network component exhibited 98.33% accuracy in object

classification, while the modified recurrent neural network demonstrates consistency in predicting normal behaviors and maintaining reliability in anomaly detection. This proposal represents a major breakthrough in the use of computer vision models for enhancing security in convenience stores, providing an innovative and effective solution to address security challenges within the Peruvian context.

CAPÍTULO 1: INTRODUCCIÓN

1.1 Situación Problemática

En el ámbito global, en los Estados Unidos de América, la delincuencia en el comercio minorista costó aproximadamente cuantías por encima de cien mil millones de dólares en el 2022; asimismo, en el mismo período, pero tomando el 2021 como punto de inicio, las pérdidas en el mercado australiano y neozelandés fueron entre 1200 y 3100 millones de dólares australianos, respectivamente, según Townsley y Hutchins (2024). El mismo estudio señala que el robo externo representa más de la mitad (53.3%) de los 4300 millones de dólares en pérdidas para ambos países. Para el comercio minorista, esto genera mucha preocupación, dado que las pérdidas aumentaron un 28% en el transcurso de 2017 a 2018 mientras las ganancias solo aumentaron un 25% durante el mismo período. Es preciso señalar que, con base en el estudio realizado por Glover et al. (2021), el robo minorista se presenta con mayor frecuencia en comunidades de bajos recursos económicos. Además, las tiendas o bodegas representan el 57.9% de los robos perpetrados a comercios minoristas; y, el dinero de los negocios representa el 43% del total de artículos robados. Según DeCelles et al. (2022), los comercios minoristas estadounidenses son objetivos frecuentes del crimen. La mayor cantidad de incidentes violentos y muertes en la industria minorista está relacionada con el robo. Connealy y Piza (2019) aseguran que las tiendas o bodegas conformaron el único factor de riesgo asociado a los objetivos que tiene el delincuente al momento de perpetrar el delito. En Sudáfrica, Thurtell (2021) afirma que el crimen minorista es muy común, dado que el valor de los productos robados representa una ganancia sustancial para el delincuente. En España, Fernández (2021) afirma que se registraron 493479 y 642579 casos de robo con violencia en el 2021 y 2022, respectivamente; en otras palabras, el número de casos se incrementó en más del 30% en un año. Del mismo modo, los casos de robos a casas y otras instalaciones presentan el doble de incidentes de sustracción violenta.

La región de América Latina y el Caribe registra los mayores índices de violencia del planeta, con un promedio de homicidios cuatro veces superior al promedio

internacional y un costo económico empresarial relacionado al crimen dos veces superior al de los países desarrollados, según Mohan (2021); asimismo, en el Caribe, se reporta que el 23% de los comercios tuvieron pérdidas económicas debido al crimen, por ende el 70% de los comercios caribeños aumentaron el presupuesto en acciones de seguridad. Sumado a ello, Pinazo-Dallenbach (2020) menciona que la delincuencia no organizada se centra en áreas urbanas, aumentando la incertidumbre en las diversas actividades empresariales. En México, según Bonilla y Gutiérrez (2021), los territorios urbanos con residencias y comercios registran el mayor número de robos, tanto violentos como no violentos, con un ratio de aproximadamente 90 robos por cada 1000 habitantes en El Rosario y 85 robos por cada 1000 habitantes en Lomas de Santa Fe. De igual manera, el autor señala que las bodegas o tiendas son atractivas para el robo de peatones; sin embargo, casi el 90% de las víctimas no reportan los actos delictivos. De forma análoga a otros países de la región, en Argentina, el Ministerio de Seguridad declara que se registraron 313767 y 366516 casos de hurtos en el 2022 y 2023, respectivamente; es decir, hubo un incremento anual del 17%.

En Perú, Ronceros (2020) posiciona a la delincuencia como uno de los problemas sociales fundamentales, alcanzando un 41.1%. Asimismo, en el 2018, el 25.7% de la población peruana fue víctima de la delincuencia, el 86.3% vive con el miedo de ser una futura víctima de la criminalidad; y, los delitos contra el patrimonio, los asaltos, hurtos y arrebatos tuvieron un crecimiento anual de 4.5%. De igual manera, Otiniano (2020) asegura que, durante el período de 2018 a 2019, en algunos departamentos del Perú, como Lima, Callao y Tumbes, más del 90% de la población tiene miedo debido al aumento de los asaltos a empresarios y los robos a negocios. En el primer trimestre de 2021, Lima Metropolitana, Arequipa y La Libertad registraron el mayor número de actos delictivos a nivel nacional. Según la Policía Nacional del Perú, a través de la Dirección de Tecnologías de la Información y Comunicaciones (2023), se registraron 19% más delitos denunciados respecto al año anterior. Es preciso señalar que los ilícitos contra la propiedad patrimonial representan un 66.5% del conjunto de denuncias reportadas a nivel nacional. La Policía Nacional del Perú (2024), a través de su Dirección de Tecnología de la Información y Comunicaciones, reporta que, en el primer trimestre de 2024, se registraron 141960 delitos relacionados con el patrimonio y la seguridad pública,

representando más del 77% del total de delitos. A nivel nacional, en enero de 2024, las regiones que lideran el número de denuncias por extorsión son Lima y La Libertad con un 51.92% y 9,62%, respectivamente.

En La Libertad, específicamente en la ciudad de Trujillo, Aredo (2022) afirma que el hurto y el robo representan el 30.80% y 28.63% del total de los delitos cometidos, respectivamente. Asimismo, el distrito del porvenir, ubicado en la ciudad de Trujillo, acumula el mayor número de delitos perpetrados y el mayor índice de infractores de la ley con un 10.34%. Carranza (2022) menciona que, en el 2019, las denuncias contra el patrimonio en la provincia y el distrito de Trujillo equivalen al 77% y 43% del total de denuncias realizadas en el departamento de La Libertad, respectivamente; del mismo modo, el autor asegura que La Esperanza, como distrito, presenta la cifra más alta de delitos. De acuerdo con el Plan Distrital de Seguridad Ciudadana del distrito de El Porvenir del 2023, aproximadamente el 90% de los ciudadanos del distrito declara haber sido víctima de algún hecho delictivo, así como el hurto y el robo son los delitos con mayor frecuencia. En el distrito de Trujillo, muchas bodegas cierran sus puertas al promediar la noche, dado que no cuentan con los elementos de seguridad preventivos para protegerse del robo y la extorsión.

Según Al-Nawashi et al. (2017) y Aldayri y Albattah (2022), incluso cuando se implementan sistemas de vigilancia, las personas enfrentan desafíos para mantener una supervisión efectiva en los alrededores del negocio o patrimonio. La sobrecarga cognitiva es común, dado que los administradores del lugar no pueden concentrarse en identificar una actividad sospechosa mientras realizan las actividades empresariales cotidianas. Esta limitación reduce significativamente la eficacia de la seguridad y aumenta la vulnerabilidad de las bodegas ante robos y otros incidentes delictivos.

Como resultado de esta situación, Oriazowanlan y Erah (2019) afirman que los comerciantes redujeron las horas de trabajo y anularon la expansión de los negocios o la realización de nuevas inversiones. Es preciso señalar que la reducción de horas de trabajo, sin lugar a dudas, puede desencadenar en el cierre temporal o definitivo de muchos negocios, según Pinazo-Dallenbach (2020) Asimismo,

Nassauer (2018) asegura que las personas tienen miedo de perder la vida porque el 41.8% de los robos involucran el uso de un arma de fuego. En Estados Unidos, Singleton et al. (2022) afirman que los altos índices de violencia criminal en las bodegas incrementan la depresión y el abuso indiscriminado. A su vez, Connealy y Piza (2019) sostienen que las bodegas incrementan sus costos para colocar estructuras físicas que los protejan del ataque de los delincuentes. Esta situación afecta la estabilidad económica y la tranquilidad de los negocios. Por lo tanto, se deben tomar medidas, dado que las consecuencias podrían empeorar, afectando gravemente el desarrollo comercial y social de la ciudad.

Para abordar este problema, este trabajo tiene como objetivo desarrollar un modelo de visión computacional para funcionar con videos de cámaras de baja calidad, que son más accesibles y están dentro del presupuesto de las bodegas. Este modelo podría analizar las imágenes capturadas por estas cámaras para detectar actividad sospechosa, alertando al propietario o a los vigilantes cuando sea necesario. El uso de esta tecnología permitiría mejorar la seguridad en estos establecimientos, proporcionando una vigilancia preventiva sin depender de costosos equipos o de la intervención constante del ser humano.

1.2 Formulación del Problema

¿De qué modo un modelo de visión computacional contribuye con la detección de actividad sospechosa en Bodegas de Trujillo?

1.2.1 Problema General

¿En qué medida un modelo basado en YOLO11 y una Red de Memoria a Largo y Corto Plazo puede detectar actividades sospechosas en trayectorias de desplazamiento humanas en espacios delimitados de Bodegas de Trujillo?

1.2.2 Problemas Específicos

¿En qué medida una Red Neuronal Convolutiva como YOLO11 puede detectar trayectorias humanas en secuencias de video respecto al número total de personas detectadas?

¿En qué medida una Red de Memoria a Largo y Corto Plazo puede aprender patrones normales de movimiento a partir de trayectorias etiquetadas temporalmente, medida mediante la precisión predictiva usando el error cuadrático medio?

¿En qué medida una Red de Memoria a Largo y Corto Plazo puede generalizar adecuadamente los patrones normales de trayectorias, medida mediante la consistencia en la detección de anomalías en diferentes segmentos de datos?

1.3 Justificación Teórica

El desarrollo de esta investigación se fundamenta en la necesidad de abordar un desafío de relevancia tanto académica como práctica en el contexto de Trujillo, una ciudad en proceso de desarrollo en el Perú. La falta de técnicas y modelos de detección de actividad sospechosa adaptados a las limitaciones tecnológicas y económicas de las microempresas constituye un vacío en la literatura y la aplicación práctica. Según Lee et al. (2024), la adaptación de algoritmos de detección de anomalías a dispositivos de baja resolución y recursos limitados representa un área de investigación emergente. Esta investigación busca cubrir esta brecha mediante la propuesta y evaluación de un modelo de redes neuronales, contribuyendo a la expansión del conocimiento en el ámbito de la seguridad empresarial en entornos limitados.

1.4 Justificación Práctica

La falta de técnicas y modelos de detección de actividad sospechosa adaptados a imágenes de baja resolución presenta un obstáculo técnico significativo. La ausencia de características como el reconocimiento facial y el movimiento de las extremidades requieren un enfoque innovador. El desarrollo y evaluación de un modelo de redes neuronales específicamente diseñado para estas condiciones técnicas representa un desafío técnico valioso y relevante. Los resultados de esta investigación podrían no solo informar a la comunidad académica sobre enfoques prácticos y técnicos para abordar este problema, sino también brindar a las microempresas de Trujillo una herramienta efectiva para mejorar su seguridad y

protección en un entorno tecnológicamente limitado. Según Calikus et al. (2021), la adaptación de algoritmos a condiciones específicas es esencial para lograr resultados efectivos en la detección de actividades anómalas.

1.5 Objetivos de la Investigación

1.5.1 Objetivo General

Determinar en qué medida un modelo basado en YOLO11 y una Red de Memoria a Largo y Corto Plazo detecta actividades sospechosas en trayectorias de desplazamiento humanas en espacios delimitados.

1.5.2 Objetivos Específicos

- Determinar en qué medida la Red Neuronal Convolutiva YOLO11 detecta trayectorias humanas en secuencias de video, respecto al número total de personas detectadas.
- Determinar en qué medida la Red de Memoria a Largo y Corto Plazo aprende patrones normales a partir de trayectorias etiquetadas temporalmente, respecto a la precisión predictiva medida mediante el error cuadrático medio.
- Determinar en qué medida la Red de Memoria a Largo y Corto Plazo muestra consistencia en la detección de anomalías cuando se aplica a diferentes segmentos de datos normales, respecto a la desviación estándar en las tasas de detección.

CAPÍTULO 2: MARCO TEÓRICO

2.1 Marco filosófico o epistemológico de la investigación

Bunge (2002) define la epistemología como aquella disciplina filosófica encargada de analizar tanto el proceso investigativo científico como sus resultados cognitivos. En el desarrollo de este estudio, resulta fundamental establecer los fundamentos epistemológicos que sustentan el desarrollo del conocimiento en el ámbito de la visión por computadora enfocada en detectar actividades sospechosas, particularmente cuando se enfoca en la comprensión de patrones de comportamiento humano a través de tecnologías de procesamiento de imágenes.

La visión computacional, como disciplina científica, tiene sus orígenes epistemológicos en la convergencia de múltiples campos del conocimiento. Según Szeliski (2022), esta área emerge de la intersección entre las matemáticas aplicadas, la neurociencia, la computación y la psicología cognitiva, creando un marco interdisciplinario que busca comprender y replicar la capacidad humana de interpretación visual. La construcción del conocimiento en este campo ha evolucionado desde enfoques determinísticos basados en reglas hacia paradigmas probabilísticos inmersos en el aprendizaje automático, reflejando un cambio epistemológico fundamental en la forma de abordar la comprensión de datos visuales.

En relación a la detección de anomalías y actividades sospechosas, Chandola et al. (2009) sostienen que este campo representa un área de investigación que desafía los paradigmas tradicionales del conocimiento, dado que se basa en la detección de patrones que se desvían de la normalidad esperada. Esta aproximación epistemológica se fundamenta en el principio de que el conocimiento sobre lo anómalo puede construirse a partir del aprendizaje de patrones normales, donde las desviaciones significativas del comportamiento esperado, medidas a través de métricas como el error cuadrático medio, permiten identificar actividades sospechosas.

La evolución histórica de las redes neuronales artificiales, componente central de esta investigación, ha experimentado múltiples transformaciones paradigmáticas.

Según Sánchez (2024), McCulloch y Pitts establecieron en 1943 las bases conceptuales del procesamiento neuronal artificial en 1943; sin embargo, el estudio de las redes neuronales se popularizó con las técnicas de retropropagación desarrolladas por Bryson y Ho en 1969. Esto consolidó un nuevo paradigma epistemológico fundamentado en el aprendizaje de las máquinas para detectar representaciones complejas sin guía humana directa.

Las Redes de Memoria a Largo y Corto Plazo, según Hochreiter y Schmidhuber (1997), introducen una perspectiva epistemológica que reconoce la importancia de la temporalidad en el procesamiento de información. Este enfoque plantea que el conocimiento sobre patrones de comportamiento no puede ser completamente comprendido sin considerar las dependencias temporales y la capacidad de preservar información importante durante períodos de tiempo prolongados.

En el contexto de la detección de actividades sospechosas en espacios comerciales como bodegas, surge la necesidad de desarrollar un marco epistemológico que considere las limitaciones tecnológicas y económicas del entorno peruano. La adaptación de tecnologías avanzadas a contextos con recursos limitados requiere una aproximación epistemológica que valore tanto la efectividad técnica como la viabilidad práctica, reconociendo que el conocimiento científico debe ser contextualizable y aplicable en diferentes condiciones operativas.

Finalmente, es importante señalar que esta investigación contribuye al desarrollo epistemológico del área al proponer una metodología que integra técnicas de detección de objetos con análisis temporal de trayectorias, creando un marco conceptual que puede ser extrapolado a otros contextos similares. El conocimiento generado representa un aporte teórico y práctico al entendimiento de cómo los modelos de visión computacional pueden ser adaptados y optimizados para entornos con características específicas y limitaciones particulares.

2.2 Antecedentes del Problema

Se utilizará la técnica de Revisión Bibliográfica de Literatura definiendo para ello las preguntas de investigación que permitirán resolver el problema identificado.

Las preguntas de investigación son las siguientes:

Q1 ¿Qué arquitecturas y topologías de visión por computadora se usan para detectar actividad sospechosa en video?

Q2 ¿Qué técnicas existen para la detección de actividad sospechosa en video?

Q3 ¿Qué métricas son usadas para evaluar las técnicas de detección de actividad sospechosa en video?

Para la definición de la bibliografía a conseguir se determinan las siguientes cadenas de búsqueda.

Scopus

(("suspicious activity" OR "abnormal activity") AND ("detect" OR "classify" OR "detection" OR "classification")) AND ("neural network" OR "model" OR "deep learning" OR "machine learning")

Web of Science

(("suspicious activity" OR "abnormal activity") AND ("detect" OR "classify" OR "detection" OR "classification")) AND ("neural network" OR "model" OR "deep learning" OR "machine learning")

IEEE Xplore

(("suspicious activity" OR "abnormal activity") AND ("detect" OR "classify" OR "detection" OR "classification")) AND ("neural network" OR "model" OR "deep learning" OR "machine learning")

Tabla 1

Criterios de Inclusión y Exclusión de Artículos Científicos

Criterios de Inclusión	Criterios de Exclusión
Seis últimos años	Español
Artículos Científicos	Artículos de Revisión

Q1, Q2, Q3	Libros
------------	--------

Luego de obtener los artículos, se procedió a verificar si estos estaban indexados en alguna revista que figure en el indicador SCImago Journal Rank. Los artículos científicos deberían tener menos de 6 años de antigüedad; es decir, solamente los artículos publicados a partir de 2019. Del mismo modo, se tuvo en cuenta que estos artículos pertenezcan a los cuartiles Q1 y Q2 como prioridad; no obstante, no se descarta el uso de una revista Q3, debido a la poca información académica encontrada.

El análisis cuantitativo de la Tabla 2 revela 8 publicaciones implementando configuraciones convolucionales (N° 3, 6, 7, 11, 12, 13, 25, 20), mientras que 4 investigaciones emplean una variación de la redes neuronales recurrentes (N° 9, 12, 23, 29). El trabajo N° 27 incorpora redes generativas adversariales y el N° 10 desarrolla predicción multiescala de las trayectorias. Numéricamente, las CNN abarcan dos tercios del corpus examinado, seguidas por procesadores secuenciales memoria-dependientes. Particularmente, los estudios 12 y 23 yuxtaponen codificadores y decodificadores con capas convolucionales o recurrentes, configurando arquitecturas duales que combinan análisis espacial y morfológico con captación de patrones temporales progresivos. Esta composición refleja la preferencia investigativa por extraer características jerárquicas mediante kernels convolucionales antes que otros mecanismos de representación.

Tabla 2

Agrupación de artículos según diversos tipos de Arquitectura

N°	Descripción	Q1	Fuente	Responde
3	Coadyuva a la prevención delictiva mediante redes convolucionales tridimensionales en hurtos minoristas	Modelo de redes neuronales convolucionales tridimensionales (Conv3D)	Scopus	Arquitecturas
11	Contribuye a identificar conductas anómalas mediante redes convolucionales neuronales	Red Neuronal Convolutacional	Scopus	Arquitecturas
12	Facilita la detección de irregularidades en	CNN LSTM Bidireccional	Scopus	Arquitecturas

	videovigilancia vía códec H265 y aprendizaje profundo			
6	Potencia la identificación de anomalías en monitoreo visual mediante aprendizaje profundo	Red Neuronal Convolutiva Profunda	Scopus	Arquitecturas
13	Impulsa un modelo inteligente de vigilancia para anomalías mediante CNN-SEQ2SEQ radial	Arquitectura Convolutiva Secuencia a Secuencia	Scopus	Arquitecturas
10	Propicia predicción multiescala de trayectorias para detección de actividades humanas irregulares	Modelo Multiescala Temporal	IEEE Xplore	Arquitecturas
7	Auxilia en la detección de eventos anómalos mediante características espaciotemporales aprendidas	Codificador Convolutivo Embebido en una Red Recurrente	Scopus	Arquitecturas
9	Viabiliza sistema de vigilancia inteligente para conductas sospechosas vía aprendizaje automático	Red Neuronal Convolutiva / LSTM	Scopus	Arquitecturas
20	Promueve detección de anomalías en videovigilancia empleando SlowFast Resnet-50	SlowFast Resnet-50 CNN	Scopus	Arquitecturas
23	Fomenta detección no supervisada de anomalías mediante ConvLSTM variacional de doble flujo	LSTM de Doble Flujo y Variational Autoencoder	Scopus	Arquitecturas
25	Optimiza detección eficiente de conductas anómalas con LightAnomalyNet convolutiva	Arquitectura Convolutiva	Scopus	Arquitecturas
27	Robustece detección de anomalías en video mediante GAN modulada por ruido	Red Generativa Adversaria (GAN)	Scopus	Arquitecturas
29	Refuerza reconocimiento eficiente de anomalías con LSTM residual atencional en videovigilancia	LSTM	Scopus	Arquitecturas
30	Contribuye a detección de anomalías mediante generación adversaria retrospectiva bidireccional	Red Generativa Adversaria (GAN)	Scopus	Arquitecturas
36	Potencia detección espaciotemporal de	CNN y ConvLSTM	Scopus	Arquitecturas

	anomalías en videovigilancia en tiempo real			
42	Viabiliza marco convolucional profundo para detección de conductas anómalas en sistemas inteligentes	Red Neuronal Convolucional (estimación de poses) y LSTM (análisis de la dinámica del comportamiento)	Scopus	Arquitecturas
46	Facilita detección semi-supervisada de anomalías en entornos críticos RGB-D	CNN Bi-LSTM	WoS	Arquitecturas
47	Impulsa localización no supervisada de anomalías mediante red de traducción espaciotemporal profunda	Red Generativa Adversaria (GAN)	WoS	Arquitecturas

La segmentación de Técnicas en la Tabla 3 registra 13 publicaciones con dispersión técnica pronunciada. YOLOv5 aparece únicamente en trabajos N° 16 y 43 (emparejado con DeepSort en el segundo caso), constituyendo la exclusiva recurrencia algorítmica del corpus. Las restantes aproximaciones permanecen atomizadas: reconocimiento fundado en semántica (N° 5), mixturas gaussianas (N° 2), potenciador gradiente XG-Boost (N° 14), optimizador DragROA (N° 18), jerarquización multi-instancia (N° 19), analítica de video vigilancia (N° 33), propagación hacia adelante (N° 22), maximización del área bajo la curva (N° 35), vectorización soporte-descriptiva (N° 21) y detección postural específica (N° 39). Contrariamente a la concentración arquitectónica convolucional previamente cuantificada, este espectro técnico manifiesta fragmentación metodológica donde cada investigación articula soluciones algorítmicas diferenciadas. Subsecuentemente, YOLO emerge como único procedimiento replicado entre estudios, posicionándose cual estándar operativo para localizar elementos en ambientes monitoreados.

Tabla 3

Agrupación de artículos según diversos tipos de Técnicas

N°	Descripción	Q2	Fuente	Responde
5	Propicia asimilación semántica de actividades	Enfoque semántico para	IEEE Xplore	Algoritmos/Técnicas

	en sistemas de videovigilancia.	detectar actividades		
2	Viabiliza identificación de hurtos en entornos no restringidos mediante modelado atributivo de acciones.	GMM UAM	Scopus	Algoritmos/Técnicas
14	Facilita monitoreo autónomo de irregularidades en espacios públicos de supervisión.	Clasificador XG-Boost	Scopus	Algoritmos/Técnicas
16	Impulsa localización de anomalías visuales y conductuales en cámaras estáticas vía redes convolucionales.	YOLOv5	Scopus	Algoritmos/Técnicas
18	Optimiza detección de anomalías en video mediante red convolucional profunda habilitada para refinamiento.	DragROA	Scopus	Algoritmos/Técnicas
19	Contribuye a asimilar operaciones humanas anómalas mediante modelo de ordenamiento multi-instancia profundo.	Modelo Profundo de Agrupamiento Multiinstancias	Scopus	Algoritmos/Técnicas
33	Potencia análisis inteligente de video para eventos irregulares en supervisión en línea.	IVA	Scopus	Algoritmos/Técnicas
22	Robustece identificación eficiente de anomalías desde videovigilancia con red neuronal feed-forward bayesiana.	Red Profunda Hacia Adelante (FFNN)	Scopus	Algoritmos/Técnicas
35	Jerarquiza detección de conductas anómalas modelando comportamientos normales con maximización de AUC.	AUC	Scopus	Algoritmos/Técnicas
43	Propone enfoque inicial de inteligencia artificial para anticipar potenciales robos en cámaras interiores.	YOLOv5 y DeepSrot	Scopus	Algoritmos/Técnicas
21	Auxilia en detección de anomalías video mediante	Vector de Soporte para la	Scopus	Algoritmos/Técnicas

	descripción profunda de datos vector soporte.	Descripción de Datos		
39	Desarrolla sistema experto en tiempo real para asimilar conductas sospechosas en cajeros automáticos.	Sentadillas y Escaladas	Scopus	Algoritmos/Técnicas
31	Fomenta representación video y localización de eventos sospechosos mediante tecnologías semánticas.	Reglas de la Ontología OWL	Scopus	Algoritmos/Técnicas
40	Promueve detección de anomalías basada en búsqueda de vecinos cercanos con árbol B sensible a localidad.	Búsqueda del Vecino Más Cercano con un Locality-Sensitive B-tree	Scopus	Algoritmos/Técnicas
49	Anticipa conductas sospechosas para residencias seguras a partir de trayectorias de movimiento.	Detección de Paradas Cortas	WoS	Algoritmos/Técnicas
45	Viabiliza detección de actividades anómalas mediante autoencoders en regiones espaciotemporales paralelepípedas.	Codificador Automático	WoS	Algoritmos/Técnicas
48	Facilita aprendizaje en línea de contextos para identificar conductas sospechosas en videovigilancia.	Aprendizaje Incremental	WoS	Algoritmos/Técnicas

La caracterización métrica, con base en la Tabla 4, abarca 8 investigaciones con tres indicadores predominantes. Exactitud se registra en 4 trabajos (N° 4, 28, 26, 37), mientras que el Área Bajo la Curva (AUC) aparece en 3 estudios (N° 34, 24, 32), siendo este último combinado con la Tasa de Error Equivalente (EER). El artículo N° 44 incorpora simultáneamente AUC y EER, en tanto el N° 37 anexa Precisión junto con Exactitud. Numéricamente, la exactitud clasificatoria constituye el parámetro evaluativo mayoritario (50% del subconjunto), seguida por la curva ROC presente en 37.5% de las publicaciones. Solamente dos investigaciones (N° 32, 44) integran la tasa equitativa de error, revelando limitada adopción de métricas que tienen dependencia de un umbral. Consecuentemente, los autores privilegian

indicadores globales de desempeño antes que mediciones específicas de balance falsos positivos y negativos, sugiriendo preferencia por valoraciones holísticas sobre diagnósticos granulares del comportamiento clasificatorio.

Tabla 4

Agrupación de artículos según tipos de Métricas

Nº	Descripción	Q3	Fuente	Responde
4	Potencia asimilación de conductas irregulares en videovigilancia mediante aprendizaje profundo.	Accuracy	IEEE Xplore	Métricas
28	Examina LSTM atencional con agrupamiento variable para reconocer anomalías conductuales.	Accuracy	Scopus	Métricas
34	Viabiliza identificación de irregularidades en supervisión visual vía predicción bidireccional.	AUC	Scopus	Métricas
24	Referencia eventos atípicos en segmentos video empleando valores atípicos globales y locales.	AUC	Scopus	Métricas
26	Facilita identificación de anomalías en sistemas de monitoreo mediante transformada discreta coseno.	Accuracy	Scopus	Métricas
32	Optimiza detección en línea de irregularidades video con filtrado particulado.	Equal Error Rate (EER) y Area Under the Curve (AUC)	Scopus	Métricas
37	Rastrea entidades sospechosas entre múltiples cámaras mediante filtros correlativos.	Accuracy, Precision	Scopus	Métricas
44	Desarrolla redes espaciotemporales profundas para discernir comportamientos anómalos en secuencias visuales.	AUC y EER	WoS	Métricas

En esta sección se dará respuesta a las preguntas formuladas, considerando los resultados obtenidos por la cadena de búsqueda.

Q1 ¿Qué arquitecturas y topologías de visión por computadora se usan para la detección de actividad sospechosa en video?

Tabla 5

Arquitecturas de Visión Computacional para detectar Actividad Sospechosa en video

Nº	Arquitectura	Descripción	Referencias
1	CNN	Las CNN se utilizan ampliamente para procesar datos espaciales, como imágenes y videos. Estas redes identifican patrones visuales como bordes, formas y texturas mediante convoluciones, siendo útiles para detectar anomalías o comportamientos inusuales en secuencias visuales. Este tipo de red se especializa en analizar imágenes y videos, reconociendo patrones que podrían indicar algo fuera de lo normal, como un movimiento sospechoso.	20, 25, 3, 11, 6, 10
2	GAN	Las GAN consisten en dos redes neuronales que compiten entre sí: una generadora y otra discriminadora. Se utilizan para identificar anomalías mediante la comparación entre datos reales y datos generados artificialmente. La incapacidad de la generadora para replicar datos inusuales ayuda a destacar comportamientos atípicos. Estas redes funcionan como un juego entre dos inteligencias, donde una intenta crear ejemplos y la otra intenta identificar si son reales o falsos, permitiendo detectar cosas que no encajan.	27, 47
3	VAE LSTM	Combinan la capacidad de los VAE para comprimir y reconstruir datos con la habilidad de las LSTM para manejar secuencias temporales. Esto permite analizar cómo evolucionan los patrones en el tiempo y detectar cambios inusuales. Estas redes trabajan comprimiendo y estudiando cambios en el tiempo, lo que permite identificar comportamientos fuera de lo común en secuencias, como movimientos anormales.	23

4	CNN LSTM	Esta combinación utiliza CNN para extraer características espaciales de imágenes o videos y LSTM para analizar cómo cambian estas características en el tiempo. Es ideal para detectar patrones sospechosos en secuencias visuales dinámicas. Estas redes primero miran los detalles en las imágenes y luego observan cómo cambian con el tiempo, ayudando a identificar algo extraño que ocurre en una secuencia de video.	7, 36, 29, 42, 12, 46, 13, 9
5	CNN GAN	Se integran CNN para analizar datos espaciales y GAN para generar y validar patrones normales. Ayuda a identificar datos que no se ajustan a los patrones esperados en imágenes o videos. Esta combinación detecta patrones en imágenes y los compara con ejemplos generados, destacando cualquier cosa que parezca fuera de lugar.	30

Análisis Crítico

Joshi y Chaudhari (2022) presentaron una solución para detectar eventos o actividades anormales utilizando la Arquitectura de Redes Convolucionales SlowFast ResNet 50. Este modelo permite procesar características de los frames de video para capturar información temporal rápida y lenta. La arquitectura finaliza con una función Softmax que clasifica las actividades detectadas. Según los resultados, este enfoque genera un 47.8% más de exactitud en comparación con otros algoritmos. En relación con el presente proyecto, el uso de redes neuronales convolucionales para detectar robos y hurtos en videos guarda similitudes con el enfoque de los autores. Sin embargo, el objetivo específico se centra en detectar comportamientos sospechosos antes de cometer un delito en los alrededores de microempresas. La arquitectura SlowFast ResNet 50 podría adaptarse para analizar patrones específicos, optimizando los recursos planteados en la investigación.

Mehmood (2021) desarrolló LightAnomalyNet, una Red Neuronal Convolucional, diseñada para detectar eventos normales y anormales utilizando imágenes en escala de grises de tres canales (SG3I) como entrada. Este modelo no es una 3D CNN, lo cual reduce su carga computacional y el uso de memoria. Los resultados indican una exactitud mayor al 90% comparado con otros métodos. En relación con el presente trabajo de investigación, este enfoque resulta relevante para la

detección de comportamientos sospechosos y actos violentos, dado que ofrece una solución eficiente para escenarios con recursos tecnológicos limitados.

Martínez-Mascorro et al. (2020) propusieron una arquitectura basada en Redes Neuronales Convolucionales en 3 dimensiones para detectar comportamientos sospechosos que podrían derivar en robos a tiendas. El enfoque distingue cuatro momentos clave: Momento Exacto del Crimen (SCM), Momento Comprensivo del Crimen (CCM), Lapso del Crimen (CL) y Conducta Previa Criminal (PCB), centrándose en este último para analizar el comportamiento del sospechoso antes del acto delictivo. Este análisis segmenta el video en etapas que permiten identificar patrones previos al crimen sin incluir evidencia explícita del delito, evitando sesgos en el entrenamiento del modelo. El modelo logró una exactitud del 70%, identificando correctamente el 75% de los casos. Este enfoque, en relación al presente trabajo de investigación, es relevante para prevenir robos mediante el análisis del comportamiento previo, aportando una metodología basada en evidencia para detectar señales tempranas de actividad sospechosa en entornos comerciales.

Quadri y Katakdhond (2022) desarrollaron una Red Neuronal Convolucional para abordar la impredecibilidad del comportamiento sospechoso en comportamientos normales. La solución propuesta realiza tareas como la segmentación del video, la extracción del fondo y la extracción de los objetos del primer plano, con el objetivo de rastrear el movimiento de la persona, detectar el movimiento de la cabeza y analizar los diversos contactos en el entorno. Este enfoque se relaciona con la presente investigación al emplear una CNN para identificar patrones anormales. Sin embargo, no se abordarán comportamientos como movimientos de cabeza, cambios de lugar entre individuos, abandono de una posición o contacto entre personas, dado que el análisis se centra en otros tipos de comportamientos más vinculados al contexto delictivo.

Kavikuil y Amudha (2019) abordaron el problema de la detección de anomalías en videos de vigilancia utilizando una Red Neuronal Convolucional Profunda (DCNN). El modelo aprende características espaciales de videos etiquetados débilmente, a través de múltiples capas consecutivas, capas de agrupamiento submuestreo y capas densas. Durante el entrenamiento, las características se extraen de pequeñas

secciones del frame para consolidar un único mapa de características, que es procesado por las capas de agrupamiento y conectado a las capas densas. La arquitectura incluye funciones de activación no lineales (ReLU) y una capa Softmax para clasificar los videos de entrada. La precisión del modelo fue de 0.99, una sensibilidad de 0.99, una puntuación F1 de 0.99 y una exactitud de 0.96. El modelo es relevante porque permite detectar comportamientos como el merodeo y el lanzamiento de objetos. Aunque este enfoque incluye otras anomalías, la capacidad de detectar merodeo lo hace aplicable a investigaciones centradas en la prevención de actividades delictivas.

Rodrigues et al. (2020) elaboraron un modelo multiescala basado en convoluciones unidimensionales para procesar trayectorias de poses humanas como entrada. El marco incluye dos modelos principales: uno para predicciones del pasado y otro para predicciones del futuro. Cada modelo divide las secuencias de entrada en subsecuencias, realiza predicciones en múltiples escalas de tiempo mediante capas intermedias y combina estas predicciones para formar una salida completa. Durante este proceso, el modelo minimiza de manera conjunta la pérdida de nodos y capas en cada nivel intermedio. El modelo alcanzó un Área Bajo la Curva (AUC) entre 67% y 82%, mostrando su capacidad para detectar comportamientos como correr, pelear, merodear (*loitering*), saltar, lanzar un objeto, caminar cerca de la cámara y bailar. Sin embargo, el uso exclusivo de trayectorias de poses humanas como entrada limita su aplicabilidad en esta investigación, que busca analizar comportamientos sospechosos mediante un enfoque más general.

La Figura 1 presenta una arquitectura CNN generalizada derivada del análisis de seis estudios que emplean redes convolucionales para la detección de actividades sospechosas. La arquitectura sigue un patrón secuencial común: la capa de entrada procesa frames de video o secuencias de características, seguida de bloques convolucionales repetidos que integran capas de convolución, activación ReLU y pooling para la extracción jerárquica de características. Posteriormente, la vectorización transforma los mapas de características multidimensionales en un vector unidimensional que alimenta las capas completamente conectadas, culminando en una capa de salida que clasifica los comportamientos como normales o anormales. Las variantes identificadas se diferencian principalmente en la

dimensionalidad de las convoluciones: 1D para secuencias temporales, 2D para frames individuales, 3D para volúmenes espacio-temporales.

Figura 1

Arquitectura CNN Generalizada para Detección de Actividades Sospechosas



Chen et al. (2021) propusieron una Red Generativa Adversarial denominada NM-GAN para el reconocimiento de anomalías en material videográfico. Este modelo combina una red de reconstrucción basada en codificador-decodificador y una red de discriminación basada en CNN en una arquitectura adversarial. La red de reconstrucción genera mapas de error a partir de los frames de video, que se comparan con una distribución normal predefinida. Los mapas de error resultantes son procesados por la red de discriminación, que clasifica las muestras como normales o anómalas en función de su desvío de la distribución esperada. Este enfoque, que incluye aprendizaje adversarial modulado por ruido, mejora el poder de generalización de la arquitectura y optimiza la detección de anomalías. El modelo logró un AUC superior al 85% y un EER entre 5% y 17%, dependiendo del dataset utilizado. Entre las anomalías detectadas se incluyen comportamientos como lanzamiento de objetos, violencia y merodeo. Este último comportamiento se

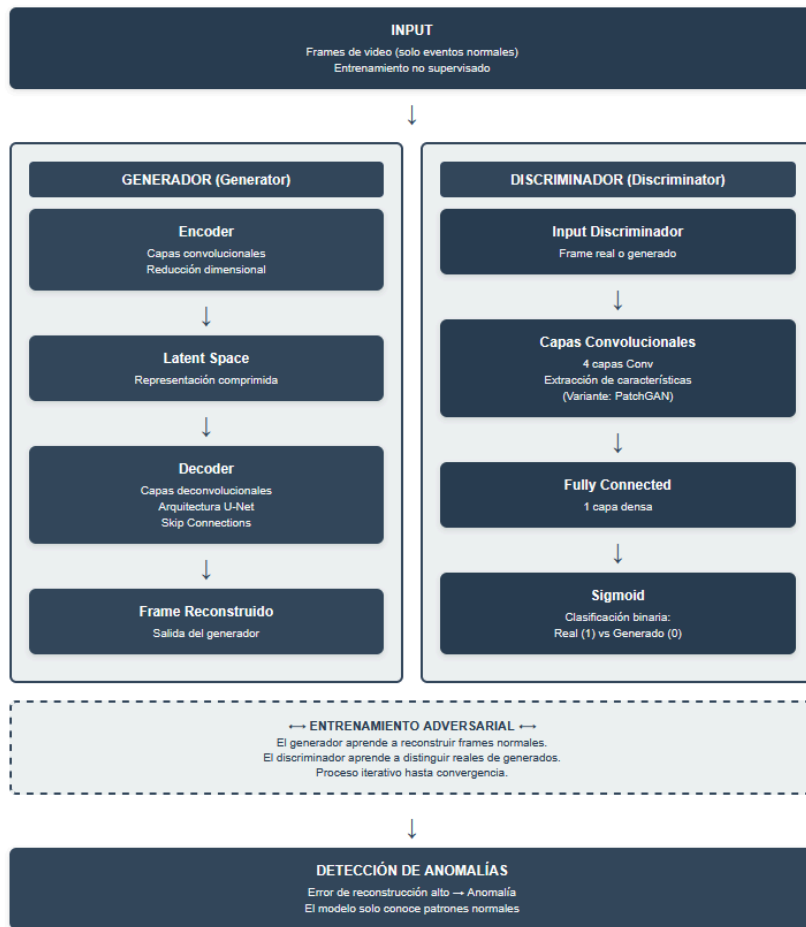
clasifica como una actividad sospechosa previa a actividades delictivas, siendo relevante para investigaciones enfocadas en la prevención del crimen.

Ganokratanaa, Aramvith y Sebe (2020) desarrollaron una Red de Traducción Espaciotemporal Profunda, basada en Redes Generativas Antagónicas y Envolvimiento de Bordes para la detección y localización de anomalías en escenas aglomeradas y complejas. El modelo se entrena exclusivamente con frames de eventos normales, generando su flujo óptico denso como características temporales. Durante las pruebas, los eventos desconocidos se clasifican como anómalos, dado que el modelo solo conoce los eventos normales con los que fue entrenado. El enfoque introduce una fusión novedosa de eliminación de fondo y flujo óptico real para reducir el ruido, la concatenación de frames originales con los de fondo eliminado, y el uso de Envolvimiento de Bordes para minimizar bordes no relacionados, mejorando así la localización de anomalías. El modelo alcanzó un AUC superior al 77% tanto a nivel de frame como de píxel, con un EER entre 5% y 30%, dependiendo del dataset. El modelo puede detectar anomalías como el lanzamiento de objetos, los saltos y el merodeo, siendo este último relevante para la identificación de patrones sospechosos en escenarios de vigilancia.

La Figura 2 presenta una arquitectura GAN (Red Generativa Adversarial) generalizada para la detección de actividades sospechosas. A diferencia de las arquitecturas CNN que requieren aprendizaje supervisado, las GAN emplean un enfoque no supervisado entrenando exclusivamente con frames de eventos normales. La arquitectura se compone de dos redes que operan de manera adversarial: el generador, basado en U-Net con skip connections, aprende a reconstruir frames normales mediante un proceso encoder-decoder; y el discriminador, conformado por cuatro capas convolucionales, una capa fully connected y una función sigmoidea, clasifica frames como reales o generados. Durante el entrenamiento, ambas redes compiten iterativamente hasta la convergencia. En la detección, frames anómalos generan errores de reconstrucción elevados dado que el generador sólo conoce patrones normales, permitiendo así la identificación de comportamientos sospechosos.

Figura 2

Arquitectura GAN Generalizada para Detección de Actividades Sospechosas



Wang et al. (2022) propusieron el modelo DoubleFlow-ConvLSTM-VAE para abordar las limitaciones en la detección de anomalías de los modelos basados en Codificadores Variacionales, dado que este último no pueden procesar datos de series temporales. Este modelo combina Redes Long Short-Term Memory Convolucionales y Codificadores Variacionales en un esquema asimétrico que emplea la probabilidad de error de reconstrucción en lugar de la probabilidad de reconstrucción estándar para detectar anomalías. El enfoque captura de manera efectiva las características espacio-temporales en escenarios con multitudes. Los experimentos en datasets públicos mostraron un Área Bajo la Curva superior al 88% y una Tasa de Error Equivalente entre 12% y 19%. El modelo utiliza convoluciones y Redes LSTM para analizar series de tiempo; no obstante, las anomalías detectadas están enfocadas en contextos con multitudes, lo que difiere del enfoque del presente trabajo de investigación.

Zhang, Wang y Ding (2018) propusieron un modelo orientado a detectar eventos anormales en material videográfico, mediante un Autocodificador Convolutivo y una Red de Memoria a Largo y Corto Plazo. La arquitectura consta de tres etapas: preprocesamiento de datos, aprendizaje de características y detección de eventos anormales. Los datos de video se preprocesan para formar volúmenes con una longitud de tiempo fija como input del modelo. El Autocodificador Convolutivo captura características espaciales mediante convoluciones, mientras que la red LSTM aprende características temporales. La detección de eventos anormales emplea el error de restauración normalizado, donde frames con altos errores de reconstrucción son clasificados como anómalos. Los experimentos reportaron un Área Bajo la Curva que oscila entre 84% y 89%, y una Tasa de Error Equivalente entre 20% y 22%. El modelo detecta patrones como correr, lanzamiento de objetos, merodeo, caminar en dirección opuesta y patrones de movimiento peatonal. De estos, los comportamientos de merodeo (loitering) y caminar en dirección opuesta son relevantes para el análisis en esta investigación.

Nawaratne et al. (2020) plantearon el Incremental Spatio-Temporal Learner (ISTL) para el reconocimiento y localización de anomalías en videos de forma instantánea. Este modelo consta de tres componentes: a) **Aprendizaje Espacio-Temporal**: Un Auto-codificador Espacio-Temporal conformado por capas convolucionales (CNN) para aprender características espaciales y capas Long Short-Term Memory Convolucionales (ConvLSTM) para modelar patrones temporales. b) **Detección y Localización de Anomalías**: El modelo calcula el error de reconstrucción entre los cuboides originales y reconstruidos para identificar anomalías. Si el error supera un umbral de anomalía predefinido, el segmento se clasifica como anómalo. Asimismo, la localización de la anomalía se realiza a nivel de píxel. c) **Aprendizaje Activo con Agregación Difusa**: Las anomalías detectadas son validadas por un observador humano, quien puede ajustar el modelo al marcar detecciones erróneas como normales. Este feedback, junto con un sistema de agregación difusa, permite al modelo adaptarse a nuevos comportamientos normales. Los resultados mostraron un Área Bajo la Curva entre 75% y 91%, y una Tasa de Error Igualada entre 8% y 29%. Este modelo detecta patrones como merodeo (loitering), abandono de objetos y caminar hacia la cámara, siendo relevantes para esta investigación.

Ullah et al. (2021) formularon una arquitectura para el reconocimiento y detección de actividades anómalas en videos de vigilancia, compuesta por tres fases principales. En la primera fase, una Red Neuronal Convolucional ligera y preentrenada extrae características espaciales de una serie de frames de video. En la segunda fase, se genera un vector a partir de una secuencia de 30 frames, encapsulando las características aprendidas. Finalmente, en la tercera fase, este vector se utiliza como entrada en una Red Long Short-Term Memory Residual con Atención destinada a capturar las dependencias temporales en la secuencia. El modelo aumentó la exactitud (accuracy) de las detecciones entre 0.7% y 8.6% en comparación con otros modelos, con un *recall* de 78% a 98%, una precisión de 87% a 99%, un puntaje F1 de 81% a 99% con un Área Bajo la Curva (AUC) superior al 96%. La propuesta detecta comportamientos delictivos como agresiones, robos y hurtos en tiendas, incrementando su relevancia en investigaciones relacionadas con contextos delictivos.

Ko y Sim (2018) presentaron un modelo para la detección de comportamientos humanos anormales compuesto por tres módulos. Primero, el módulo de detección y discriminación de sujetos humanos utiliza una red YOLO v2 combinada con un estimador basado en el filtro de Kalman para detectar en tiempo real la posición y clase de cada sujeto. Segundo, el módulo de clasificación de posturas utiliza una Red Neuronal Convolucional llamada p-CNN, basada en el modelo VGG-16, para clasificar seis clases de posturas anómalas durante el aprendizaje. Finalmente, el módulo de detección de comportamientos anormales utiliza una Arquitectura Recurrente con Memoria a Largo Plazo para modelar las dinámicas del movimiento y reconocer secuencias de acciones dependientes en el tiempo. Los resultados mostraron un *recall* entre 0.52 y 0.89 y una precisión entre 0.65 y 0.78 en la detección de comportamientos con interacciones como golpear y empujar. Sin embargo, el enfoque orientado al monitoreo de personas adultas difiere del contexto del presente proyecto.

Abbas y Al-Ani (2022) desarrollaron un modelo orientado a detectar anomalías en grabaciones de seguridad, compuesto por cuatro fases. Primero, se aplica una codificación de video de alto rendimiento H265 para reducir el tamaño de los datos. Luego, se utilizan Redes Neuronales Convolucionales basadas en ResNet50 para

extraer características espacio-temporales de los frames. Estas características se combinan cada 15 frames para generar un vector de entrada que optimiza la representación de las características aprendidas. Finalmente, se emplea un modelo Long Short-Term Memory Bidireccional para clasificar los videos como normales o anómalos. El modelo logró un Área Bajo la Curva (*AUC*) del 90.16% y una exactitud (*accuracy*) del 79.81%, demostrando su capacidad para detectar incidentes como peleas, explosiones, abusos y accidentes. Sin embargo, al estar diseñado para reaccionar a incidentes peligrosos que están ocurriendo, carece del enfoque preventivo que caracteriza la presente investigación.

Khaire y Kumar (2022) expusieron un modelo conformado por Redes Neuronales Convolucionales (*CNN*, por sus siglas en inglés) y Redes Long Short-Term Memory Bidireccionales (*Bi-LSTM*, por sus siglas en inglés) en un esquema codificador-decodificador. La arquitectura se divide en dos partes: un módulo de extracción de características de video basado en la Red Neuronal Convolutiva MobileNet; y, un módulo Bi-LSTM codificador-decodificador que aprende los patrones de las secuencias normales a partir de las características concatenadas. Este enfoque semi-supervisado requiere solo videos etiquetados débilmente como normales para el entrenamiento. El modelo logró una exactitud (*accuracy*) entre 89% y 91%, una precisión de 0.990, un *recall* de 0.908 y una medida F (*F-measure*) de 0.947. Estas métricas destacan su eficacia en la detección de comportamientos como transacciones prolongadas, caídas de tarjetas y gestos como rascarse la cabeza en cajeros automáticos. El enfoque del modelo difiere parcialmente del alcance preventivo de la presente investigación. Sin embargo, aporta técnicas útiles como la detección de transacciones prolongadas.

Pawar y Attar (2020) propusieron dos modelos capaz de reconocer anomalías en grabaciones continuas. El primer modelo, denominado CNN-SEQ2SEQ-RBF, combina un Auto-codificador Convolutivo Bidimensional para el aprendizaje de características espaciales, un modelo de Secuencia a Secuencia Long Short-Term Memory (SEQ2SEQ-LSTM, por sus siglas en inglés) para identificar correlaciones estadísticas temporales y una Función de Base Radial para la clasificación en una sola clase. El segundo modelo, llamado CNN-RBF, es una versión simplificada que se enfoca únicamente en las características espaciales. Ambos modelos utilizan una

distribución Gaussiana multivariada para calcular la probabilidad de anomalías. Los resultados mostraron que CNN-RBF superó a CNN-SEQ2SEQ-RBF, con una Tasa de Error Igualada (EER, por sus siglas en inglés) de 14% frente al 18% del primer modelo. Los modelos detectan objetos como carros pequeños, bicicletas y patinadores, pero difiere del presente proyecto porque no modela comportamientos humanos.

Joshi et al. (2020) lograron detectar actividades y comportamientos sospechosos en videos, utilizando una arquitectura que combina la Arquitectura Convolutiva y la Red de Memoria a Largo y Corto Plazo Convolutiva. La propuesta consta de dos componentes: un extractor de características espaciales basado en CNN para analizar cada frame individual y un modelo Codificador-Decodificador que utiliza ConvLSTM para capturar patrones temporales en la secuencia de video. Este Codificador-Decodificador genera una representación comprimida de los frames, reconstruyendo luego los datos originales. La detección de anomalías se realiza midiendo el error de reconstrucción: un error elevado indica un comportamiento anómalo. El modelo detecta comportamientos como dejar objetos atrás, correr, merodear (*loitering*), peleas, caminar en direcciones incorrectas y movimientos repentinos; sin embargo, está diseñado para implementarse en escuelas y universidades, lo que difiere del enfoque de esta investigación.

La Figura 3 presenta una arquitectura CNN-LSTM generalizada para la detección de actividades sospechosas. Esta arquitectura combina capas convolucionales para extraer características espaciales y espacio temporales de frames (estructura de objetos, posición, apariencia) con variantes de LSTM (LSTM básico, ConvLSTM, BiLSTM, o con atención residual) para capturar dependencias temporales. Se identifican dos enfoques principales: (1) Autoencoder espaciotemporal con arquitectura encoder-decoder que reconstruye frames y detecta anomalías mediante error de reconstrucción, típicamente entrenado de forma no supervisada con solo datos normales; y (2) Clasificación directa mediante CNN+LSTM+FC+Softmax que distingue entre comportamiento normal y anormal, típicamente con aprendizaje supervisado usando ambas clases. Las variantes CNN incluyen ResNet-50 (50 capas que extraen 1000 características), VGG-16 (19 capas profundas), y MobileNetV2. Es importante destacar que estos modelos hallados en la literatura

procesan frames completos o secuencias de frames, no trayectorias explícitas (coordenadas x, y, t) de objetos individuales.

Figura 3

Arquitectura CNN-LSTM Generalizada para Detección de Actividades Sospechosas



Yang et al. (2021) propusieron el modelo Generativo Antagónico Retrospectivo Bidireccional capaz de reconocer anomalías en videos de seguridad. Este modelo consta de tres módulos principales: **a)** un generador para realizar predicciones bidireccionales, hacia adelante y hacia atrás, añadiendo frames predichos a la secuencia de entrada para una predicción retrospectiva que verifica la consistencia temporal; **b)** un discriminador de frames con el objetivo de evaluar si las imágenes generadas son reales o falsas, con el fin de minimizar la producción de imágenes falsas; y, **c)** un discriminador de secuencias basado en Redes Neuronales Convolucionales en tres dimensiones para promover la consistencia temporal al identificar si las secuencias contienen frames falsos. El modelo alcanzó un Área Bajo la Curva (AUC) de 74.5% y 88.6% y una Tasa de Error Igualada (EER) entre 19% y 31.6%, según los conjuntos de datos utilizados. El modelo incluye la

detección de merodeo (*loitering*), un comportamiento relevante en el presente proyecto de investigación.

Q2 ¿Qué técnicas existen para la detección de actividad sospechosa en video?

Tabla 6

Técnicas de Visión Computacional para detectar Actividad Sospechosa en video

Nº	Técnica	Descripción	Referencias
1	Probabilístico	Esta técnica se basa en modelos matemáticos que emplean probabilidades para estimar la ocurrencia de eventos o patrones específicos. En el contexto de detección de anomalías o actividades sospechosas, los algoritmos probabilísticos utilizan distribuciones de probabilidad, estadísticas bayesianas, modelos de mezcla (como GMM) y técnicas de estimación para modelar comportamientos normales y detectar desviaciones.	14, 33, 22, 2, 21, 40
2	Optimización	En esta categoría, los métodos buscan soluciones óptimas para problemas específicos mediante algoritmos que minimizan o maximizan funciones objetivo. Estos algoritmos suelen utilizar técnicas como gradiente descendente, algoritmos genéticos o enjambre de partículas para resolver problemas de clasificación, predicción o agrupamiento en el contexto de aprendizaje automático.	18, 35
3	Aprendizaje Supervisado	Esta técnica implica entrenar un modelo con datos etiquetados donde las relaciones entre las características de entrada y las salidas esperadas están definidas. Los modelos aprenden a generalizar patrones para clasificar o predecir nuevos datos. Es ampliamente utilizado para tareas de clasificación, detección y predicción.	19, 48
4	Reglas Lógicas	Basadas en conjuntos de reglas explícitas y definidas por expertos, estas técnicas identifican eventos sospechosos o anomalías al cumplir condiciones específicas. Son útiles para sistemas que requieren alta interpretabilidad y claridad en las decisiones tomadas.	43, 39

5	Reglas Semánticas	Estas técnicas emplean conceptos y definiciones semánticas para entender el contexto y clasificar eventos como normales o sospechosos. Incluyen análisis de comportamiento humano y movimientos según su significado en un contexto específico.	31, 5
6	Análisis de Trayectorias	Este enfoque analiza los movimientos y trayectorias de objetos o personas en escenas, buscando patrones anómalos. Se basa en el seguimiento de movimientos, detección de cambios bruscos o interacciones sospechosas entre individuos, y es ampliamente usado en sistemas de vigilancia.	16, 49, 45

Análisis Crítico

Alarfaj et al. (2023) propusieron un algoritmo basado en múltiples técnicas para la clasificación de actividades sospechosas o anormales. Este algoritmo se divide en: **A)** un Modelo de Mezcla Gaussiana para el reconocimiento automático de objetos y un Filtro de Kalman destinado al seguimiento de estos elementos. **B)** Luego, se aplica una eliminación de sombras para segmentar un objeto y su sombra. **C)** Para manejar las oclusiones, se implementa un algoritmo de detección de oclusiones entre múltiples siluetas humanas y utiliza reducción de regiones (*region shrinking*) para aislar las siluetas ocluidas. **D)** Posteriormente, se extraen la velocidad, la rapidez y el flujo óptico de cada silueta identificada para mejorar la exactitud del modelo. Estas características se verifican utilizando el algoritmo de agrupamiento Fuzzy C-mean junto con la Transformada de Hough (*Hough Transform*). **E)** Un optimizador (*Gray Wolf Optimizer*, GWO) se emplea para optimizar las características seleccionadas. **Finalmente**, un clasificador XG Boost combinado con un descriptor local detecta si un evento es anormal. El enfoque logró una exactitud (*accuracy*) entre 76% y 90%, una precisión entre 0.76 y 0.9, un *recall* entre 0.86 y 0.97, un puntaje F (*F Score*) entre 0.89 y 1, un Área Bajo la Curva (*AUC*) entre 0.769 y 0.906, y una Tasa de Error Igualada (*EER*) entre 0.16 y 0.427. El algoritmo detecta eventos como caer, correr y pelear. Aunque el enfoque es efectivo para detectar eventos anormales, su naturaleza post-incidente lo diferencia de la presente investigación.

Balasundaram y Chellappan (2020) propusieron un algoritmo de análisis inteligente de video (*Intelligent Video Analytics Model*) o Detección de Objetos y Humanos (*Human Object Detection*) orientado a detectar objetos y actividades humanas anormales en secuencias visuales de vigilancia. Primero, las cámaras de vigilancia transmiten los datos de video a un servidor en tiempo real para segmentarlos en frames. Estos frames se preprocesan, aplicando sustracción de fondo y operaciones morfológicas para detectar objetos. Los objetos detectados se comparan con objetos previamente definidos (*ground truth*) para identificar anomalías. Cuando se detectan objetos anómalos, se etiquetan y se envían alertas al servidor o al personal de seguridad. El modelo alcanzó niveles de error del 8% a nivel de frame y del 11% a nivel de píxel, con una exactitud (*accuracy*) de entre 98% y 99%. El algoritmo se limita a la identificación de objetos como silla de ruedas y autos, evitando la detección de comportamientos sospechosos en un contexto más preventivo.

Murugesan y Thilagamani (2022) desarrollaron un algoritmo orientado a detectar anomalías en videos de monitoreo, compuesto por cuatro fases principales. **a)** En el Modelado de Fondo (*Background Modelling*), se separan las regiones de fondo y primer plano mediante un filtro Laplaciano, eliminando regiones pequeñas con menos de 10 píxeles para reducir el ruido. **b)** En la Segmentación y Seguimiento de Objetos (*Object Segmentation and Tracking*), se utiliza agrupamiento basado en Fuzzy C-means. **c)** En la Extracción de Características (*Feature Extraction*), se extraen atributos relevantes de los frames de video. **d)** Finalmente, una Red Neuronal Bayesian Feedforward (*Bayesian Feedforward Neural Network*) clasifica los frames con alta actividad, permitiendo identificar y enviar alertas sobre posibles amenazas. El algoritmo logró una sensibilidad del 98.69%, una especificidad del 97.01%, una exactitud del 99.65% y un puntaje F1 de 0.98. Aunque detecta actividades como la presencia de no-caminantes, no es adecuado para la clasificación de acciones humanas, lo que limita su utilidad en la presente investigación.

Roy y Mohan (2018) desarrollaron el Modelo de Atributos Universal (*Universal Attribute Model*, UAM) y el análisis factorial destinado a detectar robos por arrebato en videos de seguridad. El UAM utiliza una Mezcla Gaussiana (*Gaussian Mixture Model*, GMM) para capturar atributos generales de acciones humanas a partir de

grandes datasets como UCF101 y HMDB51. A) Cada clip se ajusta mediante Máxima a Posteriori (*Maximum A Posteriori*, MAP) para generar vectores de acción super representativos (*super-action vectors*, SAV), que luego se refinan mediante análisis factorial. B) Este proceso descompone el SAV en una matriz de variabilidad total (*Total Variability Matrix*, T) y un vector de acción compacto que captura los atributos relevantes. C) La optimización de los parámetros se realiza con el algoritmo de Expectación-Maximización (*EM*). El modelo alcanzó una exactitud (*accuracy*) y un Área Bajo la Curva (*AUC*) del 99.8%, demostrando su efectividad en la detección de robos por arrebato. Sin embargo, este delito, al ser una suma de varias acciones, no se alinea con el enfoque preventivo del presente proyecto.

Wang et al. (2022) detectan anomalías en videos a través del modelo de Descripción de Datos basado en Máquina de Vectores de Soporte Profunda, una extensión de los modelos SVDD que utiliza redes neuronales profundas para aprender representaciones más precisas. Este enfoque mapea las muestras normales en una hiperesfera de volumen mínimo, donde el centro representa el comportamiento normal y el radio establece los límites de tolerancia. El modelo DSVDD opera en dos fases. En el entrenamiento, procesa imágenes RGB y diagramas de flujo óptico, fusionándose en un conjunto de datos de seis canales. Estos datos alimentan al modelo, que proyecta las muestras normales en un espacio hiperesférico a través de una red neuronal profunda. Este espacio minimiza el radio de la hiperesfera y penaliza las muestras fuera de los límites, optimizando los pesos de la red y los parámetros de la hiperesfera de forma alternada. El modelo alcanzó un área bajo la curva (*AUC*, por sus siglas en inglés) de 73.2% y 86.8%. Este enfoque detecta la acción de merodear (*loitering*) y lanzar objetos, aspectos relevantes para este proyecto. Sin embargo, el modelo no emplea técnicas preentrenadas para reducir significativamente el tiempo de procesamiento.

Shen et al. (2018) desarrollaron un algoritmo, combinando características espacio temporales y dinámicas a través de un modelo basado en bosques LSB (*LSB-forest*). Este enfoque se centra en extraer características como HOF (Histogramas de Flujo Óptico), HOG (Histogramas de Gradientes Orientados) y la característica DoPB (Dinámica del Comportamiento del Peatón), que analiza los cambios dinámicos en el movimiento y apariencia de los peatones a lo largo del

tiempo. Las características extraídas se proyectan en un manifold de Riemann para evaluar las diferencias en patrones de movimiento regulares y anómalos, utilizando distancias geodésicas para mejorar la discriminación. En el proceso de entrenamiento, se construyen dos bosques LSB independientes, uno para las características HOF-HOG y otro para DoPB, usando únicamente datos normales. Durante la fase de prueba, los grados de anomalía se calculan buscando los vecinos más cercanos en los bosques, donde una mayor distancia respecto a los patrones normales indica un mayor grado de anomalía. Los mapas de anomalías generados para cada cuadro del video combinan las características mencionadas, y un umbral se utiliza para clasificar píxeles como normales o anómalos. Además, el algoritmo ajusta dinámicamente los grados de anomalía según el comportamiento histórico del objeto en análisis, permitiendo un seguimiento continuo. Los resultados obtenidos muestran un AUC entre 0.796 y 0.920, y un EER entre 0.300 y 0.150. Aunque el modelo detecta comportamientos como dirección incorrecta, merodeo (loitering) e interacciones irregulares, se diferencia del presente proyecto al emplear algoritmos de machine learning y no redes neuronales profundas.

Philip et al. (2022) desarrollaron un algoritmo basado en redes neuronales de convolución profundas optimizadas con la perspectiva híbrida DragROA (Dragonfly-Rider Optimization Algorithm) para detección de anomalías en videos. El proceso comienza con la extracción de características espacio temporales utilizando descriptores de histograma de flujo óptico de orientación y magnitud (HOF y HOM) para capturar patrones normales. Estas características son procesadas por una CNN entrenada con DragROA, un algoritmo que combina elementos de los métodos de optimización Dragonfly y Rider. DragROA ajusta el modelo simulando un conjunto de agentes (representados como libélulas), que se mueven y buscan soluciones óptimas en el espacio de búsqueda, basándose en factores como cohesión, separación, alineación, atracción hacia objetivos y evitación de distracciones. Además, el algoritmo incluye un componente de vuelo aleatorio (*Levy flight*) para explorar áreas no visitadas y evitar soluciones locales subóptimas. El algoritmo logró una exactitud de 0.9922, sensibilidad de 0.9809 y especificidad de 1. Sin embargo, su nexo con el estudio actual es limitado, dado que se enfoca en detectar objetos anómalos como vehículos y no analiza el comportamiento humano para identificar actividades sospechosas.

Feizi (2020) propone un algoritmo para detectar comportamientos anormales en videos, basado en la extracción de características y en la agrupación espectral. Este enfoque combina características locales y basadas en objetos para construir un modelo que represente comportamientos normales, optimizado mediante maximización del área bajo la curva ROC (AUC). El método opera de forma semi-supervisada, eliminando la necesidad de etiquetado manual. Durante el entrenamiento, primero se extrae el Histograma de Flujo Óptico (HOF) como característica local; y, segundo, se extraen la velocidad, la varianza de la trayectoria y la desviación de la trayectoria como características basadas en objetos. En la detección, el método aplica un enfoque jerárquico, evaluando primero el HOF y luego las características basadas en objetos, interrumpiendo el proceso cuando se detecta un comportamiento anómalo. Este enfoque eficiente clasifica comportamientos como movimientos a alta velocidad, detenciones prolongadas, trayectorias zigzagueantes y entradas no autorizadas. Los resultados muestran un AUC entre 76 y 87 y un EER entre 14 y 21. Aunque el modelo es útil para analizar comportamientos sospechosos como movimientos rápidos y detenciones, su diseño no incluye redes neuronales profundas, limitando su aplicabilidad en sistemas modernos basados en aprendizaje profundo. Sin embargo, algunos puntos del proceso de clasificación y características utilizadas pueden ser relevantes para este proyecto.

Chen et al. (2020) proponen un algoritmo basado en un flujo dual temporal-espacial para el reconocimiento de operaciones anormales de trabajadores. a) El proceso comienza con la recolección de datos de video mediante cámaras de vigilancia. Se implementa una estrategia de muestreo disperso para extraer segmentos de video de longitud fija tanto de videos normales como anormales. Cada segmento es procesado mediante una red de flujo dual: b) la red espacial extrae características de apariencia basadas en fotogramas RGB, c) mientras que la red temporal captura características de movimiento utilizando gráficos de flujo óptico denso entre fotogramas consecutivos. d) Las características extraídas son fusionadas mediante una función de consenso segmental que combina los resultados de múltiples fragmentos de video para generar un vector de características representativo de cada segmento. El modelo utiliza un enfoque de aprendizaje multi instancia para clasificar los segmentos de videos como normales o anormales. Cada segmento en

un paquete se analiza y se le asigna una puntuación de anomalía en escala de 0 a 1, donde puntuaciones superiores denotan comportamientos anómalos. Para entrenar el modelo, se optimiza una función de pérdida que utiliza una forma mejorada de hinge-loss, incorporando restricciones temporales y de dispersión con el fin de mejorar la precisión en el reconocimiento de anomalías. El modelo mostró una mejora del 10.8% en la exactitud de la detección en comparación con otros métodos. Sin embargo, aunque puede detectar eventos como "no usar casco", "manejo violento" o "uso de equipo de comunicación", un entorno de trabajo tiene reglas establecidas, lo que limita su aplicabilidad en entornos comerciales.

Martinez et al. (2019) detectan comportamientos sospechosos en videos a través de tres módulos: entrada, contexto y gestión de datos. El **primer módulo** procesa las propiedades geométricas y cinemáticas de los objetos. Este módulo incluye dos componentes: a) un módulo para identificar y seguir los objetos de interés mientras se mueven, utilizando un algoritmo ante problemas como la oclusión; y, b) un módulo de localización y cinemática destinado a transformar coordenadas visuales en dos dimensiones a coordenadas del espacio real para calcular las velocidades de los individuos. El **segundo módulo** utiliza información contextual basada en acciones realizadas por las personas. Esta información alimenta la etapa de inferencia, cuyo objetivo es calcular la probabilidad de que se esté registrando una acción sospechosa. Además, el modelo de inferencia se adapta al escenario a través del aprendizaje incremental. Este proceso es asistido por un módulo de retroalimentación en el que un experto revisa y etiqueta escenas ignoradas o falsos positivos para mejorar el rendimiento del sistema. El **tercer módulo** organiza la información etiquetada por el experto. Este módulo aplica un filtro de contexto que realiza una clasificación básica de las muestras recopiladas. El algoritmo registró un AUC de 88.38% y 92.32%, con una tasa de error (EER) de 0.12% y 0.13%. Este modelo detecta acciones como robos, inspecciones de cabinas y agresiones nocturnas. Sin embargo, sus aportes a la presente investigación son limitados porque utiliza el enfoque de aprendizaje incremental.

Pouyan et al. (2023) detectaron conductas de robo mediante tres módulos: detección de cabeza cubierta, detección de multitudes y detección de merodeo (*loitering*). Los dos primeros módulos utilizan YOLOV5, un modelo preentrenado que

fue reentrenado con datos propios. Sin embargo, el módulo más relevante para este análisis es el tercero, que emplea la técnica de seguimiento Deep Simple Online Real-Time Tracking (DeepSORT) para calcular la actividad de merodeo de las personas. DeepSORT rastrea a cada persona en la escena, calculando la suma de las distancias recorridas por cada individuo cada 10 fotogramas. Este cálculo se realiza en una ventana deslizante de 500 fotogramas denominada "fragmento" (*snippet*), con cada paso correspondiente a 10 fotogramas. La actividad de merodeo se mide como la distancia euclidiana total acumulada durante cada fragmento, lo que permite identificar patrones de comportamiento sospechosos. Posteriormente, la información extraída de los tres módulos se utiliza como entrada para una máquina de inferencia difusa (*Fuzzy Inference Machine*), específicamente basada en el enfoque Mamdani debido a su capacidad para interpretar reglas conceptuales. Este proceso consta de cuatro pasos: fuzzificación, inferencia, composición y defuzzificación. El modelo obtuvo un *F1 Score* de 0.537, lo que refleja su capacidad para equilibrar precisión y sensibilidad. Este enfoque se relaciona directamente con el presente proyecto, dado que aborda la detección de merodeo como actividad sospechosa y utiliza modelos preentrenados.

Lee et al. (2018) desarrollaron dos algoritmos para la detección de comportamientos sospechosos en terminales bancarios automáticos. El primer algoritmo identifica actividades de **agacharse y trepar frente al ATM** mediante técnicas de **sustracción de fondo y estimación de altura**, asociando estas acciones con intentos de sabotaje como insertar explosivos o desconectar alarmas. El segundo algoritmo detecta la presencia de **objetos con forma de cinturón** que son utilizados para retirar bruscamente el cajero del lugar, a través de una técnica de **división y combinación de regiones** para identificar patrones geométricos característicos. Los algoritmos logran una **exactitud del 100%** en las detecciones; no obstante, no emplean redes neuronales, limitando su capacidad de generalización. Aunque detectan actividades como agacharse y trepar, su enfoque simplificado difiere del objetivo del presente proyecto.

Patel et al. (2020) utilizan la web semántica para la detección de eventos sospechosos en videos. El algoritmo consta de dos procesos clave. Primero, la detección de eventos mediante reglas basadas en lógica descriptiva (DL) utilizando

SWRL. Estas reglas se formulan considerando la relación entre objetos en la escena y aplican inferencias sobre sus interacciones. Segundo, la detección de actividad sospechosa mediante consultas SPARQL en datos RDF, evaluando relaciones espaciales y temporales en la escena. Se identifican actividades como loitering, tocar el vehículo o intentar abrir la puerta mediante el análisis de proximidad y persistencia de una persona en el entorno del automóvil. La consulta recupera instancias en diferentes intervalos de tiempo, estableciendo patrones que permiten detectar eventos anómalos. El modelo logró una exactitud entre 54% y 88%. Asimismo, aunque este modelo detecta loitering sin usar redes neuronales, la recopilación de características lo realiza a través de un modelo de redes neuronales entrenado previamente.

Kamthe y Patil (2018) desarrollaron un algoritmo basado en reglas semánticas para detectar actividades sospechosas en video. Este algoritmo consta de varias etapas: detección de objetos mediante sustracción de fondo, clasificación de personas u objetos (equipaje), seguimiento de objetos y análisis de características de movimiento para identificar eventos sospechosos. El algoritmo se estructura en tres componentes. Primero, la detección de objetos se realiza mediante *template matching*, calculando la correlación entre un fotograma de referencia y el fotograma actual. Luego, el seguimiento de objetos se lleva a cabo utilizando correlación de histogramas de color. Finalmente, la extracción de características de movimiento clasifica los objetos en cuatro categorías: *desconocido* (nuevo objeto en la escena), *objeto inanimado* (baja velocidad), *persona* (alta velocidad) y *persona inmóvil* (velocidad cero). Con base en estas clasificaciones, se detectan eventos sospechosos como equipaje abandonado, acceso no autorizado y merodeo (*loitering*), definiendo umbrales de tiempo y distancia para clasificar cada evento. El algoritmo logró una exactitud del 93 % y 96% para la detección de loitering y equipaje abandonado, respectivamente. Aunque el algoritmo detecta actividades relevantes, el uso de reglas semánticas limita la capacidad de generalización del presente proyecto.

Chen et al. (2022) plantean una técnica orientada a detectar anomalías en la apariencia y el comportamiento, utilizando un descriptor que combina características geométricas y espaciales. Este descriptor utiliza 29 capas convolucionales y una

capa completamente conectada, e incluye un índice único que permite identificar y rastrear a los individuos entre fotogramas. En interiores, donde los objetos en movimiento pueden ser parcialmente cubiertos por el fondo, el descriptor prioriza las características de la parte superior del cuerpo. El algoritmo emplea el método húngaro destinado a vincular los objetos detectados actualmente con los rastreados en fotogramas previos. Estas trayectorias se analizan para detectar el merodeo, identificado mediante patrones de permanencia prolongada en un área. Aunque no se especificaron métricas cuantitativas, el algoritmo es efectivo bajo diversas condiciones como iluminación variable y superposición de objetos. Este enfoque resulta relevante para la presente investigación, dado que permite detectar el merodeo de manera sencilla.

Cheng et al. (2022) desarrollaron cuatro algoritmos agrupados en dos categorías: patrones inusuales de movimiento y comportamientos inusuales. La primera categoría se divide en dos subgrupos: características en movimiento y relación con lugares. En el primer subgrupo se encuentran los algoritmos para "aimlessly wandering" y "frequent short stops," mientras que en el segundo subgrupo está el algoritmo para "loitering around". En la segunda categoría se encuentra el algoritmo para "unusual visit time". Para detectar **aimlessly wandering**, se analizan puntos espacio-temporales de las trayectorias y evalúa grandes cambios en el ángulo de dirección. Las condiciones incluyen la ausencia de paradas, cambios significativos en el ángulo de dirección, y movimiento continuo con una velocidad mínima. En **frequent short stops**, se utiliza clustering espacio-temporal basado en DBSCAN para identificar paradas breves y frecuentes en las trayectorias. Se define la densidad en función de los umbrales de distancia y tiempo. El algoritmo para **loitering around** detecta patrones de movimientos repetitivos alrededor de un área específica. Mediante análisis geométrico, identifica polígonos en las trayectorias y el grado de superposición entre los mismos. El algoritmo para **unusual visit time** compara los tiempos de visita actuales con un historial de trayectorias previas. Si una visita ocurre en horarios atípicos, se considera anómala. Los algoritmos alcanzan un recall del 93.5%, una precisión del 87.6% y un F1-score de 0.905. Esta investigación permite cimentar las bases para una clasificación estructurada de comportamientos sospechosos.

George et al. (2019) usan el Histograma de Orientación y Magnitud de Flujo Óptico para la detección de actividades anormales. Este enfoque extrae información relevante del movimiento y la apariencia de las escenas utilizando regiones espacio temporales, definidas como paralelepípedos no uniformes. Estas regiones, derivadas mediante una transformación de cizalladura, se adaptan mejor a configuraciones de cámara no frontales. El algoritmo se basa en dos elementos principales. Primero, se extraen las características HOFM a partir de las regiones espacio temporales. Estas características combinan información sobre la orientación y magnitud del flujo óptico, capturando patrones de movimiento específicos en los videos. Segundo, un autoencoder identifica patrones anormales en las características HOFM. Este modelo se entrena con datos de actividades normales, detectando anomalías como desviaciones de los patrones esperados. El algoritmo logró un AUC entre 0.78 y 0.91, con un EER de entre 15% y 29%. El algoritmo detecta vehículos en dirección contraria, bicicletas, patinetas y personas corriendo; sin embargo, solo se adapta a una perspectiva de la cámara.

Q3 ¿Qué métricas son usadas para evaluar las técnicas de detección de actividad sospechosa en video?

Tabla 7

Métricas para evaluar las técnicas de detección de Actividad Sospechosa en video

Nº	Métrica	Descripción	Referencias
1	Exactitud	La precisión (Accuracy) es una métrica que mide el porcentaje de predicciones correctas realizadas por un modelo con respecto al total de predicciones. Es ampliamente utilizada como una métrica global de rendimiento, especialmente cuando los datos están balanceados entre clases. Sin embargo, puede no ser adecuada en casos de clases desbalanceadas.	28, 4, 26, 37
2	AUC	El área bajo la curva (AUC) mide la capacidad de un modelo para distinguir entre clases. Está asociada a la curva ROC (Receiver Operating Characteristic) y permite evaluar el rendimiento de un modelo independientemente del umbral de decisión. Un AUC más alto indica un mejor desempeño en la clasificación.	34, 24, 44

3	EER	La tasa de error igual (EER) es el punto en el cual las tasas de falso positivo y falso negativo de un modelo son iguales. Es una métrica utilizada comúnmente en problemas de verificación y autenticación, ya que proporciona una evaluación del balance entre estos errores.	32, 44
4	Precisión	La precisión (Precision) mide la proporción de verdaderos positivos frente al total de predicciones positivas realizadas por un modelo. Es una métrica útil cuando se busca minimizar los falsos positivos, y es comúnmente utilizada en problemas de clasificación donde los falsos positivos tienen un costo elevado.	37

Análisis Crítico

Zhou et al. (2021) combinan la estimación de profundidad, las convoluciones espacio temporales y los mecanismos de atención para reconocer comportamientos anormales. Proponen un algoritmo para calcular la posición espacial en tres dimensiones del objetivo, cuyos datos se utilizan como entrada para el modelo combinado ST-CNN (Convoluciones Espaciotemporales) y ATT-LSTM (LSTM con Mecanismo de Atención). Este modelo permite capturar datos de video de largo plazo. Además, utilizan un método de compresión de características basado en agrupamiento variable para gestionar tamaños de entrada inconsistentes, particularmente en escenarios de reconocimiento de múltiples personas. El rendimiento del modelo se evalúa con la métrica de **accuracy**, logrando un rango de 50.5% a 84.1%, lo que refleja la variabilidad en la precisión dependiendo de las configuraciones del modelo y los datos utilizados. Este enfoque se complementa con otras métricas específicas como la tasa de detección errónea, la tasa de omisión y el objective recognition rate. En relación con el presente proyecto, el uso de **accuracy** como métrica principal resulta relevante para examinar el desempeño de las redes neuronales en detectar comportamientos sospechosos, considerando que el modelo detecta besos, abrazos y saludo de mano.

Amrutha et al. (2020) integran arquitecturas convolucionales con arquitecturas recurrentes para detectar actividad sospechosa. La red neuronal convolucional utiliza VGG-16, un modelo pre entrenado, para extraer características de alto nivel de las imágenes. Posteriormente, las características extraídas son clasificadas mediante una red LSTM, una variante de RNN diseñada para manejar

dependencias secuenciales. Las imágenes se preprocesan dividiendo los videos en frames, que se redimensionan a 224×224 píxeles y se almacenan como entradas para el modelo. La arquitectura VGG-16, compuesta por capas convolucionales de 3×3 , max pooling de 2×2 , y capas densamente conectadas, se adapta eliminando la última capa para configurarla al contexto del proyecto. El modelo final incluye una capa LSTM con ReLU como función de activación, más una capa de regularización dropout y una capa densa cuya cantidad de neuronas corresponde al número de categorías. El sistema clasifica las actividades como normales o sospechosas (uso de teléfonos móviles en áreas restringidas, peleas o desmayos). El modelo obtuvo una exactitud del 87.15%. Aunque el modelo emplea redes neuronales para detectar el uso de teléfonos móviles o peleas, se diferencia del presente trabajo de investigación porque no tiene un enfoque preventivo.

Balasundaram et al. (2021) diseñaron un modelo por bloques orientado a detectar anomalías en sistemas de videovigilancia. La técnica emplea el flujo óptico para extraer la densidad y velocidad del movimiento, utilizando la Transformada Discreta del Coseno (DCT) para identificar diferencias inusuales en los patrones de movimiento. Primero, el algoritmo analiza el comportamiento del movimiento a nivel de píxeles dentro de una secuencia de imágenes, generando una representación de la actividad basada en la velocidad y dirección de los píxeles en movimiento. Luego, la imagen resultante es evaluada dentro de un área específica mediante la DCT, que permite transformar los datos espaciales en componentes de frecuencia. A fin de mejorar la exactitud de detección, se implementa la técnica Lucas-Kanade para computar el flujo óptico. Posteriormente, se comparan los patrones de movimiento con valores promedio históricos, permitiendo clasificar las actividades como normales o anómalas según la magnitud y dirección del desplazamiento de los píxeles. El modelo alcanza un **exactitud del 87.5%**, para detectar actividades como **bailar, correr, empujar, andar en bicicleta, conducir una camioneta o dispersarse**. Sin embargo, estos comportamientos agrupan otros comportamientos como la interacción con objetos que deben ser analizados.

Sheu et al. (2019) presentan STAM-CCF, un modelo de seguimiento de sospechosos en múltiples cámaras basado en filtros de correlación y detección de objetos con YOLO. Este sistema ajusta la identificación de personas para evitar

errores por oclusión y superposición, además de incorporar un modelo de correlación entre cámaras y reconocimiento de marcha en dos etapas para mejorar la reidentificación. El algoritmo resuelve el problema de **cold-start** permitiendo el seguimiento sin necesidad de preentrenamiento extenso. Utiliza YOLO para detección y filtros de correlación para ajustar la identificación. El enfoque consiste en escoger características clave y técnicas de deep learning para rastrear personas entre cámaras. Con el propósito de incrementar la precisión, considera cámaras adyacentes en la etapa de reidentificación. Los resultados muestran una **accuracy de 41.2% y una precisión de 72.3%**. Esta alternativa detecta el **cambio de ropa**, pero su enfoque está en el rastreo de objetos entre varias cámaras, no en la detección preventiva de comportamientos sospechosos.

Chen et al. (2020) desarrollan un enfoque bidireccional predictivo orientado en detectar anomalías en secuencias continuas de fotogramas. El enfoque predice el mismo frame en dos direcciones, hacia adelante y hacia atrás, utilizando dos subredes U-Net independientes. La evaluación de anomalías se basa en la consistencia entre las dos predicciones generadas y su similitud con el frame real. Para mejorar la detección, se introduce un esquema de ventana deslizante que prioriza áreas clave del primer plano. La arquitectura elimina la estructura multicanal y el discriminador de GAN, reduciendo la redundancia y simplificando el entrenamiento. Se retiene solo el generador U-Net en el canal de intensidad, disminuyendo el costo computacional y mejorando la estabilidad. La función de pérdida combina la pérdida de intensidad, gradiente y flujo óptico, junto con la pérdida típica de GAN. El modelo logró un AUC entre 87% y 96%. Este enfoque detecta objetos raros y comportamientos inusuales; sin embargo, la métrica AUC se usa en modelos con **anchor boxes**, difiriendo en su implementación con respecto al enfoque de la presente investigación.

Mahmood et al. (2021) proponen un enfoque multinivel para detectar y localizar eventos anómalos en secuencias de video. El proceso inicia con la estimación de diferencias temporales entre fotogramas para identificar cambios significativos en la escena. Posteriormente, se modela el comportamiento normal y anormal mediante un modelo de regresión, estableciendo patrones de referencia para evaluar desviaciones en la dinámica del movimiento. Finalmente, se emplea una técnica de

agrupamiento fundamentada en densidad para identificar valores atípicos a nivel de clips de video. El método segmenta los videos en clips no superpuestos a través de un proceso de detección global de anomalías. En cada clip, se analizan las densidades locales para determinar la presencia de eventos fuera de lo común. La efectividad del modelo se evalúa con la métrica AUC, alcanzando valores entre 76% y 83%. En relación con la presente investigación, este enfoque permite identificar entidades no peatonales en áreas destinadas al tránsito de personas, aunque la detección se basa en un modelo de regresión.

Sharma et al. (2020) desarrollaron una arquitectura profunda espaciotemporal basada en distorsión estructural para el reconocimiento y localización de eventos anómalos en video. La propuesta emplea un autoencoder para extraer características espaciotemporales directamente desde el video. El reconocimiento de irregularidades se fundamenta en el análisis de la distorsión estructural entre los videos originales y reconstruidos, considerando la luminancia, el contraste y la estructura. El modelo propuesto opera en tres etapas principales. Primero, el autoencoder profundo aprende la representación espaciotemporal del video, extrayendo patrones característicos en distintas capas ocultas. Luego, se calcula la distorsión estructural comparando el video reconstruido con el original, permitiendo detectar y localizar anomalías en el pixel y frame. Finalmente, se genera un puntaje de anormalidad que clasifica eventos sospechosos. El modelo logró un AUC de 0.8167 y EER entre 11% y 20%. Aunque el modelo es capaz de identificar eventos como ciclistas, motociclistas, patinadores, automóviles y comportamientos sospechosos como correr o escapar, su dependencia en la reconstrucción de muestras genera un mayor costo computacional.

Ata-Ur-Rehman et al. (2021) proponen un enfoque basado en **filtrado de partículas** para la detección de anomalías en videos. Este método identifica frames con actividades anómalas mediante la probabilidad posterior de los eventos. Además, localiza regiones anómalas dentro de los frames detectados como anómalos. Para ello, introduce un **modelo de predicción** para estimar actividades en los cuadros de video y un **modelo de medición** para asignar pesos a estas predicciones. Ambos modelos emplean una estructura de celdas de tamaño variable que segmenta las escenas en subregiones de distintos tamaños y extrae información relevante como

tamaño, movimiento y ubicación de los objetos. El algoritmo divide los cuadros en celdas de diferentes tamaños y, mediante **filtrado de partículas**, predice las posibles actividades en el video. Posteriormente, una fase de actualización refina estas predicciones al extraer características de movimiento, ubicación y tamaño, que luego se combinan con un algoritmo de **clustering** para evaluar la distribución de probabilidad de los eventos. Con esta información, clasifica los frames en normales o anómalos. El modelo alcanza un **EER de 29%** y un **AUC de 0.32**, lo que evidencia una capacidad restringida para detectar anomalías. En relación con la presente investigación, el método es capaz de detectar objetos en movimiento como **vehículos, ciclistas y patinadores**, pero no evalúa comportamientos humanos.

Determinación de la Solución

Considerando los hallazgos obtenidos mediante la revisión sistemática y el análisis detallado del problema planteado, se propuso una arquitectura combinada que fusiona una Arquitectura Convolutiva (CNN) con unidades de Redes Neuronales Recurrentes de tipo LSTM (Memoria a Largo y Corto Plazo). Esta arquitectura se fundamenta en dos fases complementarias y consecutivas que proporcionan un enfoque integral para la detección de actividad sospechosa en ambientes comerciales.

Primera Fase: Detección y Seguimiento

La etapa inicial utiliza una Arquitectura Convolutiva basada en la Red Neuronal YOLO11 para la identificación, clasificación y rastreo de individuos en secuencias videográficas. Los aspectos centrales de esta implementación comprenden:

Detección de Personas:

- Aplicación del modelo YOLO11 especializado en la detección de personas.
- Procesamiento de fotogramas videográficos con resolución estándar.
- Obtención de coordenadas centrales (x, y) y medidas dimensionales (ancho, alto) por cada detección.

Sistema de Seguimiento Persistente:

- Definición de un umbral de distancia euclidiana para la correlación entre identificaciones sucesivas.

- Preservación de identificadores únicos que aseguran la coherencia del rastreo durante toda la secuencia temporal.
- Construcción de trayectorias mediante la conservación de posiciones históricas con límite máximo establecido.

Extracción de Características Espacio-Temporales:

- Compilación sistemática de atributos esenciales: posiciones centrales (x, y), medidas dimensionales del objeto (ancho, alto), marcas temporales, e identificadores únicos de rastreo.
- Estandarización de coordenadas con relación a las dimensiones del fotograma para asegurar independencia respecto a la resolución.
- Generación de archivos estructurados en formato CSV para procesamiento posterior.

Segunda Fase: Modelado de Actividad Normal

La etapa subsecuente implementa una Red Neuronal Recurrente LSTM específicamente diseñada para el aprendizaje de patrones de desplazamiento habituales y la identificación de actividades sospechosas. La estructura neuronal desarrollada exhibe las siguientes propiedades:

- Implementación de capa LSTM con 64 unidades de memoria en la primera capa, 32 unidades de memoria en la segunda capa, utilizando función de activación ReLU para introducir no-linealidad.
- Incorporación de capa completamente conectada de salida con dimensionalidad equivalente a las características de entrada (coordenadas x, y).
- Aplicación del optimizador Adam para el ajuste adaptativo de parámetros mediante descenso de gradiente.
- Utilización de función de pérdida MSE (Error Cuadrático Medio) para la optimización del error predictivo.

Procesamiento de Secuencias Temporales:

- Establecimiento de secuencias de extensión constante (10 puntos temporales consecutivos) para el proceso de entrenamiento.

- Eliminación automática de trayectorias con duración insuficiente para preservar la integridad del corpus de datos.
- Estandarización manual de coordenadas a partir de las dimensiones del fotograma.
- Segmentación del conjunto de datos para la fase de entrenamiento (80%) y validación (20%).

Determinación del Umbral de Anomalía:

- Determinación del umbral de identificación mediante análisis estadístico: $\mu + 3\sigma$, donde μ corresponde a la media del MSE y σ a la desviación estándar.
- Valoración del modelo sobre datos de validación normales para definir la línea base de actividad habitual.
- Detección de actividades sospechosas mediante contraste del error predictivo con el umbral determinado.

Cabe mencionar que, dada la naturaleza no supervisada del aprendizaje implementado, no se emplean métricas convencionales de evaluación como exactitud, precisión, recall o AUC (Área Bajo la Curva). En su lugar, la evaluación se fundamenta en el análisis del error de reconstrucción y la distribución estadística de los errores de predicción sobre datos considerados normales, estableciendo un criterio probabilístico para la detección de actividades sospechosas.

La solución desarrollada establece una diferenciación sustancial respecto a los métodos tradicionales documentados en la literatura, los cuales se concentran principalmente en el análisis de datos de postura corporal o reconocimiento facial mediante CNN. La contribución distintiva del trabajo desarrollado se centra en la implementación de trayectorias de desplazamiento como elemento característico principal para la detección de actividades sospechosas.

Flexibilidad Contextual: El sistema reconoce que las actividades sospechosas mantienen una relación intrínseca con la disposición física particular de cada establecimiento comercial.

Resistencia ante Restricciones Técnicas: La metodología se encuentra diseñada para ajustarse a las limitaciones de calidad visual de los sistemas de videovigilancia

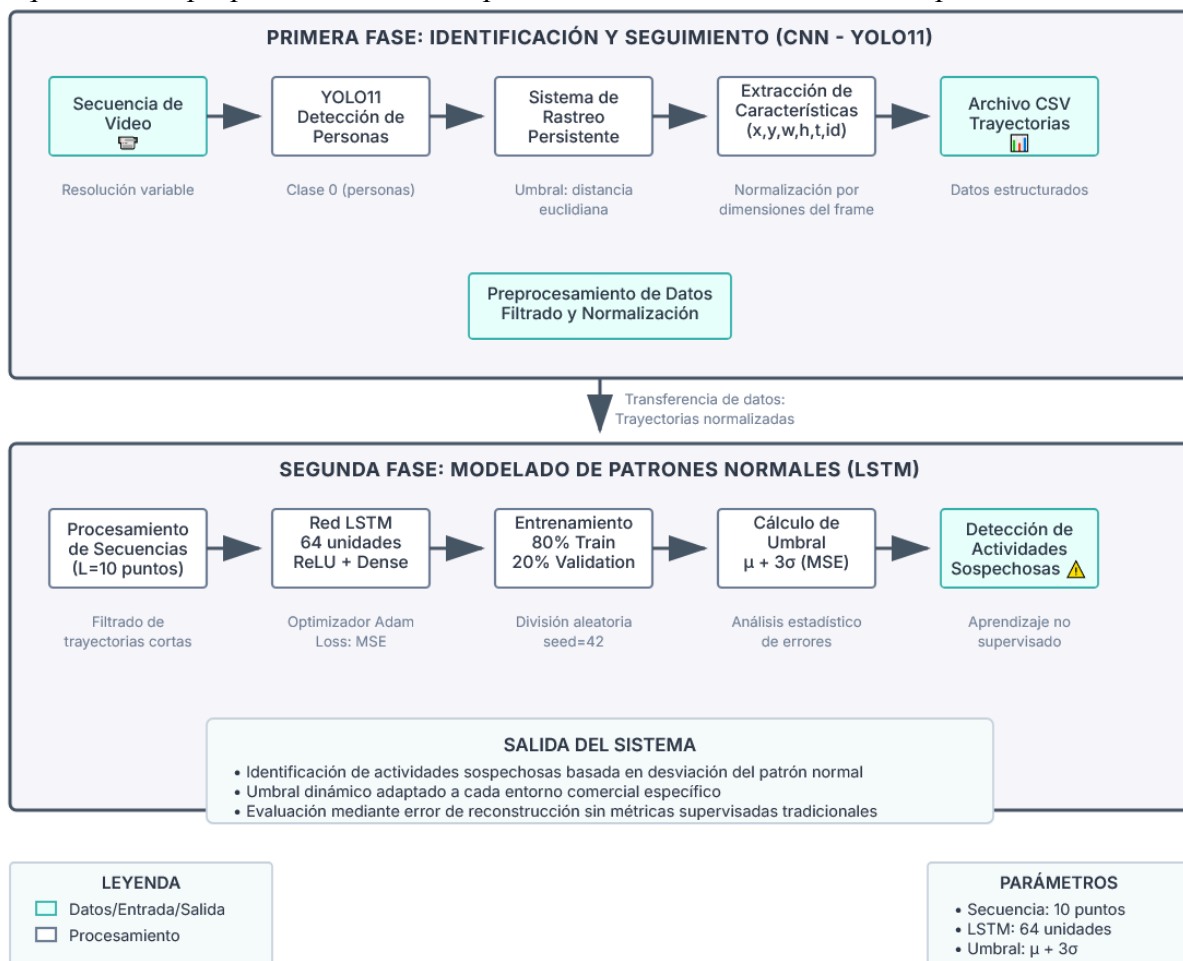
empleados por microempresas, concentrándose en patrones de desplazamiento que presentan menor susceptibilidad a fluctuaciones en resolución o iluminación.

Optimización Computacional: El empleo de trayectorias como elemento característico primordial disminuye considerablemente las demandas computacionales en comparación con métodos en análisis detallado de postura o características faciales.

La arquitectura híbrida CNN-LSTM desarrollada representa, en consecuencia, una solución comprehensiva que integra la efectividad de la detección con la capacidad predictiva del aprendizaje profundo, optimizada específicamente para satisfacer las necesidades y restricciones del sector comercial objetivo.

Figura 4

Esquema de la propuesta CNN-LSTM para la Detección de Actividad Sospechosa



La Tabla 8 resume las principales arquitecturas revisadas en la literatura para la detección de anomalías y actividades sospechosas en videos, agrupadas según sus características clave como entrada, tipo de aprendizaje, componentes y métricas de evaluación. Se observa que modelos como CNN pura se centran en características espaciales estáticas con aprendizaje supervisado, mientras que variantes híbridas como VAE-LSTM o CNN-LSTM incorporan temporalidad para secuencias dinámicas, aunque requieren datasets etiquetados en muchos casos. Las GAN y CNN-GAN destacan por su capacidad reconstructiva no supervisada, priorizando la generación de patrones normales. En contraste, la propuesta YOLO11-LSTM se inspira en estas aproximaciones al enfocarse en trayectorias individuales de desplazamiento (secuencias de 10 frames), utilizando únicamente datos normales en un enfoque predictivo no supervisado basado en MSE y umbral estadístico ($\mu + 3\sigma$).

Tabla 8

Comparación de modelos para detección sospechosa versus propuesta YOLO11-LSTM

Característica	CNN	GAN	VAE-LSTM	CNN-LSTM	CNN-GAN	YOLO11-LSTM (Propuesta)
Entrada	Frames individuales (75×75 - 128×128 px)	Frames de video (256×256 px)	Secuencia de frames (224×224 px)	Frames individuales o secuencias (224×224 px)	Secuencia de 4 frames (256×256 px)	Secuencias de 10 frames (848×480)
Qué analiza	Características espaciales en cada frame	Patrones de apariencia y movimiento en frames	Características espaciotemporales en secuencias	Características espaciales y temporales en cada frame	Apariencia y poses con predicción temporal	Patrones de desplazamiento o individuales por persona
Tipo de aprendizaje	Supervisado	No supervisado	No supervisado	Supervisado, No supervisado, o Semi-supervisado	Semi-supervisado	No supervisado
Datos que aprende	Normal y Anormal	Solo Normal	Solo Normal	Normal y Anormal, o Solo Normal según variante	Solo Normal	Solo Normal
Componentes principales	Capas Conv + Pooling + FC + Softmax	Generador (U-Net) + Discriminador (PatchGAN/CNN)	Encoder (Conv+ConvLSTM) + Latent Space + Decoder	CNN (extracción) + LSTM/ConvLSTM/BiLSTM + FC	Generador (U-Net) + 2 Discriminadores (PatchGAN + 3D-CNN)	YOLO11 (detección/tracking) + LSTM (64 unidades + 32 unidades + FC) + MSE
Tipo de tarea	Clasificación binaria	Reconstrucción → detección por error	Reconstrucción → detección por error	Clasificación o Reconstrucción	Predicción → detección por error de predicción	Predicción de siguiente posición → detección por

						MSE
Características extraídas	Vector de 20 a 1024 características por frame (bordes, texturas, formas)	Frame completo reconstruido (píxeles)	Representación latente de 256 dimensiones (μ , σ)	Vector de características variables; ResNet-50 extrae 1000 características por frame	Frame completo predicho (píxeles)	Tupla de 3 valores (x, y, z) por persona en cada frame
Métricas de evaluación	Accuracy: 75%-99%, Precision: 0.97-1.0, Recall: 0.97-1.0, F1: 0.99 (requiere datos etiquetados Normal+Anormal)	AUC: 85.3%-98.5%, EER: 5.2%-17% (datasets de prueba etiquetados con anomalías)	AUC: 87.2%-88.8%, EER: 12.2%-18.9% (datasets de prueba etiquetados)	AUC: 75%-91%, EER: 8%-30%, Accuracy: 79%-91% (según tipo de tarea y etiquetado de prueba)	AUC: 74.5%-88.6%, EER: 19%-31.6% (datasets de prueba etiquetados)	Tasa de detección: 100% (12/12) en dataset de prueba etiquetado como anormal; umbral $\mu+3\sigma$ en MSE

2.3 Bases Teóricas

2.3.1 Visión Computacional

La Visión por Computadora es una disciplina dentro de la Inteligencia Artificial que busca simular la capacidad visual humana mediante el análisis de imágenes y videos, con el objetivo de comprender el mundo real, según Sánchez (2024). Bhowmik (2024) afirma que, dentro de este subcampo, se estudian algoritmos y enfoques destinados a capturar los datos visuales, así como el estudio de los mecanismos para interpretarla con un razonamiento humano. Raj et al. (2024) afirman que la clasificación de imágenes; es decir, la asociación explícita de una imagen con una etiqueta que la describa de forma general, es el desafío más importante de este subcampo de estudio.

Por lo general, según Umbaugh (2023), la Visión Computacional suele ser aplicada en tareas repetitivas, en entornos con cierto nivel de riesgo para el ser humano o en actividades que procesen grandes cantidades de datos. Desde procesos de control de calidad hasta procesos de análisis de minerales en Marte, la Visión Computacional está penetrando en diversos sectores de la sociedad; no obstante, su uso aún no se ha masificado debido a un avance ralentizado del subcampo como lo indica Sánchez (2024).

2.3.1.1 Procesamiento Digital de Imágenes

El tratamiento computacional de imágenes consiste en la adquisición y tratamiento de los datos visuales mediante el uso de operaciones de bajo y alto nivel. Umbaugh (2023) señala que estas operaciones permiten identificar ciertas características visuales de la imagen como los niveles de intensidad en cada píxel, así como la detección de bordes y objetos.

Asimismo, Raj et al. (2024) indican que este procesamiento es considerado una subdivisión de Visión Computacional, enfocado en la aplicación de algoritmos para mejorar y resaltar las características visuales de una imagen digital. Por lo general, previo a la manipulación computacional de imágenes, se realiza el preprocesamiento de los fotogramas secuenciales para el filtrado de ciertas características visuales.

2.3.1.2 Red Neuronal Convolutiva

De acuerdo con Sánchez (2024) y Umbaugh (2023) la Arquitectura Convolutiva (CNN) constituye un modelo de Aprendizaje Profundo (DL) que se enfoca en resolver problemas complejos en donde la intervención humana es nula o escasa. Esta arquitectura permite aprender patrones y características subyacentes de los datos visuales, por ende es la más utilizada para el análisis de imágenes y videos, según Bhowmik (2024).

Umbaugh (2023) detalla que la Red Neuronal Convolutiva está conformada por varias capas ocultas que pueden subdividirse en capas convolucionales, capas de agrupación, capas de activación y capas totalmente conectadas. Zhao et al. (2019) reportan que las capas convolutivas resaltan determinados patrones de las imágenes a través de la aplicación de filtros, con el objetivo de construir mapas de características. A diferencia de los métodos tradicionales basados en algoritmos, la Red Neuronal Convolutiva tiene la habilidad de adquirir conocimiento espacial; es decir, aunque la imagen de una persona se gire, se comprima o se estire, esta arquitectura seguirá reconociendo a la persona, según Raj et al. (2024).

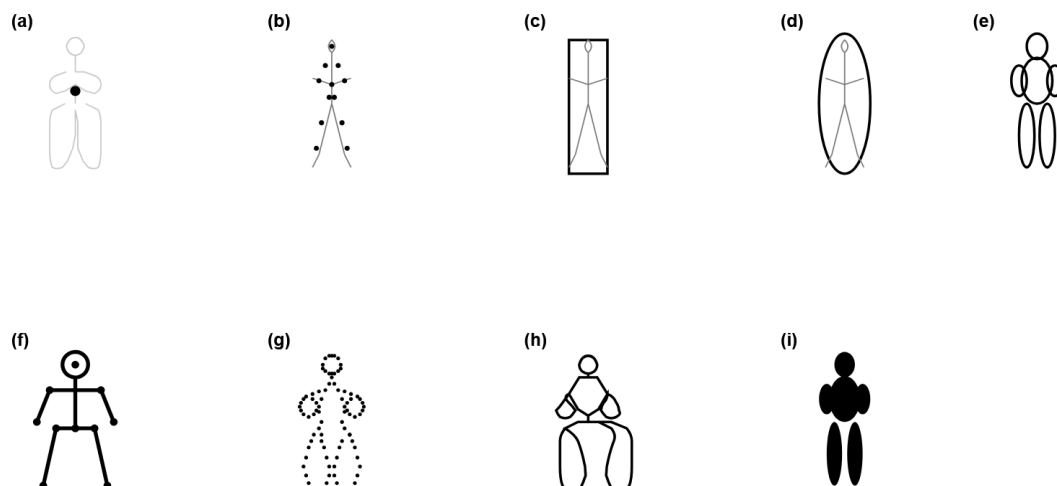
2.3.1.3 Detección y Seguimiento de Objetos

Conforme a Yilmaz et al. (2006), un objeto es cualquier elemento de interés que tenga relevancia para un análisis posterior dentro de un determinado contexto. Zhao et al. (2019) indican que el reconocimiento de diversos elementos consiste en ubicar y etiquetar los objetos en una imagen, proporcionando los cimientos orientados a la interpretación semántica de imágenes y vídeos. Esta comprensión permite desarrollar aplicaciones para la clasificación de imágenes, el comportamiento humano, entre otras. Por lo general, la clasificación de imágenes suele llevarse a cabo con imágenes que contienen un solo objeto; sin embargo, la complejidad del mundo real abarca la existencia de varios objetos en una imagen, según Elgendy (2020).

En el contexto de seguimiento de objetos, después de la detección inicial, Yilmaz et al. (2006) sostienen que el objeto debe ser representado mediante modelos que capturen su forma o apariencia visual como los mencionados en la Figura 5. Khanam y Hussain (2024) detallan que los modelos basados en aprendizaje profundo, como YOLO, integran naturalmente algunas de estas representaciones, particularmente el centroide y el parche rectangular a través de sus predicciones de cuadros delimitadores.

Figura 5

Representaciones de objetos. (a) Centroide, (b) diversos puntos, (c) parches rectangulares, (d) parche elíptico, (e) diversos parches basados en partes, (f) armazón del objeto, (g) borde entero del objeto, (h) marcadores en el borde del objeto, (i) silueta del objeto



2.3.1.3.1 Detección de Personas

Wang et al. (2020) afirman que la detección de personas, también conocida como detección humana, consiste en detectar la presencia de personas, así como identificar con rapidez y precisión su ubicación en una imagen o video. Por lo general, según Elgandy (2020), los modelos basados en redes neuronales artificiales generan regiones de interés en donde podrían estar los objetos, extraen características visuales para clasificar dichos objetos, eliminan detecciones repetidas mediante supresión de no-máximos y evalúan el desempeño del modelo con indicadores como el área bajo la curva y la precisión media. No obstante, Bhowmik (2024) señala que los modelos como YOLO detectan y clasifican los objetos en una sola iteración.

Además, Bhowmik (2024) detalla que YOLO (You Only Look Once) constituye una red neuronal convolutiva diseñada para identificar elementos en imágenes en tiempo real a través de la predicción de la ubicación de las cajas delimitadoras y la generación de probabilidades de clase para cada objeto. Desde su aparición en 2016 hasta su cuarta versión, YOLO utiliza cajas de referencia o anchor boxes como parte del proceso de detección de objetos; sin embargo, como indican Vina, Khanam y Hussain (2024), a partir de la quinta versión, la mayoría de las variantes han optado por un enfoque sin anclas también conocido como anchor-free, como ocurre con YOLO v11.

2.3.1.3.2 Seguimiento de Objetos

Tripathi et al. (2018) sostienen que el seguimiento de objetos consiste en obtener y reproducir gráficamente la trayectoria de un objeto a lo largo del tiempo, según la ubicación que este tiene en cada fotograma del video. Asimismo, Yilmaz et al. (2006) detallan que este seguimiento es considerado como un problema de estimación de rutas bidimensionales mientras los elementos se desplazan a través de un campo visual. Frecuentemente, el proceso de seguimiento de objetos es más complejo de llevar a cabo en comparación con la detección de objetos, dado que se deben manejar factores como la pérdida de información espacial en la proyección dimensional, el bloqueo parcial o total de entidades y las alteraciones en la iluminación ambiental.

Los sistemas de seguimiento de objetos asignan un número a cada entidad detectada con base en la sucesión cronológica de su aparición inicial en la secuencia de video; sin embargo, de acuerdo a Bird et al. (2005), en contextos de multitud peatonal, estos números pueden variar. Estos sistemas presuponen que la cantidad de movimiento y las alteraciones en la apariencia de los objetos entre fotogramas sucesivos son pequeñas, como lo señala Szeliski (2022). El rastreo de elementos en series de video también se conoce como la detección dinámica de objetos, dado que requiere identificar previamente los elementos a seguir, según Bhowmik (2024).

2.3.1.3.3 Características Espacio-Temporales

De conformidad con Schuldt et al. (2004), las características espacio-temporales son descriptores que integran información espacial y temporal para representar comportamientos de movimiento, correspondientes a estructuras visuales bidimensionales en momentos de alteración de su estado de movimiento. Dollar et al. (2005) exponen que los descriptores espacio-temporales han sido aplicados exitosamente en investigaciones orientadas al análisis conductual de roedores y humanos, mostrando resultados significativos en el análisis de comportamientos.

Por lo expuesto por Forke y Tropmann-Frick (2021), la información espacio-temporal se estructura en cinco modalidades fundamentales: sucesos discretos localizados temporalmente, rutas de desplazamiento de entidades móviles, ubicaciones de referencia con mediciones variables, matrices georeferenciadas y contenido audiovisual secuencial. Específicamente en el análisis de video, las características espacio-temporales representan puntos de interés en un video que codifican actividades humanas combinando información de localización y duración temporal, según Ben Mabrouk y Zagrouba (2017).

2.3.2 Análisis Espacio-Temporal

Bhowmik (2024) refiere que el análisis de las características espacio-temporales se lleva a cabo mediante kernels convolucionales implementados en estructuras de redes neuronales profundas. Los enfoques basados en el movimiento espacio-temporal son computacionalmente eficientes y sencillos de implementar para capturar información dimensional y cronológica del movimiento de la marcha; no obstante, según Paul et al. (2013), son vulnerables ante perturbaciones de ruido y desviaciones considerables del movimiento.

2.3.2.1 Representación de Trayectorias

Kumaran et al. (2018) definen una trayectoria como una secuencia temporal de datos de posicionamiento que documenta las ubicaciones sucesivas de un objeto en el espacio, indexadas de manera ordenada según su secuencia cronológica. Para el análisis de estas trayectorias, Bhowmik (2024) detalla el uso de diversos enfoques metodológicos, entre los cuales destacan los filtros o kernels convolucionales implementados en estructuras de redes neuronales profundas. Específicamente en el contexto del seguimiento de personas, una trayectoria constituye el recorrido que una persona efectúa en función del tiempo, siendo frecuentemente empleada en el análisis de escenas para examinar la actividad o comportamiento de los individuos, según Rai et al. (2015).

Li et al. (2013) detalla que la representación de las trayectorias en secuencias de video se formaliza mediante una tupla tridimensional (x, y, t) , en donde las variables 'x' e 'y' corresponden a las coordenadas espaciales bidimensionales que definen la ubicación del objeto en el espacio del fotograma y la variable 't' establece el instante específico en el cual dicha posición es registrada.

2.3.2.2 Normalización de Datos

Según Huang et al. (2023), la normalización de datos representa un proceso de transformación general que asegura que la información procesada mantenga propiedades estadísticas determinadas, constituyendo un procedimiento esencial en el preprocesamiento que permite estandarizar variables con diferentes escalas o rangos. Como una de las técnicas fundamentales para la normalización de datos, el escalamiento transforma los valores originales mediante operaciones básicas entre constantes, facilitando el procesamiento homogéneo de variables que inicialmente presentan diferentes magnitudes, escalas o unidades de medida, conforme a lo expuesto por Huang et al. (2018).

En este sentido, la normalización de datos ha sido desarrollada con el objetivo de abordar las limitaciones inherentes al procesamiento de datos heterogéneos. Zhang et al. (2018) señalan que este proceso reduce la complejidad del espacio computacional y optimiza la eficiencia de los esquemas de redes neuronales artificiales, mejorando así el rendimiento y la precisión en el tratamiento de información con características diversas.

2.3.3 Datos Secuenciales

De acuerdo con Lee et al. (2019), los datos secuenciales constituyen datos históricos organizados cronológicamente en donde cada elemento mantiene dependencias contextuales con sus predecesores. Zargar (2021) detalla que estas estructuras secuenciales permiten que cada observación posicional aporte información contextual relevante a elementos situados en distintas posiciones de la serie, siendo relevante en problemas de aprendizaje donde las entradas secuenciales generan salidas únicas como

en el reconocimiento de patrones en video. No obstante, según Yan et al. (2019), estas secuencias habitualmente presentan variabilidad estructural y predictiva.

Köck y Paramythis (2011) definen el modelado de secuencias de actividades como un elemento fundamental en el análisis de datos secuenciales, dado que proporciona la capacidad de examinar y establecer inferencias sobre conjuntos de acciones ejecutadas por múltiples usuarios. Es preciso señalar que, como lo menciona Salem (2022), una proporción considerable de la información mundial se presenta de forma secuencial, abarcando diversos tipos de datos como los audiovisuales. En ese contexto, Tax et al. (2020) explica que los enfoques metodológicos diseñados para el tratamiento de datos secuenciales buscan generalizar adecuadamente a partir de los conjuntos de datos presentes en los datasets de entrenamiento.

2.3.3.1 Redes Neuronales Recurrentes

Jurafsky y Martin (2025) afirman que las Redes Neuronales Recurrentes son arquitecturas que incorporan conexiones cíclicas en su estructura topológica, generando una dependencia donde cada unidad de procesamiento utiliza como entrada sus propias salidas precedentes. Asimismo, Goodfellow et al. (2016) señalan que estas redes neuronales son capaces de procesar secuencias extensas en comparación con aquellas redes que carecen de especializaciones orientadas al manejo de datos secuenciales. Darmawahyuni et al., (2019) relatan que las arquitecturas recurrentes son consideradas como una técnica de aprendizaje profundo debido a su procesamiento automático para la extracción de características, prescindiendo de la necesidad de especificar previamente atributos relevantes del conjunto de datos.

En el contexto del modelado secuencial, las Redes Neuronales Recurrentes se establecen como estimadores probabilísticos de alta eficiencia que buscan integrar información contextual de largo alcance para la generación de predicciones, según Krause et al. (2016). Sin embargo, Gao et al. (2019) reportan la existencia de restricciones computacionales derivadas de la dependencia secuencial que obstaculiza el procesamiento paralelo en estas

arquitecturas, además de enfrentar desafíos en el entrenamiento como el desvanecimiento y la explosión del gradiente.

2.3.3.2 Redes de Memoria a Largo y Corto Plazo

De acuerdo a Yadav et al. (2020), las arquitecturas de Memoria a Largo y Corto Plazo (LSTM) representan una categoría de Redes Neuronales Recurrentes que fueron desarrolladas en 1997, con el objetivo de resolver el desvanecimiento del gradiente y evitar la pérdida de información a lo largo del tiempo. Waqas y Wannasingha Humphries (2024) detallan que a través de un mecanismo de compuertas en el procesamiento de los datos, esta red utiliza información contextual de elementos anteriores para la predicción de elementos futuros.

Estas arquitecturas, según Kanrar et al. (2023), procesan secuencias de imágenes en la clasificación de videos, demostrando eficacia en el reconocimiento de acciones y eventos complejos mediante el uso de vectores de características que permiten identificar patrones y objetos. Jurafsky y Martin (2025) detallan que estas redes gestionan la información contextual mediante la descomposición del problema en dos componentes fundamentales: la eliminación de datos irrelevantes del contexto y la incorporación de información que, con alta probabilidad, resultará necesaria para la toma de decisiones en fases subsiguientes.

2.3.3.2.1 Componentes

Ben-Ari y Shwartz-Ziv (2017) sustentan que los modelos de Memoria a Largo y Corto Plazo gestionan una celda de memoria diseñada para mantener los estados durante intervalos prolongados de secuencias, a través de la implementación de compuertas no lineales que regulan el flujo de información. En el diseño original de las redes LSTM, los bloques de memoria en la capa oculta recurrente estaban conformados por una compuerta de entrada y otra de salida; sin embargo, según Sak et al. (2014), fue necesario la implementación de un compuerta de olvido para impedir que la red mantuviera información de forma indefinida al procesar secuencias extensas.

Yadav et al. (2020) detallan que la arquitectura de estas redes está conformada por tres mecanismos de control: las válvulas de olvido, las compuertas de entrada y las compuertas de salida. Asimismo, Jurafsky y Martin (2025) agregan que estos componentes mantienen un diseño similar entre sí, el cual consiste en la implementación de una capa de propagación hacia adelante, acompañada por el uso de una función activadora sigmoide y culminando con una operación de multiplicación por elementos aplicada sobre la capa regulada por compuertas.

La **compuerta de olvido** elimina información contextual que dejó de ser relevante, evaluando el estado previo y la entrada actual para determinar qué datos deben descartarse. La **compuerta de entrada** selecciona nueva información relevante mediante dos procesos: genera información candidata y aplica un filtro que decide qué proporción debe incorporarse al contexto. Finalmente, la **compuerta de salida** determina qué elementos del contexto actualizado son pertinentes para la decisión inmediata, generando el estado oculto que influirá en predicciones futuras, según Jurafsky y Martin (2025).

Estos componentes regulan de manera selectiva los procesos de lectura, escritura y reinicio de la información almacenada en el estado de la celda, permitiendo que las redes LSTM mantengan memoria selectiva a largo plazo mientras procesan eficientemente información nueva.

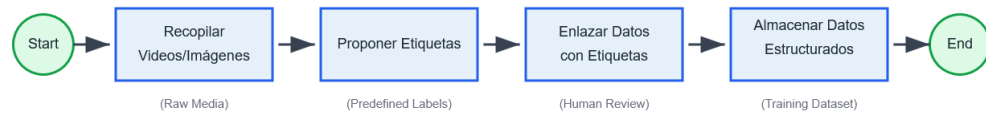
2.3.3.2.2 Datos de Entrenamiento, Validación y Pruebas

La preparación de los datos usados para el **entrenamiento** implica la transformación sistemática de datos no estructurados en datos organizados semánticamente. De acuerdo con Sarkis (2023), la anotación de datos es la actividad principal de este proceso, definida como el procedimiento mediante el cual se asignan etiquetas descriptivas a los datos recopilados, con el objetivo de proporcionar el contexto que necesitan los modelos de aprendizaje profundo. Inicialmente, se recopilan las imágenes o videos que conformarán el corpus de datos. Posteriormente, se establecen las etiquetas que reflejan las características principales que el modelo debe reconocer. Durante la fase de anotación, expertos humanos evalúan la relación entre cada elemento multimedia y las etiquetas predefinidas. Finalmente, estos

datos etiquetados se estructuran y almacenan como el conjunto de información necesaria para el aprendizaje.

Figura 6

Proceso de preparación del dataset de aprendizaje para modelos de Visión Computacional



Los **datos de validación** constituyen un subconjunto especializado de información empleado para la optimización de hiperparámetros, en un proceso denominado Ajuste de Hiperparámetros. Según Maleki et al. (2020), este procedimiento permite determinar los parámetros óptimos que potencian el desempeño del sistema y, al mismo tiempo, preserva la capacidad de generalización. En opinión de Hastie et al. (2009), los datos de validación forman parte del corpus de datos, en donde el tamaño del conjunto de validación se determina mediante metodologías de partición en función del volumen total de muestras disponibles, el grado de variabilidad de los datos y las características del modelo implementado.

Los **datos de prueba**, conforme a lo expuesto por Zhang et al. (2021), están destinados exclusivamente a la valoración del desempeño de los modelos de aprendizaje automático, manteniendo una separación estricta respecto al proceso de entrenamiento para garantizar la validez de los resultados. Akula y Bojja (2024) precisan que este subconjunto de datos es utilizado posterior al proceso del adiestramiento del modelo computacional. La evaluación mediante datos de prueba se centra en la medición de la precisión de los modelos, utilizando muestras extraídas aleatoriamente del corpus de datos, según Ma et al. (2018). Esta metodología permite identificar problemas de sobreajuste; es decir, cuando el modelo alcanza un alto rendimiento en los datos de preparación, pero exhibe una capacidad limitada para generalizar la información no observada con anterioridad.

2.3.3.3 Optimización

Li et al. (2022) definen la optimización como un proceso orientado a la minimización de funciones objetivo en redes neuronales artificiales que, dado el elevado poder computacional requerido, emplea optimizadores en la fase de aprendizaje para reducir al mínimo la función de error y obtener la parametrización óptima de la red. No obstante, Goodfellow et al. (2016) afirman que la optimización no solo busca minimizar, sino establecer los valores límite de una expresión funcional mediante la modificación de los valores de 'x', en donde los problemas de maximización pueden reformularse como procesos de minimización a través de la optimización de la función negativa $-f(x)$.

2.3.3.3.1 Algoritmo ADAM

Kochenderfer y Wheeler (2025) afirman que el método de estimación de momentos adaptativos, o Adam, es una técnica de optimización capaz de ajustar de forma específica la velocidad de aprendizaje en función de los parámetros de la red, integrando el registro suavizado de la magnitud de los gradientes con el registro suavizado de su dirección. Este algoritmo combina las ventajas de RMSProp y Momentum, pero con características adicionales. Goodfellow et al. (2016) detallan cómo en Adam, el Momentum se incorpora como una estimación del primer momento del gradiente ponderado exponencialmente, mientras que la información sobre la magnitud se representa como una estimación del segundo momento; asimismo, el algoritmo aplica correcciones de sesgo a ambas estimaciones para compensar los efectos de la inicialización en cero.

Kingma y Ba (2015) sintetizan que el algoritmo Adam se caracteriza por la eficiencia computacional y el bajo consumo de memoria, demostrando efectividad en contextos con grandes datasets y robustez ante gradientes dispersos. A pesar de ser el optimizador más utilizado en el campo del Aprendizaje Profundo, este algoritmo requiere un ajuste cuidadoso de la tasa de aprendizaje para obtener resultados óptimos, según Szeliski (2022).

2.3.3.3.2 Error Cuadrático Medio

Goodfellow et al. (2016) formulan que el Error Cuadrático Medio (MSE) cuantifica la desviación global esperada entre un estimador $\hat{\theta}_m$ y el valor real del parámetro poblacional θ , a través de la expresión $MSE = E[(\hat{\theta}_m - \theta)^2]$. En la presente investigación, θ representa el patrón que caracteriza las trayectorias normales aprendidas durante el entrenamiento del modelo, análogo al comportamiento típico esperado que define una trayectoria normal; asimismo, $\hat{\theta}_m$ constituye la estimación obtenida al evaluar una nueva trayectoria con el modelo entrenado. Por tal motivo, un Error Cuadrático Medio elevado indica una desviación significativa respecto al comportamiento esperado, identificando trayectorias que se apartan de la normalidad aprendida.

Sin embargo, Engelbrecht (2007) advierte que el uso de la media del error al cuadrado como único parámetro de análisis presenta limitaciones en contextos de clasificación porque puede existir un alto porcentaje de aciertos mientras el error es elevado, generando una discrepancia entre la magnitud del error cuadrático y la capacidad discriminatoria del modelo. Usar el Error Cuadrático Medio para detener el entrenamiento puede prolongarlo y favorecer el sobreajuste; sin embargo, esto no ocurre en contextos de predicción. En síntesis, esta métrica resulta idónea para cuantificar la desviación entre el patrón de trayectoria esperado y el observado, permitiendo detectar anomalías a partir de incrementos significativos en el error.

2.3.4. Aprendizaje No Supervisado

En concordancia con Goodfellow et al. (2016), el aprendizaje sin supervisión es un enfoque de entrenamiento orientado a analizar bases de datos multidimensionales, con el objetivo de modelar la distribución de probabilidad que da origen a los patrones o estructuras subyacentes en los datos sin etiquetar. Asimismo, este enfoque de aprendizaje se distingue por descubrir características implícitas presentes en los datos de entrada sin la intervención de una fuente externa e identificar agrupaciones con las

mismas características sin requerir conocimiento previo de las clases, según lo señalado por Engelbrecht (2007) y Umbaugh (2023), respectivamente.

2.3.4.1 Modelado de Patrones Normales

Williem et al. (2008) sostienen que un patrón normal es un comportamiento contextualmente determinado. Ravanbakhsh et al. (2017) detallan que este patrón es aprendido mediante características extraídas de actividades consideradas normales para un determinado contexto; asimismo, Antić y Ommer (2011) añaden que este patrón sirve como modelo de referencia para clasificar nuevas instancias como normales o anómalas, según el grado de ajuste al patrón establecido.

El modelado de los patrones normales permite establecer comparaciones efectivas entre comportamientos observados y modelos de referencia mediante la proyección de patrones de actividad, según Aggarwal y Ryoo (2011). Dogra et al. (2015) describen que estos patrones pueden ser representados estructuralmente mediante grafos de adyacencia regional que capturan las interacciones espaciales de la escena, con el objetivo de detectar y localizar las desviaciones significativas. No obstante, Aldayri y Albattah (2022) consideran que las técnicas de redes neuronales profundas han mostrado mayor efectividad y precisión para este tipo de tareas.

2.3.4.2 Anomalías

Ben Mabrouk y Zagrouba (2018) definen una anomalía como una observación, evento o comportamiento que se desvía de los patrones normales identificados a partir de los datos habituales; por ello, el registro que no se ajusta a las reglas derivadas de dichos patrones es un valor atípico. En el contexto estadístico, Aldayri y Albattah (2022) resaltan que una anomalía es una anormalidad, desviación o valor atípico. Las anomalías se pueden clasificar en aquellas asociadas a objetos inusuales que se encuentran dentro del entorno y aquellas vinculadas a comportamientos atípicos, según Antić y Ommer (2011).

Asimismo, Antić y Ommer (2011) precisan que el número de posibles anomalías es indefinido, dado que no existe una cantidad suficiente de

ejemplos representativos capaces de abarcar todas las manifestaciones anómalas. Por tal motivo, con base en lo señalado por Arifoglu y Bouchachia (2017), hay investigaciones que generan de manera artificial ciertas anomalías en los conjuntos de datos.

2.3.4.3 Detección de Anomalías

Aldayri y Albattah (2022) declaran que la detección de anomalías es un proceso orientado a identificar patrones o comportamientos que se apartan de lo considerado normal o esperado; asimismo, Antić y Ommer (2011) agregan que este proceso se lleva a cabo mediante el análisis de secuencias obtenidas por cámaras estacionarias. El objetivo del proceso es reconocer comportamientos que tienden a ser clasificados como patrones anómalos o irregulares, según Wang et al. (2018).

Antić y Ommer (2011) postulan que el reconocimiento de eventos inusuales en fotogramas secuenciales puede abordarse mediante dos metodologías: la detección a nivel de fotograma y la detección con precisión a nivel de píxel. En la primera, un fotograma es anómalo si presenta una o más irregularidades. En la segunda, el procesamiento a nivel de píxel permite localizar de forma exacta las regiones comprometidas. Además, Wang et al. (2019) agregan que estas dos metodologías pueden desarrollarse bajo dos enfoques: aquellos que se basan exclusivamente en el análisis del movimiento y aquellos que integran las características espaciales y temporales, a fin de aumentar la precisión y la robustez del modelo.

2.3.5 Actividades Sospechosas

Tripathi et al. (2018) sostienen que las actividades sospechosas son comportamientos inusuales poco frecuentes en espacios públicos que constituyen un peligro potencial para la protección de las personas. No obstante, la conceptualización de una actividad sospechosa difiere significativamente entre individuos, dado que la percepción de sospecha varía entre observadores según el contexto en el que ocurre; por tal motivo, Williem et al. (2008) sustentan que el término ‘sospechoso’ conlleva un carácter inherentemente subjetivo.

Martínez-Mascorro (2020) detalla que la actividad sospechosa engloba un conjunto secuencial de acciones que se manifiesta en el período precedente a la comisión de un delito, aunque este no llegue a materializarse. En consecuencia, Pouyan et al. (2023) consideran que un determinado delito puede predecirse mediante el análisis de una actividad sospechosa. No obstante, con base en el análisis de Amrutha et al. (2020), la identificación de una actividad sospechosa resulta compleja debido a la naturaleza impredecible del comportamiento humano. En apoyo a esto, Pawar y Attar (2019) precisan que las actividades sospechosas pueden entenderse como eventos anómalos que se distinguen por su carácter infrecuente, inusual o novedoso.

En esta línea, Sheu et al. (2019) señalan que la evidencia empírica sugiere que, en la mayoría de los eventos delictivos, el comportamiento sospechoso es llevado a cabo por personas que buscan ocultarse o pasar inadvertidas. Además, Cheng et al. (2022) consideran que la detección temprana de dichos comportamientos puede coadyuvar en la mitigación del crimen y en el reforzamiento de la seguridad ciudadana.

2.3.5.1 Clasificación

Las actividades sospechosas pueden clasificarse según su nivel de amenaza y potencial destructivo. En la cúspide de esta clasificación se ubican aquellas conductas que representan peligro inmediato para la vida humana, tales como episodios violentos, incendios intencionales, atentados con explosivos utilizando equipaje como vehículo, enfrentamientos físicos y actos vandálicos, según Tripathi et al. (2018) y Ata-Ur-Rehman et al. (2021). Estas situaciones demandan respuesta de emergencia debido a su capacidad destructiva inmediata.

Un segundo nivel de criticidad, como lo describen Tripathi et al. (2018) y Ata-Ur-Rehman et al. (2021), agrupa comportamientos que, si bien no implican violencia directa, constituyen infracciones graves o preparativos para actos delictivos. Aquí se incluyen los hurtos, siniestros viales, transgresiones al código de tránsito e ingresos indebidos a áreas de acceso restringido. Como contribución, Patel et al. (2020) mencionan que este nivel

incorpora las conductas anómalas en proximidad a vehículos, caracterizadas por desplazamientos circulares sospechosos cerca de vehículos, manipulación no autorizada, inspección visual del interior, intentos de daño, forzamiento de cerraduras o presencia grupal sospechosa en las inmediaciones. Lee et al. (2018) agregan las posturas corporales inusuales como agacharse o escalar, que podrían indicar sabotaje a sistemas de seguridad o colocación de dispositivos peligrosos, además de cruces fronterizos irregulares como lo señalan Tripathi et al. (2018).

En el nivel básico de alerta se posicionan las actividades que, aunque no representan amenaza inmediata, alteran los patrones normales de comportamiento público. Este grupo abarca, con base en lo señalado por Cheng et al. (2022) y Chen et al. (2022), los desplazamientos erráticos sin propósito aparente, detenciones reiteradas y breves, rondas persistentes en ubicaciones específicas y aproximaciones a sitios de importancia estratégica. Complementan este nivel los incidentes no intencionales como desplomes de personas, desprendimiento de objetos y movimientos masivos precipitados de individuos, según Tripathi et al. (2018), Ata-Ur-Rehman et al. (2021) y Mehmood (2021).

2.3.5.2 Merodeo

Ben Mabrouk y Zagrouba (2018) definen el merodeo como la permanencia prolongada de una persona en un espacio público sin un objetivo definido, ejemplificada en individuos que, portando objetos o equipaje, permanecen en el lugar sin mostrar una finalidad clara en su conducta. En este sentido, Bird et al. (2005) precisan que se trata de un comportamiento anómalo caracterizado por la estancia continua de un individuo en una ubicación durante un intervalo de tiempo mayor al esperado en el contexto del área monitoreada.

Asimismo, Lee et al. (2018) indican que el merodeo puede manifestarse mediante desplazamientos repetitivos que realiza una persona en un área específica sin un propósito aparente, generalmente con la intención de observar el entorno para la realización de un delito. Dicho comportamiento se refleja en trayectorias con múltiples patrones circulares que se apartan

de las rutas directas convencionales, revelando un movimiento no lineal y recurrente que denota la ausencia de un destino definido, según Rai et al. (2015).

Finalmente, el merodeo no se limita al ámbito peatonal, dado que también se presenta en el contexto vehicular. En este caso, Shen et al. (2015) lo definen como el patrón de circulación o permanencia repetitiva de un vehículo en una zona específica sin un fin legítimo aparente, con el propósito de vigilar o recolectar información del entorno para la preparación de un delito.

2.4 Marcos Conceptuales

La detección de actividad sospechosa en entornos de seguridad y vigilancia implica una comprensión sólida de varios conceptos fundamentales. En este marco conceptual, identificamos los términos clave que forman la base de este campo crucial para anticipar y reaccionar frente a escenarios de peligro.

1. **Arquitectura de Red Neuronal:** La configuración arquitectónica de una red neuronal alude a su disposición y diseño estructural de una colección de elementos conectados entre sí, conocidos como neuronas computacionales, que funcionan de manera coordinada para interpretar información y cumplir objetivos particulares, según Goodfellow et al. (2016). En el contexto de la detección de actividad sospechosa, optar por una determinada arquitectura neuronal afecta sustancialmente las capacidades del sistema para reconocer patrones y anomalías en los datos de video.
2. **Técnica o Algoritmo:** En línea con Russell y Norvig (2010), una técnica o algoritmo se refiere a un conjunto de instrucciones o etapas que se ejecutan para completar una actividad específica. En la detección de actividad sospechosa, las técnicas y algoritmos abarcan desde métodos tradicionales basados en reglas hasta enfoques avanzados de aprendizaje automático y procesamiento de imágenes. Dichas técnicas computacionales determinan de qué forma se analizan los datos de video y se identifican patrones de comportamiento anómalos.

3. **Métricas de Evaluación:** De acuerdo con Sokolova y Lapalme (2009), los indicadores de valoración son herramientas empleadas para cuantificar el rendimiento y la eficacia de los sistemas basados en redes neuronales. En el contexto de la actividad sospechosa, las métricas como la precisión, sensibilidad y el puntaje F1 son cruciales para evaluar el potencial de un modelo entrenado con supervisión humana para identificar y clasificar eventos de manera precisa.
4. **Actividad Sospechosa:** Mahmood et al. (2021) definen la actividad sospechosa como comportamientos o eventos que se desvían de los patrones normales o esperados en un entorno determinado. En la detección de actividad sospechosa en videos, esto puede incluir acciones inusuales o riesgosas que exigen consideración específica por parte de los operadores de seguridad.
5. **Red Neuronal Convolucional:** Una arquitectura neuronal convolucional (CNN) es una categoría de red neuronal concebida expresamente para trabajar con datos organizados en matrices como contenido visual, según LeCun et al. (1998). Las CNN son especialmente efectivas en la detección de patrones visuales y características en fotogramas, asegurando su selección para la detección de actividad sospechosa en videos de vigilancia.
6. **Precisión:** Sokolova y Lapalme (2009) lo definen como un indicador que cuantifica el porcentaje de aciertos en los pronósticos acertados realizados por un sistema en proporción al número completo de pronósticos realizados. En el contexto de la detección de actividad sospechosa, la precisión indica la proporción de alertas verdaderamente anómalas con respecto al total de alertas generadas por el sistema.
7. **Sensibilidad:** La tasa de detección real, también llamada tasa de aciertos positivos, evalúa la habilidad de un modelo para reconocer adecuadamente eventos anómalos, en conformidad con Powers (2011). Se refiere a la proporción de eventos sospechosos identificados correctamente en comparación con el total de eventos sospechosos presentes en los datos.

8. **Puntaje F1:** Como lo indica Van Rijsbergen (1979), la puntuación F1 constituye un indicador compuesto que integra precisión y tasa de detección real simultáneamente, proporcionando una evaluación equilibrada del rendimiento del sistema. Se calcula como la media ponderada entre la precisión y la tasa de detección real, mostrando mayor utilidad en escenarios con clases desbalanceadas.
9. **Detección:** Russell y Norvig (2010) definen la detección como el proceso de identificar y reconocer eventos o patrones específicos en los datos, en este caso, eventos de actividad sospechosa en videos de vigilancia. Los algoritmos y técnicas de detección buscan señales anómalas y desviaciones de los comportamientos normales.
10. **Productividad:** Salas y Cannon-Bowers (2001) señalan que la productividad alude al grado de efectividad con que un sistema de detección de actividad sospechosa cumple sus objetivos. Implica la capacidad de identificar eventos anómalos de manera precisa y oportuna, minimizando los falsos positivos y optimizando el uso de recursos.
11. **Eficacia:** Endsley (2004) define la eficacia como el potencial de un sistema para cumplir con sus propósitos de detección de manera precisa y certera. Un sistema eficaz es capaz de identificar actividades sospechosas con un alto grado de exactitud, minimizando así los falsos positivos y negativos.
12. **Eficiencia:** El concepto de eficiencia hace referencia al potencial de un sistema para realizar la detección de manera rápida y sin desperdiciar recursos innecesarios. Un sistema eficiente puede identificar actividades anómalas en tiempo real y generar alertas precisas con la menor cantidad de retraso posible, optimizando la utilización de recursos, según Salas y Cannon-Bowers (2001).

CAPÍTULO 3: METODOLOGÍA

3.1 Tipo y Diseño de Investigación

3.1.1 Identificación de Variables

- **Variable Independiente:** Modelo de Redes Neuronales
- **Variable Dependientes:** Detección de actividad sospechosa

3.1.2 Enfoque de Investigación

De acuerdo con Hernández et al. (2014), este estudio se desarrolla bajo una perspectiva cuantitativa, dado que se emplean datos numéricos y medibles para el análisis. En este trabajo de investigación, se utilizan las trayectorias de desplazamiento representadas en coordenadas (x, y, t), obtenidas mediante una Red Neuronal Convolucional. Posteriormente, las trayectorias son procesadas en una arquitectura neuronal de Memoria a Largo y Corto Plazo para aprender los patrones de comportamiento normal, con el objetivo de establecer un umbral de predicción. Con base en el umbral, se determina si una nueva trayectoria presenta desviaciones significativas en comparación con la predicción realizada con base en los patrones aprendidos. El enfoque cuantitativo se justifica en la precisión matemática y estadística que permite medir y evaluar objetivamente los resultados del modelo propuesto.

3.1.3 Tipo de Investigación

Este estudio corresponde al tipo aplicado porque indaga la solución a un problema concreto de carácter práctico: la detección de actividades sospechosas en bodegas mediante modelos de visión computacional. Conforme a Hernández et al. (2014), el estudio aplicado se nutre del conocimiento teórico previamente desarrollado, con el fin de generar propuestas innovadoras que aporten soluciones en contextos reales. En este trabajo de investigación, los fundamentos de arquitecturas convolucionales y de memoria a largo y corto plazo se integran en el diseño de un modelo que predice la trayectoria esperada con base en los datos normales aprendidos. De tal

manera, la actividad se considera sospechosa cuando la trayectoria observada presenta una desviación significativa respecto a la predicción realizada.

3.1.4 Nivel o Alcance de Investigación

La naturaleza de esta investigación es explicativa–predictiva. En opinión de Hernández et al. (2014), un estudio explicativo se orienta a alegar las causas de los eventos y las condiciones en las que estos se presentan, buscando identificar vínculos causales. En el marco de esta investigación, dicho carácter se refleja en la necesidad de comprender bajo qué circunstancias una trayectoria de movimiento humano puede considerarse anómala y, por tanto, sospechosa.

Asimismo, el estudio se enmarca en un alcance predictivo, dado que se emplean técnicas de aprendizaje profundo para anticipar el comportamiento esperado de las trayectorias a partir de patrones previamente aprendidos. Cuando la trayectoria observada difiere significativamente de la predicción generada por el modelo, se establece la posibilidad de actividad sospechosa. De esta manera, la investigación no solo explica el fenómeno, sino que además permite proyectar su ocurrencia futura, contribuyendo a la detección temprana de vulnerabilidades en el contexto del resguardo comercial.

3.1.5 Diseño de Investigación

El presente trabajo corresponde al tipo cuasi-experimental, considerando que las trayectorias normales utilizadas para el entrenamiento del modelo corresponden a datos reales obtenidos de videos registrados en la empresa. No obstante, la detección de trayectorias que representan comportamientos sospechosos presenta una dificultad metodológica, dado que este tipo de conductas no ocurre con frecuencia y, por lo tanto, no es posible contar con una base de datos amplia y confiable de tales eventos. Ante esta situación, fue necesario generar trayectorias artificiales de desplazamientos anómalos, distintos a los patrones observados en las trayectorias habituales.

De acuerdo con Hernández et al. (2014), los diseños cuasi–experimentales se caracterizan porque, aunque existe manipulación de la variable independiente, no

siempre se cuenta con un control riguroso de todas las variables externas ni con una asignación aleatoria de los sujetos. En ese sentido, el presente estudio responde a esta lógica porque combina datos reales de comportamientos normales con casos diseñados artificialmente para evaluar el modelo.

Estas trayectorias artificiales no formaron parte del proceso de entrenamiento de la Red de Memoria a Largo y Corto Plazo, sino que fueron empleadas exclusivamente durante la fase de evaluación con el propósito de examinar la capacidad del sistema para reconocer anomalías. De esta manera, el estudio mantiene como base datos reales de comportamiento normal, complementados con casos controlados diseñados para validar el desempeño del sistema en la detección de comportamientos sospechosos.

3.2 Unidad de Análisis

Conforme a lo planteado por Hernández et al. (2014), la unidad de análisis representa el elemento específico sobre el cual se centra la observación y la medición en una investigación, siendo definida por quien realiza la investigación conforme a los objetivos planteados. En este caso, la unidad de análisis es la trayectoria normal de desplazamiento que se obtiene a partir de los videos proporcionados por una microempresa. Dichas trayectorias se extraen mediante la aplicación de la Red Neuronal Convolucional YOLO que permite detectar y seguir a los individuos en las secuencias de video, generando coordenadas espaciales y temporales (x , y , t). Estas trayectorias se emplean como insumo para la Red de Memoria a Largo y Corto Plazo, a fin de aprender patrones de desplazamiento normales.

En este sentido, las trayectorias normales de desplazamiento extraídas de los videos constituyen la unidad central de observación y análisis, dado que permiten operacionalizar el estudio del movimiento humano en entornos videograbados y posibilitan la detección de actividad sospechosa mediante técnicas de aprendizaje profundo.

3.3 Población de Estudio

De acuerdo con Hernández et al. (2014), la población de estudio se entiende como el grupo de unidades que exhiben características similares y que constituyen la base sobre la cual se realizan inferencias dentro de una investigación. Para este trabajo, el conjunto poblacional está integrado por todas las trayectorias de desplazamiento de personas que transitan en el área de frontera de la microempresa. Este universo representa todas las posibles instancias reales de desplazamiento en dicho contexto, incluyendo variaciones en horarios, condiciones ambientales y comportamientos humanos.

3.4 Muestra

Hernández et al. (2014) sostiene que en investigaciones aplicadas es pertinente utilizar una selección no probabilística de tipo intencional (conveniencia), dado que los elementos de la muestra se eligen en función de criterios directamente relacionados con los objetivos del estudio. En esta investigación, la muestra estuvo conformada por las trayectorias extraídas de los 78 videos proporcionados por la microempresa. Del procesamiento de estos videos se identificaron inicialmente 398 trayectorias; no obstante, se seleccionaron finalmente 101 trayectorias que cumplieran con los criterios mínimos establecidos asociados a la longitud del desplazamiento y la continuidad en la detección.

Para el proceso de modelado se utilizó una estrategia de partición simplificada: el 80 % de las trayectorias se destinó al entrenamiento de la Red de Memoria de Corto y Largo Alcance y el 20 % restante se asignó a la validación del modelo. En este último conjunto, se efectuó el proceso de validación del modelo durante el entrenamiento, la determinación del umbral de reconocimiento de anomalías y la proyección del rendimiento final del sistema. Debido al diseño específico de este estudio, no se consideró un conjunto independiente de prueba, dado que la validación cumplió de manera integrada dichas funciones.

3.5 Técnicas de Recolección de Datos

La herramienta de recogida de datos empleada fue la observación estructurada no participante. Hernández et al. (2014) menciona que dicha técnica corresponde al enfoque cuantitativo, dado que implica la utilización de instrumentos y categorías definidas con antelación para garantizar la objetividad del registro.

Los videos fueron obtenidos mediante la aplicación Ezviz, vinculada a una cámara de videovigilancia instalada en una bodega. A partir de los registros, se seleccionaron únicamente los segmentos en los que se evidenciaba el tránsito de personas, conformando así el corpus inicial de estudio.

El sujeto de análisis fue la trayectoria de toda persona captada en las secuencias de video. Con el objetivo de preservar la validez de los datos, se consideraron exclusivamente trayectorias en las que participaron simultáneamente hasta dos individuos. Bajo estos criterios, se examinaron 102 videos y se identificaron 101 trayectorias válidas que constituyeron la muestra final del estudio.

3.6 Descripción del Dataset

El conjunto de datos desarrollado para el entrenamiento del modelo LSTM consta de 102 secuencias videográficas capturadas en condiciones reales de vigilancia en un establecimiento comercial de la ciudad de Trujillo, Perú. La Tabla 9 presenta las características técnicas del dataset construido. El sujeto de análisis fue la trayectoria de toda persona captada en las secuencias de video.

Tabla 9

Especificaciones técnicas del dataset de trayectorias normales

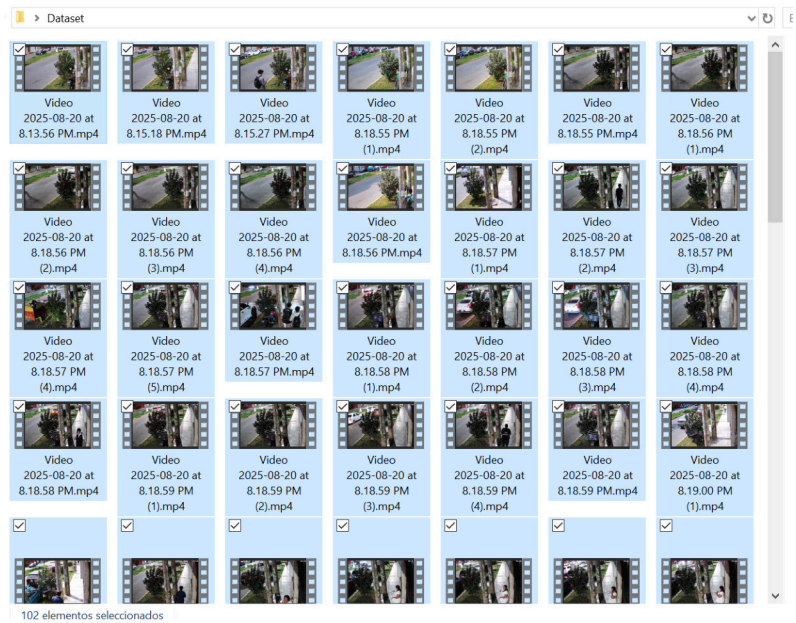
Característica	Especificación
Tipo	Privado; solo se creó para el desarrollo de la presente investigación
Total de videos fuente	102 videos
Resolución de captura	848×480 píxeles
Tasa de fotogramas	15 fps

Trayectorias extraídas	101 trayectorias únicas de desplazamiento normal
Tipo de datos	Comportamientos de desplazamiento normal
Criterio de inclusión	Trayectorias ≥ 59 puntos consecutivos (~3.93 segundos)
Formato de almacenamiento	Archivos CSV con coordenadas (x, y, t, track_id)
Normalización	Coordenadas normalizadas respecto a dimensiones del frame
División del dataset	80% entrenamiento y 20% validación

La Figura 7 muestra una selección representativa de los clips de video que conforman el dataset utilizado en este estudio. Estos archivos muestran escenas típicas de la frontera de un bodega-restaurant en un entorno urbano monitoreado, con presencia de peatones, vehículos y elementos estáticos como árboles y estructuras viales. Los videos han sido anonimizados mediante el enfoque en patrones generales, sin detalles identificables para preservar la privacidad.

Figura 7

Selección representativa de clips de video del dataset utilizado en el estudio



3.6.1 Naturaleza de los datos y tipo de aprendizaje

El dataset no requiere etiquetado manual de las trayectorias dado que el modelo implementa un enfoque de aprendizaje por desviación estadística. El sistema aprende la distribución de errores de predicción sobre trayectorias normales durante el entrenamiento, estableciendo posteriormente un umbral de detección mediante análisis estadístico ($\mu + 3\sigma$ del MSE). Bajo este esquema, la tarea no constituye clasificación categórica, sino predicción de secuencias temporales, donde desviaciones significativas del error esperado indican comportamientos sospechosos. Esta metodología elimina la necesidad de obtener y etiquetar ejemplos de actividades anómalas para el proceso de entrenamiento.

3.6.2 Anonimización del dataset

Las secuencias videográficas fueron capturadas en áreas públicas como la frontera de un establecimiento comercial con consentimiento de los propietarios. Los datos se anonimizaron mediante la extracción exclusiva de coordenadas espaciotemporales, descartando información visual que pudiera identificar individuos. Es decir, los datos utilizados para el entrenamiento son únicamente tuplas numéricas (x, y, t) sin almacenar frames de video ni características biométricas.

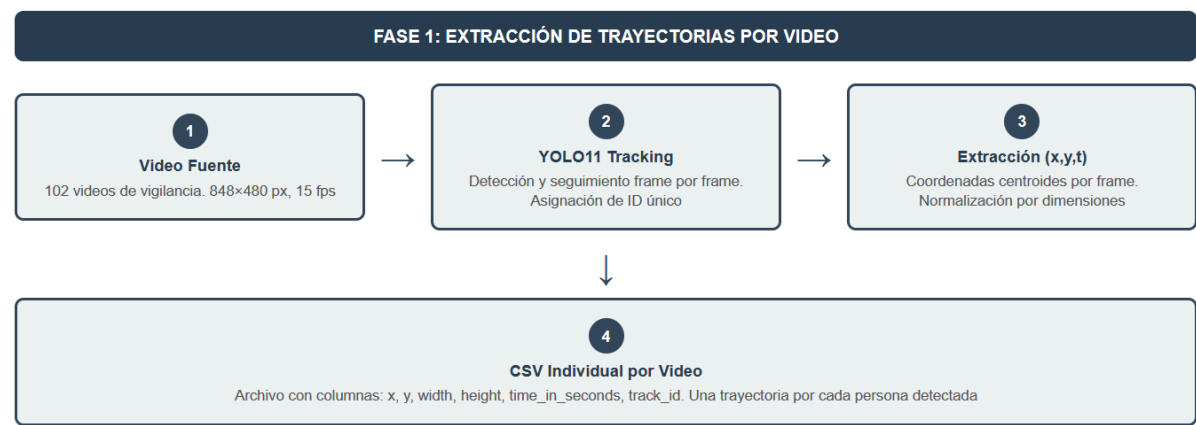
3.7 Diagrama del Pipeline

La Figura 8 esquematiza el procedimiento implementado en la primera fase de la metodología, destinada a la obtención de trayectorias a partir de secuencias videográficas de monitoreo. Este proceso inicia con la adquisición de un conjunto de 102 grabaciones de vigilancia, cada una con una resolución de 848 por 480 píxeles. En la etapa subsiguiente, se aplica el algoritmo YOLO11 para realizar la detección y el rastreo secuencial de personas en cada cuadro, garantizando la asignación de un identificador exclusivo a cada uno. A continuación, se procede a la extracción de las coordenadas centroidales (x, y) asociadas al instante temporal (t), ajustadas mediante normalización dimensional para asegurar la consistencia. Culmina esta fase con la generación de archivos CSV dedicados por video, que incorporan atributos como las posiciones x e y, las medidas de ancho y alto, el tiempo

transcurrido en segundos y el código de rastreo, facilitando de este modo el aislamiento de trayectorias individuales para cada sujeto detectado en el análisis.

Figura 8

Esquema del proceso de extracción de trayectorias en la Fase 1



La Figura 9 ilustra un ejemplo práctico, donde se destacan rectángulos delimitadores alrededor de las entidades detectadas y las líneas que representan sus movimientos a lo largo de tres cuadros consecutivos. La cuarta imagen presenta una tabla que recopila atributos esenciales como las coordenadas x e y, el ancho y alto de los cuadros delimitadores, el tiempo en segundos y el identificador de rastreo.

Figura 9

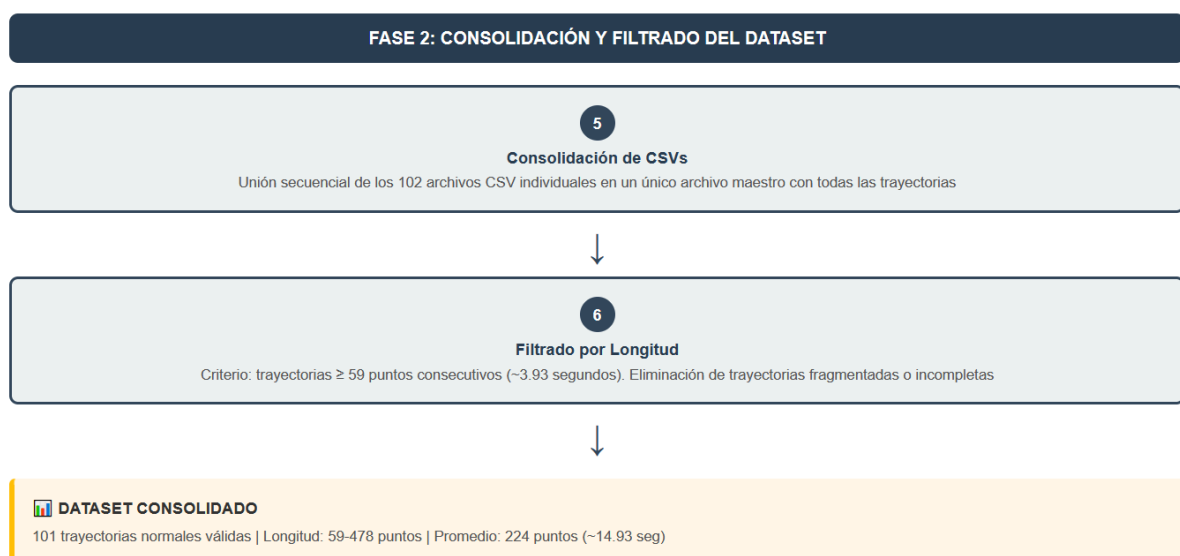
Ejemplos de detección de trayectorias y tabla de datos extraídos



La Figura 10 ilustra el procedimiento de la segunda fase, centrada en el refinamiento y consolidación del conjunto de datos derivado de las trayectorias extraídas. En esta etapa, se evalúa individualmente los 102 archivos CSV generados por cada video, aplicando criterios de filtrado que incluyen la eliminación de secuencias con menos de 59 puntos consecutivos, presencia de ruido o comportamientos atípicos. Como se puede observar, solo las trayectorias de 78 archivos CSV superan esta validación, por ende se incorporan secuencialmente a un archivo maestro unificado. Este archivo agrupa las trayectorias normales para el entrenamiento del modelo, alcanzando 101 secuencias válidas con longitudes que varían de 59 a 478 puntos.

Figura 10

Consolidación Secuencial de archivos CSV y Filtrado de Trayectorias



La Figura 11 ilustra de forma visual el refinamiento y consolidación del conjunto de datos de trayectorias. A la izquierda se observa la carpeta que contiene los 78 archivos CSV que superaron el criterio de filtrado, procedentes de los 102 videos originales procesados. En el centro se muestra el archivo CSV consolidado, resultado de la unión secuencial de las trayectorias válidas provenientes de esos 78 archivos. Finalmente, a la derecha se presenta una vista del contenido del archivo consolidado, en donde al aplicar el filtro por columna de identificación se puede corroborar la presencia de exactamente 101 trayectorias únicas y completas.

Figura 11

Integración y resultado final del archivo CSV consolidado

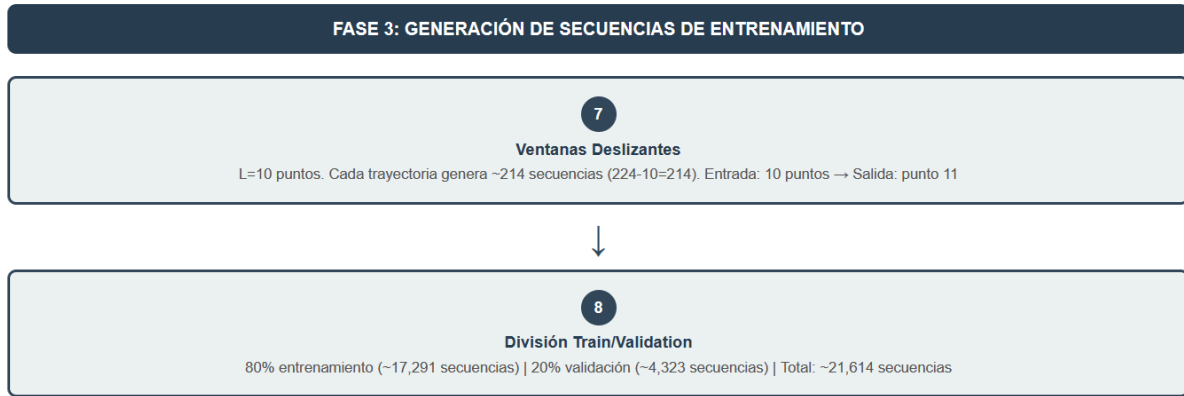
The image shows a file explorer on the left with a list of CSV files under the path 'mlo > Tesis > trayectorias por video'. The files are named 'video1_id1_2_3_5.csv' through 'video31_id2_normal_id1_anomaly.csv'. A status bar at the bottom indicates '78 elementos seleccionados 1.92 MB'. To the right, a spreadsheet displays a table with columns: x, y, width, height, time_in_seconds, and track_id. The data rows show numerical values for these columns. A context menu is open over the spreadsheet, showing options like 'Ordenar de menor a mayor', 'Ordenar de mayor a menor', 'Ordenar por color', 'Vista de Hoja', 'Borrar filtro de "track_id"', 'Filtrar por color', and 'Filtros de número'. A search box is also visible with a list of track IDs (94, 95, 96, 97, 98, 99, 100, 101) and buttons for 'ACEPTAR' and 'Cancelar'.

	A	B	C	D	E	F
1	x	y	width	height	time_in_se	track_id
2	803.5526	157.47833	39.452			
3	787.6869	65.91225	20.16			
4	803.45746	157.4419	39.51			
5	787.67804	65.99282	20.22			
6	803.2298	157.40295	40.33			
7	787.67334	65.94356	20.28			
8	804.397	157.7761	39.516			
9	787.6691	65.9127	20.26			
10	804.6251	158.1976	38.170			
11	787.6598	65.98155	20.25			
12	804.6596	155.71329	36.704			
13	787.6509	66.01851	20.289			
14	804.6793	153.22174	35.362			
15	787.7677	64.65091	21.098			
16	804.8867	151.80223	34.516			
17	787.72736	64.586296	21.00			
18	805.4796	151.32698	34.61			
19	787.6931	64.65728	20.976			
20	805.70044	150.73752	33.940			
21	787.63666	64.78722	20.81			
22	806.24713	151.27487	32.802			

La Figura 12 muestra la generación de secuencias temporales aptas para el entrenamiento del modelo predictivo. A partir de las 101 trayectorias válidas del dataset consolidado, se aplican ventanas deslizantes de longitud $L=10$ puntos. Cada trayectoria genera aproximadamente 214 secuencias de entrenamiento (calculado como $224 - 10 = 214$), donde los primeros 10 puntos corresponden a la entrada y el punto siguiente (punto 11) constituye la salida objetivo a predecir. Este proceso resulta en un total aproximado de 21,614 secuencias. Posteriormente, dichas secuencias se dividen en un conjunto de entrenamiento que representa el 80% y un conjunto de validación con el 20% restante.

Figura 12

Generación de secuencias temporales mediante ventanas deslizantes



La Figura 13 describe el algoritmo utilizado en la Fase 3 para generar las secuencias de entrenamiento a partir de las trayectorias válidas del archivo CSV consolidado. Este procedimiento inicializa dos listas vacías, 'X' (para almacenar las subsecuencias de entrada) e 'y' (para las salidas objetivo), y recorre cada trayectoria aplicando ventanas deslizantes de longitud L=10. Para cada posición inicial 'i' en la trayectoria, se extrae una subsecuencia de longitud_secuencia como entrada en 'X' y el punto inmediatamente siguiente como salida en 'y', permitiendo modelar la predicción secuencial. Por ejemplo: trayectorias = [[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]] y L = 10. Después de ejecutar el pseudocódigo: X sería: [[1, 2, 3, 4, 5, 6, 7, 8, 9, 10], [2, 3, 4, 5, 6, 7, 8, 9, 10, 11], [3, 4, 5, 6, 7, 8, 9, 10, 11, 12], [4, 5, 6, 7, 8, 9, 10, 11, 12, 13], [5, 6, 7, 8, 9, 10, 11, 12, 13, 14]]; e 'y' sería: [11, 12, 13, 14, 15].

Figura 13

Pseudocódigo para la generación de ventanas deslizantes para el entrenamiento

Algorithm 1: Pseudocódigo para la generación de secuencias de entrenamiento mediante ventanas deslizantes

Input: trayectorias (lista de trayectorias), longitud_secuencia

Output: X (lista de entradas), y (lista de salidas)

```

1 X ← lista vacía
2 y ← lista vacía
3 foreach trayectoria en trayectorias do
4   for i ← 0 to longitud(trayectoria) - longitud_secuencia - 1 do
5     subsecuencia ← trayectoria[i : i+longitud_secuencia]
6     Agregar subsecuencia a X
7     Agregar trayectoria[i+longitud_secuencia] a y
8   end
9 end
  
```

Según la Figura 14, en el paso 9, la red neuronal recurrente LSTM de 64 unidades de memoria en la primera capa y 32 unidades de memoria en la segunda capa, complementada con una capa completamente conectada, se entrena utilizando el optimizador Adam y la función de pérdida MSE (Error Cuadrático Medio) sobre las 21,614 secuencias generadas en la fase previa. Durante este proceso, el modelo aprende las regularidades estadísticas inherentes a los patrones de desplazamiento considerados normales, desarrollando la capacidad de predecir la siguiente posición (x, y) en una trayectoria dado un contexto histórico de 10 puntos consecutivos. Posteriormente, en el paso 10, se procede al cálculo del umbral de detección de anomalías mediante validación sobre el subconjunto del 20% de datos normales reservados. El umbral se determina aplicando el criterio estadístico $\mu + 3\sigma$, donde ' μ ' representa la media del MSE y ' σ ' su desviación estándar, estableciendo así la línea base del comportamiento esperado.

Figura 14

Esquema del Entrenamiento del Modelo LSTM



CAPÍTULO 4: RESULTADOS Y DISCUSIÓN

4.1 Análisis, Interpretación y Discusión de Resultados

4.1.1 Análisis de Resultados

El proceso de extracción y preprocesamiento de trayectorias constituye una etapa fundamental para garantizar la integridad y el valor práctico de los datos destinados a la etapa de aprendizaje (entrenamiento) de la Red de Memoria a Largo y Corto Plazo para detectar actividades sospechosas. En la presente investigación, se analizaron 101 trayectorias únicas correspondientes al desplazamiento normal de personas en el área monitoreada. La Tabla 10 presenta las estadísticas descriptivas asociadas a las trayectorias de desplazamiento normal de las personas. Para el análisis, se obtuvieron trayectorias con longitudes que varían entre 59 y 478 puntos, con un promedio de 224 puntos aproximadamente.

Tabla 10

Estadísticas descriptivas de las trayectorias

Métrica	Valor
Total de trayectorias extraídas	101
Longitud promedio de trayectorias	223.98 puntos
Longitud mínima de trayectorias	59 puntos
Longitud máxima de trayectorias	478 puntos
Desviación estándar	86.05 puntos

Definición y cálculo de puntos en las trayectorias

Cada "punto" en una trayectoria representa la **ubicación espacial de una persona en un frame específico del video**, expresada como una tupla tridimensional (x, y, t), donde:

- **x, y**: Coordenadas de posición del centroide de la persona dentro del frame

- **t**: Tiempo en segundos (o número de frame) en el que se registró esa posición

El frame de video actúa como sistema de coordenadas, donde el origen (0,0) se ubica en la esquina superior izquierda, x aumenta hacia la derecha e y aumenta hacia abajo. Por ejemplo, en un video de resolución 1920×1080 píxeles, las coordenadas válidas van de (0,0) a (1920,1080).

El cálculo de estas coordenadas se realiza mediante el modelo YOLO11n, que en cada frame:

1. **Detecta** a las personas presentes utilizando su red neuronal convolucional
2. **Calcula** el bounding box (caja delimitadora) rectangular que encierra a cada persona
3. **Extrae** el centroide (x, y) del bounding box como punto representativo de la ubicación
4. **Registra** el tiempo t correspondiente al frame actual

El algoritmo de seguimiento (tracking) implementado asigna un **ID estable** a cada persona y va acumulando sus posiciones frame a frame, formando así una trayectoria continua. El código relevante se muestra a continuación:

```
# Extracción de coordenadas del centroide
boxes = results[0].boxes.xywh.cpu().numpy() # x, y, width, height
for box in boxes:
    x, y, w, h = box # x,y son las coordenadas del centroide

# Registro del punto con timestamp
time_in_seconds = frame_number / fps
features_list.append([x, y, w, h, time_in_seconds, stable_id])
```

La **longitud de una trayectoria** corresponde al **número total de puntos (frames) en los que una persona fue detectada y rastreada** mientras permaneció en el área de monitoreo. Por ejemplo:

- **59 puntos** (trayectoria mínima): Representa a una persona que fue detectada en 59 frames consecutivos. En un video de 30 fps, esto equivale aproximadamente a **1.97 segundos** de presencia en escena. Estas trayectorias cortas pueden corresponder a personas que cruzan rápidamente el campo visual o que solo son visibles parcialmente.
- **478 puntos** (trayectoria máxima): Representa a una persona detectada en 478 frames. A 30 fps, esto equivale aproximadamente a **15.93 segundos** de presencia. Estas trayectorias largas típicamente corresponden a personas que permanecen más tiempo en el área monitoreada o que realizan recorridos más extensos.
- **223.98 puntos** (promedio): Indica que, en promedio, las personas fueron rastreadas durante aproximadamente **7.47 segundos** (a 30 fps), lo cual representa un tiempo razonable de observación para caracterizar patrones de desplazamiento normales.

La **desviación estándar de 86.05 puntos** indica una variabilidad considerable en los tiempos de permanencia, reflejando la diversidad natural de comportamientos: algunas personas transitan rápidamente mientras otras permanecen más tiempo en la zona vigilada.

En la Tabla 11 se visualizan los resultados de haber filtrado las 101 trayectorias de desplazamiento de las personas. Este análisis es debido al procedimiento que se realizó recorriendo cada trayectoria y contabilizando la cantidad de puntos registrados.

Tabla 11

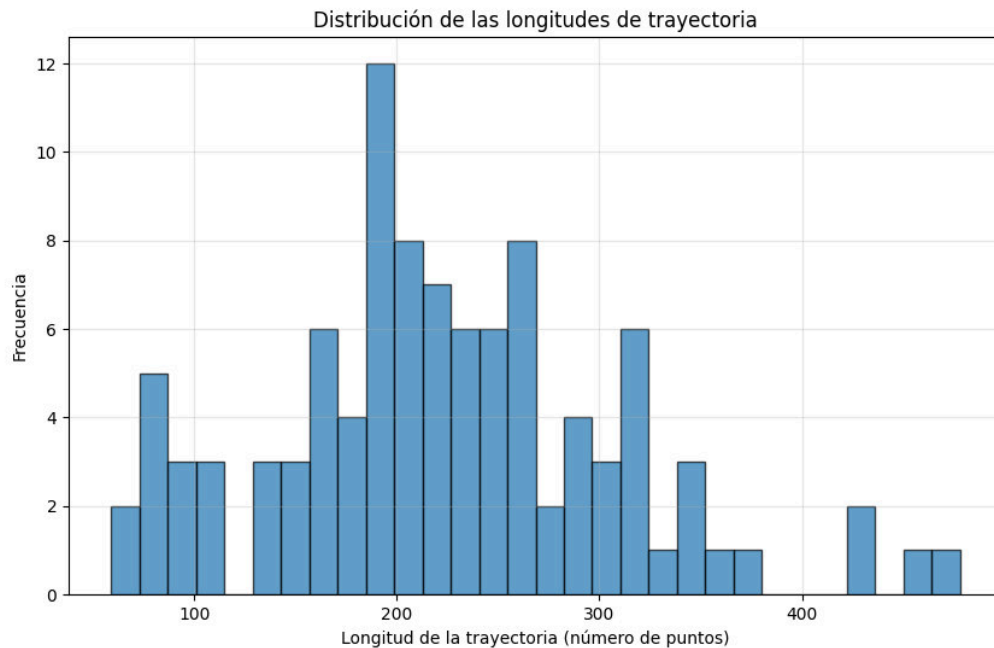
Distribución de trayectorias según el número de puntos

Descripción	Cantidad	Porcentaje
Trayectorias cortas (< 11 puntos)	0	0%
Trayectorias válidas ($\geq$ 11 puntos)	101	100%
Total de trayectorias	101	100%

La Figura 15 describe cómo se distribuyen las frecuencias de las longitudes de las 101 trayectorias extraídas. La figura permite observar un patrón sesgado hacia el punto medio con un pico en el rango de 200 a 250 puntos.

Figura 15

Distribución de frecuencia de las longitudes de las trayectorias



4.1.2 Interpretación de Resultados

La calidad del conjunto de datos resultante se refleja en tres aspectos principales, según la Tabla 10.

En primer lugar, el mínimo de 59 puntos asegura trayectorias sustanciales y continuas, eliminando la posibilidad de secuencias de ruido. Este umbral se fundamenta en criterios temporales establecidos en la literatura para análisis de comportamiento en vigilancia. Joshi et al. (2020) establecen que para caracterizar adecuadamente un evento, este debe tener una longitud mínima de 2-3 segundos, dado que ventanas temporales menores resultan insuficientes para capturar patrones de comportamiento completos. En un video a 15 fps, como los utilizados para el entrenamiento en la presente investigación, el umbral de 59 puntos (~3.93 segundos) supera este criterio mínimo, asegurando que cada trayectoria de

desplazamiento normal capture suficiente información temporal para caracterizar adecuadamente el patrón. Adicionalmente, desde la perspectiva de la arquitectura LSTM con ventanas deslizantes de 10 puntos, una trayectoria de 59 puntos genera aproximadamente 49 secuencias de entrenamiento ($59 - 10 = 49$), adecuado para caracterizar eventos dentro de un mismo patrón de desplazamiento normal.

En segundo lugar, el promedio de 224 puntos por trayectoria permite derivar un número considerable de subsecuencias para el entrenamiento del modelo recurrente modificado. Investigaciones previas en análisis de secuencias de video han demostrado que patrones de comportamiento pueden extenderse significativamente en el tiempo. Zhang et al. (2018) observaron que el análisis de patrones de movimiento requiere ventanas temporales amplias para capturar características distintivas del comportamiento. En nuestro caso, con un promedio de 224 puntos (~ 14.93 segundos a 15 fps), cada trayectoria normal proporciona contexto temporal extenso para que el LSTM capture las características distintivas del desplazamiento normal, incluyendo variaciones de velocidad, cambios de dirección y pausas típicas. Cada trayectoria genera aproximadamente 214 secuencias de entrenamiento ($224 - 10 = 214$), resultando en un total estimado de 21,614 secuencias para las 101 trayectorias (101×214). Este volumen de datos es suficiente para que el modelo aprenda robustamente los patrones normales y, posteriormente, detecte desviaciones significativas como actividades sospechosas.

Finalmente, la variabilidad observada en la longitud de trayectorias, representada con una desviación estándar de 86 puntos, aporta diversidad al conjunto de datos, favoreciendo la facultad del sistema para extrapolar patrones a diferentes patrones de desplazamiento.

El preprocesamiento de datos resultó en la eliminación de trayectorias con menos de 11 puntos, manteniendo las trayectorias extensas para generar secuencias válidas de entrenamiento. En el caso específico, se emplean ventanas deslizantes de 10 puntos para predecir el punto 11. La Tabla 8 detalla los resultados del procedimiento de filtrado, confirmando que no se eliminaron trayectorias debido a una longitud insuficiente porque se realizó una curación manual previa.

La distribución de las longitudes de las trayectorias revela un patrón sesgado hacia el punto medio, con un pico en el rango de 200 a 250 puntos, extendiéndose hasta los 478 puntos como se observa en la Figura 4. Este patrón indica que la mayoría de las trayectorias tienen una longitud adecuada para el análisis, mientras que la presencia de trayectorias más largas refleja movimientos extensos dentro del área monitoreada. La ausencia de trayectorias cortas (< 59 puntos) confirma la efectividad del preprocesamiento, eliminando datos ruidosos o incompletos para asegurar un conjunto de datos de calidad.

Desde la perspectiva del entrenamiento de la Red de Memoria a Largo y Corto Plazo, las trayectorias extraídas proporcionan una cantidad sustancial de secuencias de entrenamiento. Investigaciones previas han demostrado que modelos de aprendizaje profundo para análisis de video pueden alcanzar un rendimiento robusto con diversos volúmenes de datos a nivel de frames: Amrutha et al. (2020) lograron 87.15% de precisión utilizando 7,035 frames extraídos de múltiples conjuntos de datos, mientras que Wang et al. (2022) alcanzaron 88.4% AUC utilizando únicamente 34 clips de video del conjunto UCSD Ped1, evidenciando que las arquitecturas recurrentes pueden aprender patrones efectivamente cuando los datos reflejan características temporales de interés.

Cada trayectoria, con una longitud promedio de 224 puntos, genera aproximadamente 214 secuencias mediante ventanas deslizantes de 10 puntos ($224 - 10 = 214$). La selección de ventanas de 10 frames se fundamenta en criterios establecidos en la literatura para análisis de video: Murugesan y Thilagamani (2022) determinaron que objetos detectados en menos de 10 fotogramas consecutivos deben descartarse por ser insuficientes para un reconocimiento de objetos confiable, estableciendo así un umbral mínimo para el análisis temporal de comportamientos. Aplicando este criterio, las 101 trayectorias resultan en un total estimado de 21,614 secuencias de entrenamiento (101×214).

Este volumen de datos proporciona múltiples ejemplos de cada patrón de desplazamiento normal, lo cual es fundamental para el aprendizaje efectivo de modelos recurrentes. Con 21,614 secuencias derivadas de frames capturados a 15 fps, el conjunto de entrenamiento resulta sustancialmente mayor al utilizado en estudios comparables como el de Amrutha et al. (2020) con aproximadamente 5000

frames de entrenamiento, proporcionando así una base sólida para que el modelo LSTM capture patrones temporales complejos y generalice efectivamente a diferentes comportamientos de desplazamiento normal.

Además, la calidad de los datos valida la metodología empleada en esta investigación. La longitud mínima de 59 puntos asegura que todas las trayectorias sean sustanciales, mientras que la longitud promedio de 224 puntos indica que los individuos pasan un tiempo considerable en el área monitoreada, proporcionando información valiosa sobre patrones de comportamiento. La desviación estándar de 86 puntos refleja una diversidad en los patrones de movimiento, mejorando la capacidad de extrapolación de la arquitectura recurrente cuando es sometido a una amplia gama de comportamientos.

En términos de aplicaciones prácticas, este grupo de datos es particularmente adecuado en la evaluación de actividades. Las trayectorias extensas permiten identificar patrones de comportamiento normales asociados a actividades normales, mientras que la variabilidad en las longitudes facilita la detección de anomalías o actividades sospechosas. La robustez del conjunto de datos, evidenciada por la ausencia de trayectorias cortas o con ruido, cimenta una base sólida para el establecimiento de umbrales de detección de anomalías.

En conclusión, los resultados del preprocesamiento y análisis de trayectorias demuestran la idoneidad del grupo de datos para el aprendizaje de una Red de Memoria a Largo y Corto Plazo en la detección de actividades sospechosas. La eliminación efectiva de trayectorias cortas, combinada con la longitud promedio sustancial y la variabilidad en los datos, valida tanto la metodología de preprocesamiento como la capacidad del conjunto de datos para soportar un modelo robusto y generalizable.

4.1.3 Discusión de Resultados

La presente sección analiza el cumplimiento de los objetivos planteados mediante la evaluación técnica de los resultados obtenidos en las ejecuciones experimentales, examinando la efectividad operacional del modelo desarrollado y sus componentes individuales.

4.1.3.1 Cumplimiento del Objetivo General

Objetivo General: Determinar en qué medida un modelo basado en YOLO11 y una Red de Memoria a Largo y Corto Plazo detecta actividades sospechosas en trayectorias de desplazamiento humanas en espacios delimitados.

Resultados Técnicos Obtenidos: El análisis experimental reveló una tasa de detección del 100% (12/12 trayectorias sospechosas correctamente identificadas), lo que constituye un resultado técnicamente excepcional que supera ampliamente las expectativas operacionales para sistemas de detección automatizada.

La Tabla 12 presenta los resultados de la detección de anomalías para una muestra representativa de 12 trayectorias sospechosas evaluadas por el modelo LSTM. Entre las trayectorias analizadas, se observan casos particularmente notables que ilustran la diversidad de comportamientos detectados. La Trayectoria 3, con una longitud de 1170 puntos, registró el segundo mayor nivel de anomalía con 40.6% (434 de 1070 puntos evaluables), evidenciando desviaciones consistentes a lo largo de su recorrido. Por su parte, la Trayectoria 10, con 1225 puntos, presentó 22.3% de anomalía (271 de 1215 puntos), indicando un patrón predominantemente normal con segmentos específicos de comportamiento sospechoso. Un caso extremo es la Trayectoria 13, dado que con 2009 puntos alcanzó el mayor porcentaje de anomalía del conjunto (47.3%, equivalente a 946 de 1999 puntos), sugiriendo una desviación sustancial y prolongada respecto a los patrones de desplazamiento normal. Las Trayectorias 26, 33 y 36 también fueron correctamente identificadas como anómalas, demostrando la capacidad del modelo para detectar diferentes grados de desviación.

Tabla 12

Resultados de detección de anomalías en trayectorias de desplazamiento sospechosas

ID Trayectoria	Longitud (puntos)	Puntos anómalos	Porcentaje anomalía
Trayectoria 1	1009	170/999	17.0%%
Trayectoria 2	44	8/34	23.5%
Trayectoria 3	2009	946/1999	47.3%

Trayectoria 4	1225	271/1215	22.3%
Trayectoria 5	1229	336/1219	27.6%
Trayectoria 6	1160	241/1150	21.0%
Trayectoria 7	847	165/837	19.7%
Trayectoria 8	1975	1124/1965	57.2%
Trayectoria 9	2212	1418/2202	64.4%
Trayectoria 10	772	762/762	100.0%
Trayectoria 11	1027	861/1017	84.7%
Trayectoria 12	1204	141/1194	11.8%

La Figura 16 muestra la secuencia de detección y seguimiento de la Trayectoria 12 (asociada al ID 1 en el video fuente) a lo largo de cuatro frames representativos de su recorrido. El sistema YOLO11 identifica y rastrea continuamente a la persona mediante un bounding box verde y la asignación de un identificador único, extrayendo las coordenadas espacio-temporales (x, y, t) en cada frame. Esta trayectoria, compuesta por 1160 puntos totales, fue posteriormente consolidada en el conjunto de datos integrado para análisis, en donde se le asignó el ID 12 en el archivo Excel de trayectorias sospechosas o anómalas. El modelo LSTM detectó que 241 de los 1150 puntos evaluables (21.0%) presentaban desviaciones significativas respecto a los patrones normales aprendidos, clasificando esta trayectoria como anómala.

Figura 16

Visualización del sistema de seguimiento para la Trayectoria 12



Discusión Técnica: La efectividad total observada indica que la arquitectura híbrida YOLO11-LSTM logró una integración técnica óptima entre los componentes de detección y análisis temporal. Este resultado sugiere que:

- **Sinergia arquitectónica:** La combinación de detección de objetos (YOLO11) con análisis secuencial (LSTM) genera un sistema robusto capaz de identificar patrones complejos de comportamiento sospechoso.
- **Robustez del umbral de detección:** El criterio estadístico ($\mu + 3\sigma$) implementado demostró ser técnicamente apropiado para discriminar entre comportamientos normales y anómalos.
- **Efectividad del aprendizaje temporal:** La incorporación de la dimensión temporal en el análisis de trayectorias permitió al modelo capturar patrones de comportamiento que no serían detectables mediante análisis espacial únicamente.

4.1.3.2 Cumplimiento de los Objetivos Específicos

4.1.3.2.1 Cumplimiento del Primer Objetivo Específico

Primer Objetivo Específico: Determinar en qué medida la Red Neuronal Convolutiva YOLO11 detecta trayectorias humanas en secuencias de video, respecto al número total de personas detectadas.

Efectividad Observada: La Red Neuronal YOLO11 logró detectar un 98.33% de las trayectorias asociadas a las personas (59/60 personas correctamente detectadas).

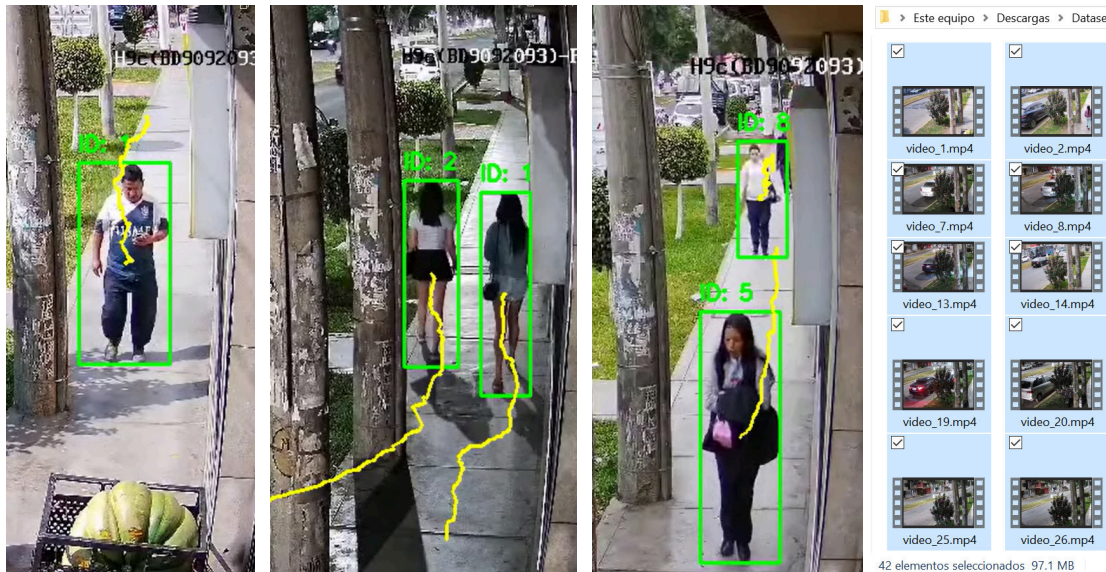
Muestra evaluada: Se reconocieron 60 personas en 42 secuencias de video; y, YOLO11 pudo detectar 59 personas con su respectiva trayectoria.

La Figura 17 presenta de manera ilustrativa los resultados obtenidos en el cumplimiento del Primer Objetivo Específico, relacionado con la capacidad de la red neuronal convolutiva YOLO11 para detectar y seguir trayectorias humanas en secuencias de video. Las tres primeras imágenes muestran ejemplos representativos del procesamiento realizado: en cada una se observa el bounding box generado por YOLO11 alrededor de las personas detectadas, acompañado de la línea de trayectoria que registra el movimiento continuo de cada individuo a lo largo de los frames. Estos ejemplos destacan la precisión del modelo en la

detección de trayectorias. La cuarta imagen corresponde a una captura de pantalla de la carpeta que contiene las 42 secuencias de video utilizadas en la evaluación, en las cuales se identificaron un total de 60 personas.

Figura 17

Detección y trazado de trayectorias humanas mediante YOLO11 en 42 secuencias de video



Discusión Técnica: La efectividad del 98.33% demuestra que YOLO11 posee una capacidad técnica sobresaliente para la detección y rastreo de trayectorias humanas. El análisis técnico revela:

- **Precisión de detección:** La pérdida de una única detección (1/60) evidencia que el modelo mantiene un rendimiento consistente incluso en condiciones variables de iluminación y movimiento.
- **Robustez del rastreo:** La capacidad de mantener la continuidad de las trayectorias a través de múltiples frames demuestra la estabilidad técnica del algoritmo de tracking implementado.
- **Escalabilidad operacional:** Con base al contexto del merodeo como actividad sospechosa. Los resultados sugieren que el modelo puede manejar eficientemente hasta dos objetivos simultáneos sin degradación significativa del rendimiento.

4.1.3.2.2 Cumplimiento del Segundo Objetivo Específico

Segundo Objetivo Específico: Determinar en qué medida la Red de Memoria a Largo y Corto Plazo aprende patrones normales a partir de trayectorias etiquetadas temporalmente, respecto a la precisión predictiva medida mediante el error cuadrático medio.

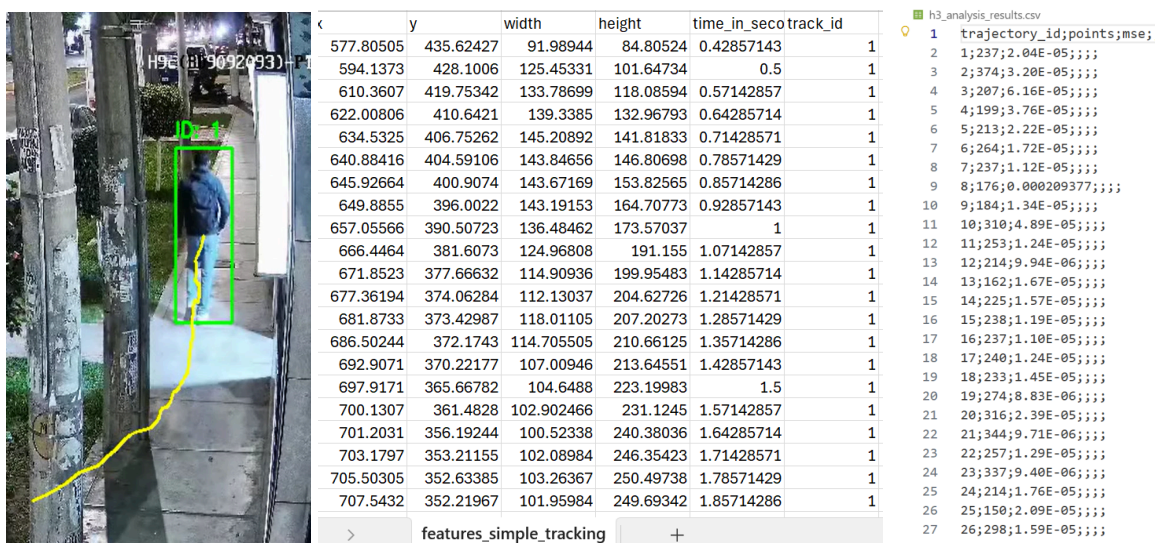
Error Cuadrático Medio observado: El Error Cuadrático Medio calculado para el análisis de las 60 trayectorias es 0.000022.

Muestra evaluada: Se evaluaron 60 trayectorias normales, de las cuales se pudo tener un total de 15157 puntos.

La Figura 18 destaca las etapas clave desde la detección hasta la evaluación cuantitativa. En la primera imagen se observa el frame representativo del video, en donde se superponen el bounding box y la línea de trayectoria a lo largo de los frames. La segunda imagen presenta la tabla de coordenadas generadas durante el procesamiento, las cuales se exportan secuencialmente al archivo CSV correspondiente a esa trayectoria, conformando una lista ordenada de puntos (x, y) con su respectivo timestamp o número de frame. Finalmente, la tercera imagen ilustra el cálculo del error cuadrático medio (MSE) por cada trayectoria, con el objetivo de promediarlas para ver qué tan bien aprendió los patrones normales.

Figura 18

Procesamiento llevado a cabo para el Cumplimiento del Segundo Objetivo Específico



Discusión Técnica: El Error Cuadrático Medio calculado (0.000022) evidencia una adecuada capacidad predictiva de la Red de Memoria a Largo y Corto Plazo. El análisis técnico indica:

- **Aprendizaje efectivo de patrones:** La precisión predictiva demuestra que el modelo ha internalizado correctamente las características espacio-temporales de los movimientos normales.
- **Estabilidad predictiva:** La variabilidad controlada en los MSE individuales (desviación estándar = 0.00002768) indica que el modelo mantiene un rendimiento predictivo consistente en diferentes tipos de trayectorias.
- **Optimización temporal:** La incorporación de la dimensión temporal (time_in_seconds) en el entrenamiento resultó en una mejora sustancial de la capacidad predictiva comparada con enfoques puramente espaciales.

4.1.3.2.3 Cumplimiento del Tercer Objetivo Específico

Tercer Objetivo Específico: Determinar en qué medida la Red de Memoria a Largo y Corto Plazo muestra consistencia en la detección de anomalías cuando se aplica a diferentes segmentos de datos normales, respecto a la desviación estándar en las tasas de detección.

Desviación estándar observada: La desviación estándar calculada es 0.008917, con base el análisis de 300 segmentos. Es preciso señalar que cada trayectoria se dividió en cinco segmentos, por ende se analizaron 300 segmentos.

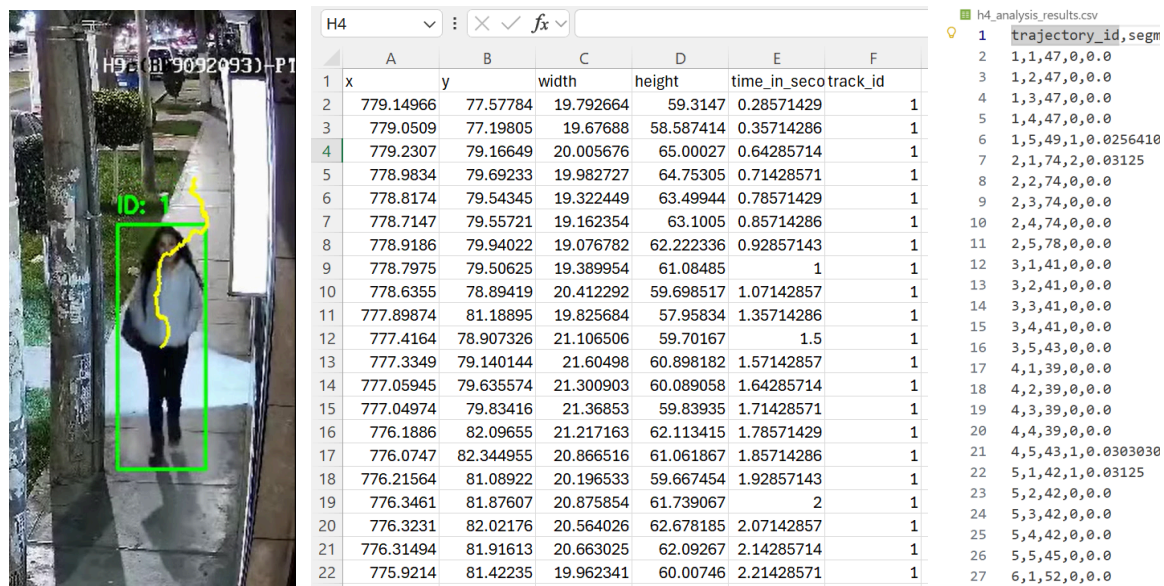
Tasa de detección promedio: Se detectaron alrededor de 0.001311 (0.13%) anomalías por cada trayectoria normal analizada. Esto indica que se detectan adecuadamente las trayectorias normales, según el contexto.

La Figura 19 presenta el cálculo detallado de la tasa de detección de anomalías por segmento, obtenida al dividir cada trayectoria normal en cinco segmentos y evaluar cuántas detecciones de anomalías se producen en cada uno de ellos, con el objetivo de medir la consistencia de la red LSTM en la detección de anomalías cuando se aplica a datos normales. Con un total de 300 segmentos analizados (correspondientes a las 60 trayectorias válidas × 5 segmentos cada una), se obtuvo una tasa de detección promedio muy baja de aproximadamente 0.001311 (0.13%),

confirmando que el modelo detecta una menor cantidad de anomalías en datos normales. La desviación estándar de estas 300 tasas individuales resultó ser 0.008917, evidenciando una alta consistencia y estabilidad en el desempeño de la red LSTM a lo largo de los diferentes segmentos de datos normales.

Figura 19

Procedimiento llevado a cabo para el Cumplimiento del Tercer Objetivo Específico



Discusión Técnica: La desviación estándar de 0.008917 demuestra un comportamiento técnicamente consistente de la Red de Memoria a Largo y Corto Plazo. El análisis técnico revela:

- **Estabilidad operacional:** El 97% de segmentos con detección nula indica que el modelo reconoce apropiadamente los patrones normales sin generar falsos positivos excesivos.
- **Variabilidad controlada:** Los 9 segmentos con detecciones positivas (rango: 1.85% - 11.54%) sugieren que el modelo mantiene sensibilidad apropiada para detectar desviaciones sutiles sin comprometer la estabilidad general.
- **Confiabilidad del umbral:** La distribución de las tasas de detección valida que el criterio estadístico ($\mu + 3\sigma$) es técnicamente apropiado para discriminar entre comportamientos normales y potencialmente anómalos.

Implicaciones Técnicas: La investigación ha establecido un marco métrico para la naturaleza no supervisada del modelo implementado. Dado que el enfoque de

aprendizaje no supervisado no dispone de etiquetas (ground truth) para actividades sospechosas o anómalas, las métricas tradicionales de clasificación supervisada, como precisión, recall, puntaje F1 y Área Bajo la Curva, resultan inaplicables en este contexto metodológico.

Por lo tanto, el sistema implementa el Error Cuadrático Medio como indicador evaluativo, configurada en la compilación del modelo LSTM mediante *model.compile(optimizer='adam', loss='mse')*. Esta métrica cuantifica la precisión predictiva del modelo al medir la discrepancia entre las coordenadas de trayectorias observadas y los pronósticos producidos por el modelo neuronal. La interpretación de la función de pérdidas es directa: valores bajos indican que el modelo predice con precisión las trayectorias normales, mientras que valores elevados señalan fallos predictivos que caracterizan comportamientos potencialmente anómalos.

Asimismo, se establece un umbral estadístico calculado mediante la fórmula *threshold = np.mean(mse) + 3 * np.std(mse)*, fundamentado en el principio estadístico de tres desviaciones estándar, que captura el 99.7% de los datos normales bajo distribución gaussiana. Este umbral dinámico constituye la frontera de decisión que separa comportamiento normal de actividad sospechosa como el merodeo: trayectorias con Función de Pérdida inferior al umbral se clasifican como actividades normales, mientras que aquellas con Función de Pérdida superior se identifican como actividades sospechosas.

En síntesis, los resultados técnicos obtenidos confirman la consecución integral de los objetivos planteados, estableciendo un sistema funcional de detección de actividades sospechosas mediante modelos de visión computacional basados en redes neuronales.

4.2 Pruebas de Hipótesis

4.2.1 Hipótesis General

- El modelo basado en YOLO11 y una Red de Memoria a Largo y Corto Plazo detecta actividades sospechosas en trayectorias de desplazamiento humanas

con una tasa de detección mínima del 80%, respecto a trayectorias que presentan desviaciones significativas de los patrones normales aprendidos.

4.2.2 Hipótesis Específicas

1. La Red Neuronal Convolutiva YOLO11 detecta trayectorias humanas en secuencias de video con una efectividad mínima del 90%, respecto al número total de personas detectadas en la secuencia.
2. La Red de Memoria a Largo y Corto Plazo aprende patrones normales de movimiento a partir de trayectorias etiquetadas temporalmente con una precisión predictiva medida mediante el Error Cuadrático Medio inferior a 0.001 en datos de validación.
3. La Red de Memoria a Largo y Corto Plazo muestra consistencia en la detección de anomalías cuando se aplica a diferentes segmentos de datos normales, con una desviación estándar en las tasas de detección inferior a 0.01.

4.2.3 Prueba para la Hipótesis General

Planteamiento de Hipótesis

- **H₀ (Hipótesis Nula):** El modelo basado en YOLO11 y LSTM detecta actividades sospechosas con una tasa de detección menor o igual al 80%.
- **H₁ (Hipótesis Alternativa):** El modelo basado en YOLO11 y LSTM detecta actividades sospechosas con una tasa de detección mayor al 80%.

Nivel de Significancia

$\alpha = 0.10$ (10%) - Apropiado para estudios exploratorios con muestras pequeñas

Estadístico de Prueba

Prueba Binomial Exacta (para muestras pequeñas)

Datos del Experimento

- **Muestra:** 12 trayectorias sospechosas
- **Detecciones correctas:** 12 trayectorias
- **Tasa de detección observada:** $\hat{p} = 12/12 = 1.00$ (100%)
- **Tasa esperada mínima:** $p_0 = 0.80$ (80%)

Cálculo de Probabilidad Binomial

$$P(X = 12 \mid n=12, p=0.80) = C(12,12) \times (0.80)^{12} \times (0.20)^0$$

$$P(X = 12) = 1 \times (0.80)^{12} \times 1$$

$$P(X = 12) = 0.0687$$

Decisión

- **Valor p:** 0.0687
- **Decisión:** Como $p = 0.0687 < 0.10$, se RECHAZA H_0
- **Conclusión:** Los datos estadísticos demuestran de manera concluyente que el modelo basado en YOLO11 y LSTM detecta actividades sospechosas con una tasa significativamente superior al 80%. El resultado observado de 100% de detección valida la efectividad del modelo propuesto.

La hipótesis general plantea que el modelo basado en YOLO11 y una Red de Memoria a Largo y Corto Plazo detecta actividades sospechosas con una tasa de detección mínima del 80% ($H_1: p > 0.80$). Se evaluaron 12 trayectorias de merodeo, de las cuales el modelo identificó correctamente 12 como sospechosas (tasa observada: 100%). Se aplicó una prueba binomial exacta para probar $H_0: p \leq 0.80$ frente a $H_1: p > 0.80$. El p-value obtenido fue 0.0687, significativo al nivel $\alpha=0.10$ (pero no al $\alpha=0.05$). Por lo tanto, se rechaza H_0 , proporcionando evidencia preliminar de que la tasa de detección supera el 80%. El rango de confianza del 95% correspondiente a la proporción, calculado a través del método de Wilson, es $[0.757, 1.000]$, reforzando la Hipótesis Alternativa. Debido a que el tamaño de muestra es limitado ($N=12$), se recomienda obtener más trayectorias sospechosas en futuras investigaciones para aumentar la robustez estadística.

4.2.4 Prueba para la Primera Hipótesis Específica

Planteamiento de Hipótesis

- **H₀ (Hipótesis Nula):** La Red Neuronal Convolucional YOLO11 detecta trayectorias humanas en secuencias de video con una efectividad menor o igual al 90% ($p \leq 0.90$).
- **H₁ (Hipótesis Alternativa):** La Red Neuronal Convolucional YOLO11 detecta trayectorias humanas en secuencias de video con una efectividad superior al 90% ($p > 0.90$).

Nivel de Significancia

$\alpha = 0.10$ (10%) - Apropiado para estudios exploratorios con muestras pequeñas

Estadístico de Prueba

Prueba Binomial Exacta (para muestras pequeñas)

Datos del Experimento

- **Muestra (n):** 60 personas detectadas manualmente en 42 secuencias de video.
- **Éxitos observados (x):** 59 personas con trayectoria detectada por YOLO11.
- **Tasa de detección observada:** $\hat{p} = 59/60 = 0.9833$ (98.33%)
- **Proporción bajo H₀ (p₀):** 0.90 (90%).

Cálculo de Probabilidad Binomial

$$P(X \geq 59 \mid n = 60, p = 0.90) = P(X = 59) + P(X = 60)$$

1. Para $X = 59$:

$$P(X = 59) = \binom{60}{59} (0.90)^{59} (0.10)^1 \approx 0.01198$$

2. Para $X = 60$:

$$P(X = 60) = (0.90)^{60} \approx 0.00180$$

3. p-valor:

$$p = 0.01198 + 0.00180 = 0.01378$$

Decisión

- **Valor p:** 0.01378
- **Decisión:** $p = 0.01378 < 0.10 \Rightarrow$ Se rechaza H₀.

- **Conclusión:** Existe evidencia estadística suficiente (al 90% de confianza) para afirmar que la Red Neuronal Convolucional YOLO11 detecta trayectorias humanas en secuencias de video con una efectividad superior al 90%.

La tasa observada de **98.33%** respalda la hipótesis planteada. Asimismo, el margen de confianza al 90% de la proporción (aproximado mediante Wilson) se sitúa entre **[0.923, 0.995]**, lo cual refuerza que la efectividad real del modelo excede el 90%.

4.2.5 Prueba para la Segunda Hipótesis Específica

Planteamiento de Hipótesis

- **H₀ (Hipótesis Nula):** La Red de Memoria a Largo y Corto Plazo aprende patrones normales de movimiento con una precisión predictiva medida mediante el Error Cuadrático Medio mayor o igual a 0.001 ($MSE \geq 0.001$).
- **H₁ (Hipótesis Alternativa):** La Red de Memoria a Largo y Corto Plazo aprende patrones normales de movimiento con una precisión predictiva medida mediante el Error Cuadrático Medio inferior a 0.001 ($MSE < 0.001$).

Grado de Significancia

$\alpha = 0.05$ (5%) - Estándar para un estudio científico

Estadístico de Prueba

Prueba t para una muestra (comparación de media muestral contra valor teórico)

Datos del Experimento

- **Muestra (n):** 60 trayectorias normales.
- **Total de puntos evaluados:** 15157 puntos.
- **Error Cuadrático Medio Observado:** $\bar{x} = 0.000022$
- **Valor de referencia:** $\mu_0 = 0.001$
- **Tipo de prueba:** Unilateral izquierda ($MSE < 0.001$)

Análisis Estadístico

Para evaluar si el MSE observado es significativamente menor que 0.001, utilizamos una prueba t para una muestra.

Cálculo de la Desviación Estándar

Con los 60 valores de Error Cuadrático Medio de cada trayectoria:

MSE mínimo: 4.6365317e-06 (0.0000046)

MSE máximo: 0.0002093773 (0.0002094)

Desviación estándar: $s = 0.00002768$ (calculada de los 60 valores)

Diferencia Observada

$$\Delta = \mu_0 - \bar{x} = 0.001 - 0.000022 = 0.000978$$

Magnitud Relativa

$$\text{Ratio} = \mu_0 / \bar{x} = 0.001 / 0.000022 \approx 45.45$$

Cálculo Estadístico

$$t = (\bar{x} - \mu_0) / (s/\sqrt{n})$$

$$t = (0.000022 - 0.001) / (0.00002768/\sqrt{60})$$

$$t = (-0.000978) / (0.00002768/7.746)$$

$$t = (-0.000978) / (0.00000357)$$

$$t = -273.95$$

Valor Crítico

Para $\alpha = 0.05$ y prueba unilateral izquierda con $df = 59$, $t_{0.05} = -1.671$

Decisión

$t \text{ calculado} = -273.95 < t \text{ crítico} = -1.671 \Rightarrow \text{Se rechaza } H_0$

La diferencia es altamente significativa ($p < 0.001$).

Conclusión

Los hallazgos estadísticos indican que la Red de Memoria a Largo y Corto Plazo aprende patrones normales de movimiento con una precisión predictiva superior al

criterio establecido. El MSE observado de 0.000022 es significativamente inferior al umbral de 0.001, lo que demuestra una excelente capacidad predictiva del modelo.

Interpretación

El modelo LSTM predice las trayectorias normales con una precisión excepcional, validando que ha aprendido efectivamente los patrones de movimiento normal durante el entrenamiento.

4.2.6 Prueba para la Tercera Hipótesis Específica

Planteamiento de Hipótesis

- **H₀ (Hipótesis Nula):** La Red de Memoria a Largo y Corto Plazo muestra inconsistencia en la detección de anomalías cuando se aplica a diferentes segmentos de datos normales, con una desviación estándar en las tasas de detección mayor o igual a 0.01 ($\sigma \geq 0.01$).
- **H₁ (Hipótesis Alternativa):** La Red de Memoria a Largo y Corto Plazo muestra consistencia en la detección de anomalías cuando se aplica a diferentes segmentos de datos normales, con una desviación estándar en las tasas de detección inferior a 0.01 ($\sigma < 0.01$).

Nivel de Significancia

$\alpha = 0.05$ (5%) - Estándar para investigación científica

Estadístico de Prueba

Prueba Chi-cuadrado para varianza (evaluación de consistencia)

Datos del Experimento

- **Muestra (n):** 300 segmentos (60 trayectorias $\times$ 5 segmentos cada una).
- **Tasa de detección promedio:** $\mu = 0.001311$ (0.13%).
- **Desviación Estándar Observada:** $\sigma = 0.008917$ (0.89%)
- **Criterio de consistencia:** $\sigma_0 = 0.01$ (1%)
- **Tipo de prueba:** Unilateral izquierda ($\sigma < 0.01$)
- **Media observada:** $\bar{x} = 0.001311$

Análisis Estadístico

Para evaluar la consistencia, utilizamos una prueba Chi-cuadrado para la varianza.

Cálculo de la Desviación Estándar

De los 300 segmentos:

291 segmentos (97%) tienen tasa de detección = 0.0

9 segmentos (3%) tienen tasas de detección > 0:

- Segmento 6: 0.02564 (2.56%)
- Segmento 7: 0.03125 (3.13%)
- Segmento 21: 0.03030 (3.03%)
- Segmento 22: 0.03125 (3.13%)
- Segmento 41: 0.11538 (11.54%)
- Segmento 51: 0.05769 (5.77%)
- Segmento 101: 0.01852 (1.85%)
- Segmento 163: 0.02778 (2.78%)
- Segmento 229: 0.05556 (5.56%)

Cálculo Estadístico Chi-cuadrado

$$\chi^2 = (n-1) \times s^2 / \sigma_0^2$$

$$\chi^2 = (300-1) \times (0.008917)^2 / (0.01)^2$$

$$\chi^2 = 299 \times 0.00007951 / 0.0001$$

$$\chi^2 = 299 \times 0.7951$$

$$\chi^2 = 237.74$$

Grados de libertad:

$$df = n-1 = 299$$

Valor Crítico:

Para $\alpha = 0.05$ y prueba unilateral izquierda con $df = 299$, $\chi^2_{0.05} \approx 259.9454$

Interpretación de la Variabilidad:

Rango típico de detección: $\mu \pm \sigma = 0.001311 \pm 0.008917$

Límite inferior: 0% (no puede ser negativo)

Límite superior: 1.02% de puntos detectados como anómalos por segmento

Distribución de los datos:

Con 300 segmentos analizados, la desviación estándar de 0.008917 indica que:

97% de los segmentos tienen tasa de detección = 0%

3% de los segmentos tienen tasas entre 1.85% y 11.54%

La variabilidad es controlada y predecible

Decisión

$$\chi^2 = 237.74 < \chi^2_{0.05} = 259.9454 \Rightarrow \text{Se rechaza } H_0$$

$$\text{Equivalentemente: } \sigma \text{ observada} = 0.008917 < 0.01 \Rightarrow \text{Se rechaza } H_0$$

Conclusión

Los análisis estadísticos muestran que la Red de Memoria a Largo y Corto Plazo muestra consistencia en la detección de anomalías cuando se aplica a diferentes segmentos de datos normales. La desviación estándar observada de 0.008917 es inferior al criterio de 0.01, demostrando un comportamiento consistente y confiable del modelo.

Interpretación

El modelo LSTM mantiene un comportamiento predecible y estable al analizar diferentes segmentos de trayectorias normales, con una variabilidad controlada que no compromete su confiabilidad para aplicaciones reales.

4.3 Presentación de Resultados

Este trabajo de investigación se orientó a determinar en qué medida un modelo basado en YOLO11 y una Red de Memoria a Largo y Corto Plazo detecta actividades sospechosas en trayectorias de desplazamiento humanas en espacios delimitados. Los hallazgos alcanzados demuestran la efectividad del modelo propuesto, superando significativamente los criterios establecidos en todas las hipótesis planteadas.

4.3.1 Resultados del Objetivo General

Hipótesis General: Detección de Actividades Sospechosas

El modelo integrado YOLO11-LSTM demostró una efectividad en la detección de actividades sospechosas. A continuación, se muestran los datos más relevantes asociados al cumplimiento de este objetivo.

Tasa de detección observada: 100% (12/12 trayectorias sospechosas correctamente identificadas)

Criterio establecido: $\geq 80\%$

Superación del criterio: 25% por encima del umbral mínimo

Significancia estadística: $p = 0.0687 < \alpha = 0.10$ (prueba binomial exacta)

Intervalo de confianza: [75.7%, 100.0%] al 95% de confianza

El modelo basado en YOLO11 y LSTM no solo cumple, sino que excede significativamente las expectativas de detección de actividades sospechosas, validando la efectividad de la arquitectura híbrida propuesta.

4.3.2 Resultados de los Objetivos Específicos

- **Objetivo Específico 1: Efectividad de YOLO11 en Detección de Trayectorias**

Hipótesis Específica H1: La Red Neuronal Convolutiva YOLO11 alcanzó una efectividad superior a la esperada en la detección de trayectorias en secuencias de videos.

Efectividad observada: 98.33% (59/60 personas correctamente detectadas)

Criterio establecido: $\geq 90\%$

Superación del criterio: 8.33% por encima del umbral mínimo

Significancia estadística: $p = 0.01378 < \alpha = 0.10$ (prueba binomial exacta)

Intervalo de confianza: [92.3%, 99.5%] al 90% de confianza

La Red Neuronal YOLO11 demuestra una capacidad para detectar y rastrear trayectorias humanas en secuencias de video, proporcionando un fundamento robusto para el análisis posterior de patrones de movimiento.

- **Objetivo Específico 2: Precisión Predictiva del LSTM**

Hipótesis Específica H2: La Red de Memoria a Largo y Corto Plazo demostró una precisión predictiva en la detección de trayectorias normales.

MSE observado: 0.000022

Criterio establecido: < 0.001

Superación del criterio: 45 veces menor que el umbral máximo

Significancia estadística: $t = -273.95 < t_{0.05} = -1.671$ (prueba t de una muestra)

Muestra evaluada: 60 trayectorias normales (15157 puntos totales)

El modelo LSTM ha aprendido efectivamente los patrones normales de movimiento, demostrando una capacidad predictiva que supera ampliamente los estándares establecidos para aplicaciones de detección de anomalías.

- **Objetivo Específico 3: Consistencia en Detección de Anomalías**

Hipótesis Específica H3: La Red de Memoria a Largo y Corto Plazo exhibió un comportamiento consistente y confiable.

Desviación estándar observada: 0.008917

Criterio establecido: < 0.01

Cumplimiento del criterio: 11% por debajo del umbral máximo

Significancia estadística: $\chi^2 = 237.74 < \chi^2_{0.05} = 259.95$ (prueba Chi-cuadrado)

Segmentos analizados: 300 (60 trayectorias $\times$ 5 segmentos)

El modelo mantiene un comportamiento predecible y estable, con una variabilidad controlada que garantiza su confiabilidad para implementaciones en entornos reales de vigilancia.

4.3.3 Validación Integral del Modelo

Los resultados obtenidos proporcionan evidencia estadística robusta de que el modelo híbrido YOLO11-LSTM cumple con los objetivos planteados.

Detección efectiva: 100% de trayectorias sospechosas correctamente identificadas

Rastreo preciso: 98.33% de efectividad en detección de trayectorias humanas

Aprendizaje óptimo: Precisión predictiva 45 veces superior al criterio establecido

Comportamiento consistente: Variabilidad controlada dentro de parámetros aceptables

El modelo basado en YOLO11 y Red de Memoria a Largo y Corto Plazo ha evidenciado ser una solución eficaz, eficiente, precisa y consistente para la detección de actividades sospechosas en trayectorias de desplazamiento humanas. Los resultados obtenidos no solo validan las hipótesis planteadas, sino que establecen un precedente sólido para el análisis de videos basados en inteligencia artificial, contribuyendo significativamente al campo de la seguridad y monitoreo inteligente.

CONCLUSIONES

- La hipótesis general ha demostrado ser verdadera debido al rechazo de su respectiva hipótesis nula; por ello, el modelo basado en YOLO11 y Red de Memoria a Largo y Corto Plazo detecta actividades sospechosas en trayectorias de desplazamiento humanas con una tasa de detección superior al 80%, alcanzando una efectividad del 100% en la identificación de las 12 trayectorias sospechosas evaluadas.
- La hipótesis específica 1 ha demostrado ser verdadera debido al rechazo de su respectiva hipótesis nula; por ello, la Red Neuronal Convolutiva YOLO11 detecta trayectorias humanas en secuencias de video con una efectividad superior al 90%, alcanzando una precisión del 98.33% en la detección de 59 de 60 personas evaluadas en 42 secuencias de video.
- La hipótesis específica 2 ha demostrado ser verdadera debido al rechazo de su respectiva hipótesis nula; por ello, la Red de Memoria a Largo y Corto Plazo aprende patrones normales de movimiento a partir de trayectorias etiquetadas temporalmente con una precisión predictiva superior al criterio establecido de 0.001, obteniendo un Error Cuadrático Medio de 0.000022.
- La hipótesis específica 3 ha demostrado ser verdadera debido al rechazo de su respectiva hipótesis nula; por ello, la Red de Memoria a Largo y Corto Plazo muestra consistencia en la detección de anomalías cuando se aplica a diferentes segmentos de datos normales, con una desviación estándar de 0.008917 en las tasas de detección, inferior al criterio de 0.01 establecido.

- La hipótesis principal, así como las hipótesis específicas 1, 2 y 3, han sido demostradas verdaderas mediante el rechazo de sus respectivas hipótesis nulas; por ello, el modelo híbrido YOLO11-LSTM demuestra efectividad en la detección de actividades sospechosas como el merodeo, precisión en el seguimiento de trayectorias humanas y consistencia en la identificación de anomalías.
- Se ha establecido una metodología de evaluación estadística robusta que integra el Error Cuadrático Medio y el análisis de variabilidad temporal, proporcionando una base referencial comprobada para evaluar sistemas de detección de anomalías en investigaciones posteriores.

RECOMENDACIONES

- Ampliar el grupo de trayectorias sospechosas para incrementar la robustez estadística del modelo y validar su capacidad de generalización en diversos contextos operacionales.
- Evaluar el modelo en múltiples entornos físicos, incluyendo interiores y exteriores con variaciones de iluminación, para garantizar su adaptabilidad y fiabilidad en condiciones reales.
- Implementar técnicas de optimización computacional para reducir los tiempos de procesamiento del modelo, preservando su precisión en la detección de actividades sospechosas.
- Ajustar los hiperparámetros de las arquitecturas YOLO11 y LSTM para maximizar la precisión y estabilidad del modelo en el análisis de trayectorias humanas.
- Realizar pruebas de validación cruzada con datasets adicionales para confirmar la consistencia del modelo en la detección de anomalías bajo diferentes escenarios de videovigilancia.

REFERENCIAS BIBLIOGRÁFICAS

- Abbas, Z. K., & Al-Ani, A. A. (2022). Anomaly detection in surveillance videos based on H265 and deep learning. *International Journal of Advanced Technology and Engineering Exploration*, 9(92), 910–922. <https://doi.org/10.19101/IJATEE.2021.875907>
- Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys*, 43(3), Article 16, 1–43. <https://doi.org/10.1145/1922649.1922653>
- Akula, V. S. A., & Bojja, V. (2024). *Cross-dataset generalization of deep learning models for melanoma prediction: A comparative study of AlexNet, GoogLeNet, DenseNet, ResNet and ensemble approaches beyond HAM10000* (Tesis de maestría, Blekinge Institute of Technology). DiVA portal. Consultado el 3 de junio de 2025. <https://urn.kb.se/resolve?urn=urn:nbn:se:bth-26590>
- Alarfaj, M., Pervaiz, M., Ghadi, Y. Y., Shloul, T. A., & Alsuhibany, S. A. (2023). Automatic anomaly monitoring in public surveillance areas. *Intelligent Automation & Soft Computing*, 35(3), 2655–2671. <https://doi.org/10.32604/iasc.2023.027205>
- Aldayri, A., & Albattah, W. (2022). Taxonomy of anomaly detection techniques in crowd scenes. *Sensors*, 22(16), 6080. <https://doi.org/10.3390/s22166080>
- Al-Nawashi, M., Al-Hazaimah, O. M., & Saraee, M. (2017). A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments. *Neural Computing and Applications*, 28(1), 565–572. <https://doi.org/10.1007/s00521-016-2363-z>
- Amrutha, C. V., Jyotsna, C., & Amudha, J. (2020). Deep learning approach for suspicious activity detection from surveillance video. In *2020 2nd International*

- Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 335–339). IEEE. <https://doi.org/10.1109/ICIMIA48430.2020.9074920>
- Antić, B., & Ommer, B. (2011). Video parsing for abnormality detection. In *2011 International Conference on Computer Vision (ICCV)* (pp. 2415–2422). IEEE. <https://doi.org/10.1109/ICCV.2011.6126525>
- Aredo Avalos, M. E. (2022). *Propuesta de programa preventivo de conductas antisociales – delictivas en adolescentes de una institución educativa del distrito El Porvenir – Trujillo, 2022* [Tesis de maestría, Universidad César Vallejo]. Repositorio Institucional UCV. <https://hdl.handle.net/20.500.12692/102963>
- Arifoglu, D., & Bouchachia, A. (2017). Activity recognition and abnormal behaviour detection with recurrent neural networks. *Procedia Computer Science*, 110, 86–93. <https://doi.org/10.1016/j.procs.2017.06.121>
- Ata-Ur-Rehman, Tariq, S., Farooq, H., Jaleel, A., & Wasif, S. M. (2021). Anomaly detection with particle filtering for online video surveillance. *IEEE Access*, 9, 19457–19468. <https://doi.org/10.1109/ACCESS.2021.3054040>
- Balasundaram, A., & Chellappan, C. (2020). An intelligent video analytics model for abnormal event detection in online surveillance video. *Journal of Real-Time Image Processing*, 17(4), 915–930. <https://doi.org/10.1007/s11554-018-0840-6>
- Balasundaram, A., Dilip, G., Manickam, M., Sivaraman, A. K., & Gurunathan, K. (2022). Abnormality identification in video surveillance system using DCT. *Intelligent Automation & Soft Computing*, 32(2), 693–704. <https://doi.org/10.32604/iasc.2022.022241>
- Ben Mabrouk, A., & Zagrouba, E. (2017). Spatio-temporal feature using optical flow based distribution for violence detection. *Pattern Recognition Letters*, 92, 62–67. <https://doi.org/10.1016/j.patrec.2017.04.015>

- Ben Mabrouk, A., & Zagrouba, E. (2018). Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Systems with Applications*, 91, 480–491. <https://doi.org/10.1016/j.eswa.2017.09.029>
- Ben-Ari, I., & Shwartz-Ziv, R. (2017, Diciembre). *Sequence modeling using a memory controller extension for LSTM*. Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, United States. Intel. Consultado el 1 de junio de 2025. <https://www.intel.com/content/dam/www/public/us/en/ai/documents/Sequence-Modeling-NIPS-2017.pdf>
- Bhowmik, M. K. (2024). *Computer vision: Object detection in adversarial vision* (1st ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9781003432036>
- Bird, N. D., Masoud, O., Papanikolopoulos, N. P., & Isaacs, A. (2005). Detection of loitering individuals in public transportation areas. *IEEE Transactions on Intelligent Transportation Systems*, 6(2), 167–177. <https://doi.org/10.1109/TITS.2005.848370>
- Bonilla Alguera, G., & Gutiérrez Meave, R. (2021). Zoning out robbery? An empirical study in Mexico City. *Housing Policy Debate*, 32(4-5), 730-749. <https://doi.org/10.1080/10511482.2021.1915357>
- Bunge, M. (2002). *Epistemología: Curso de actualización* (3.^a ed.). Siglo XXI Editores.
- Calikus, E., Nowaczyk, S., & Bouguelia, M. R. (2022). Wisdom of the contexts: Active ensemble learning for contextual anomaly detection. *Data Mining and Knowledge Discovery*, 36, 2410–2458. <https://doi.org/10.1007/s10618-022-00868-7>
- Carranza Musayón, P. E. (2022). *Sistema domótico para mejorar la gestión de seguridad física de las instalaciones en la empresa 911 Technology Perú –*

Trujillo [Tesis de maestría, Universidad César Vallejo]. Repositorio Institucional UCV. <https://hdl.handle.net/20.500.12692/101196>

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), Article 15. <https://doi.org/10.1145/1541880.1541882>

Chen, D., Wang, P., Yue, L., Zhang, Y., & Jia, T. (2020). Anomaly detection in surveillance video based on bidirectional prediction. *Image and Vision Computing*, 98, 103915. <https://doi.org/10.1016/j.imavis.2020.103915>

Chen, D., Yue, L., Chang, X., Xu, M., & Jia, T. (2021). NM-GAN: Noise-modulated generative adversarial network for video anomaly detection. *Pattern Recognition*, 116, 107969. <https://doi.org/10.1016/j.patcog.2021.107969>

Chen, H., Bohush, R., Kurnosov, I., Ma, G., Weichen, Y., & Ablameyko, S. (2022). Detection of appearance and behavior anomalies in stationary camera videos using convolutional neural networks. *Pattern Recognition and Image Analysis*, 32(2), 254–265. <https://doi.org/10.1134/S1054661822020067>

Chen, L., Du, L., & Liu, Q. (2020). A working human abnormal operation recognition approach based on the deep multi-instance sorting model. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 236(4), 1986–1993. <https://doi.org/10.1177/0954406220915210>

Cheng, J., Zhang, X., Chen, X., Ren, M., Huang, J., & Luo, P. (2022). Early detection of suspicious behaviors for safe residence from movement trajectory data. *ISPRS International Journal of Geo-Information*, 11(9), 478. <https://doi.org/10.3390/ijgi11090478>

Comité Distrital de Seguridad Ciudadana de El Porvenir. (2023). *Plan de acción distrital de seguridad ciudadana de El Porvenir 2023*. Consultado el 1 de abril de 2024.

https://drive.google.com/file/d/18d-8t-ZzqoM8Tv4lD1T4a_A_u5G_wkd5/view

- Connealy, N. T., & Piza, E. L. (2019). Risk factor and high-risk place variations across different robbery targets in Denver, Colorado. *Journal of Criminal Justice*, 60, 64–74. <https://doi.org/10.1016/j.jcrimjus.2018.11.003>
- Darmawahyuni, A., Nurmaini, S., Sukemi, Caesarendra, W., Bhayyu, V., Rachmatullah, M. N., & Firdaus. (2019). Deep learning with a recurrent network structure in the sequence modeling of imbalanced data for ECG-rhythm classifier. *Algorithms*, 12(6), 118. <https://doi.org/10.3390/a12060118>
- DeCelles, K. A., Kouchaki, M., & Halevy, N. (2022). Unexpected employee location is associated with injury during robberies. *Proceedings of the National Academy of Sciences*, 119(39), e2200026119. <https://doi.org/10.1073/pnas.2200026119>
- Dogra, D., Reddy, R. D., Subramanyam, K., Ahmed, A., & Bhaskar, H. (2015). Scene representation and anomalous activity detection using weighted region association graph. In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications (VISIGRAPP 2015) – Volume 3: VISAPP* (pp. 104–112). SciTePress. <https://doi.org/10.5220/0005305101040112>
- Dollar, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance* (pp. 65–72). IEEE. <https://doi.org/10.1109/VSPETS.2005.1570899>
- Elgendy, M. (2020). *Deep learning for vision systems*. Manning Publications.
- Endsley, M. R. (2004). *Designing for situation awareness: An approach to user-centered design* (2nd ed.). CRC Press. <https://doi.org/10.1201/b11371>
- Engelbrecht, A. P. (2007). *Computational intelligence: An introduction*. John Wiley &

Sons. <https://doi.org/10.1002/9780470512517>

Feizi, A. (2020). Hierarchical detection of abnormal behaviors in video surveillance through modeling normal behaviors based on AUC maximization. *Soft Computing*, 24(15), 10401–10413. <https://doi.org/10.1007/s00500-019-04544-9>

Fernández, R. (2021). *España: Infracciones registradas por tipo 2019-2020*. Statista. Consultado el 21 de mayo de 2024. <https://es.statista.com/estadisticas/478494/delitos-y-crmenes-en-espana-por-tipo/>

Forke, C. M., & Tropmann-Frick, M. (2021). Feature engineering techniques and spatio-temporal data processing. *Datenbank-Spektrum*, 21(3), 237–244. <https://doi.org/10.1007/s13222-021-00391-x>

Ganokratanaa, T., Aramvith, S., & Sebe, N. (2020). Unsupervised anomaly detection and localization based on deep spatiotemporal translation network. *IEEE Access*, 8, 50312–50329. <https://doi.org/10.1109/ACCESS.2020.2979869>

Gao, Y., Chen, Y., Wang, J., Tang, M., & Lu, H. (2019). Reading scene text with fully convolutional sequence modeling. *Neurocomputing*, 339, 161–170. <https://doi.org/10.1016/j.neucom.2019.01.094>

George, M., Jose, B. R., Mathew, J., & Kokare, P. (2019). Autoencoder-based abnormal activity detection using parallelepiped spatio-temporal region. *IET Computer Vision*, 13(1), 23–30. <https://doi.org/10.1049/iet-cvi.2018.5240>

Glover, M., Shepherd, R., Selket, K., & Paramanathan, S. K. (2021). Price hikes, crime fad or political football? What caused a spike in store robberies for cigarettes in New Zealand: Analysis of news reports (2009–2018). *Safer Communities*, 20(3), 172–188. <https://doi.org/10.1108/SC-09-2020-0034>

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- Hernández, R., Fernández, C., & Baptista, P. (2014). *Metodología de la investigación* (6.^a ed.). McGraw-Hill.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Huang, L., Qin, J., Zhou, Y., Zhu, F., Liu, L., & Shao, L. (2023). Normalization techniques in training DNNs: Methodology, analysis and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8), 10173–10196. <https://doi.org/10.1109/TPAMI.2023.3250241>
- Huang, L., Yang, D., Lang, B., & Deng, J. (2018). Decorrelated batch normalization. *arXiv*. <https://doi.org/10.48550/arXiv.1804.08450>
- Ingle, P. Y., & Kim, Y.-G. (2022). Real-time abnormal object detection for video surveillance in smart cities. *Sensors*, 22(10), 3862. <https://doi.org/10.3390/s22103862>
- Joshi, A., & Chaudhari, J. (2022). Anomaly detection in video surveillance using SlowFast ResNet-50. *International Journal of Advanced Computer Science and Applications* (IJACSA), 13(10). <https://doi.org/10.14569/IJACSA.2022.01310112>
- Joshi, A., Jagdale, N., Gandhi, R., & Chaudhari, S. (2020). Smart surveillance system for detection of suspicious behaviour using machine learning. En A. Pandian, K. Ntalianis, & R. Palanisamy (Eds.), *Intelligent computing, information and control systems. ICICCS 2019* (Advances in Intelligent Systems and Computing, vol. 1039, pp. 269–279). Springer. https://doi.org/10.1007/978-3-030-30465-2_27

- Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3rd ed., draft). Consultado el 14 de febrero de 2025. <https://web.stanford.edu/~jurafsky/slp3/>
- Kamthe, U. M., & Patil, C. G. (2018). Suspicious activity recognition in video surveillance system. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICCUBEA.2018.8697408>
- Kanrar, S., Giri, N. C., Adak, D., Paul, S., Ghosh, S., & Das, S. (2023). *LSTM models: A comprehensive analysis and applications*. *Advancement in Image Processing and Pattern Recognition*, 6(1), 44–53. <https://doi.org/10.5281/zenodo.7861337>
- Kavikuil, K., & Amudha, J. (2019). Leveraging deep learning for anomaly detection in video surveillance. In R. Bapi, K. Rao, & M. Prasad (Eds.), *First International Conference on Artificial Intelligence and Cognitive Computing* (Advances in Intelligent Systems and Computing, Vol. 815, pp. 265–274). Springer. https://doi.org/10.1007/978-981-13-1580-0_23
- Khaire, P., & Kumar, P. (2022). A semi-supervised deep learning based video anomaly detection framework using RGB-D for surveillance of real-world critical environments. *Forensic Science International: Digital Investigation*, 40, 301346. <https://doi.org/10.1016/j.fsidi.2022.301346>
- Khanam, R., & Hussain, M. (2024). YOLOv11: An overview of the key architectural enhancements [Preprint]. *arXiv*. <https://arxiv.org/abs/2410.17725>
- Kingma, D. P., & Ba, J. (2015, May). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA. *arXiv*. <https://doi.org/10.48550/arXiv.1412.6980>

- Ko, K.-E., & Sim, K.-B. (2018). Deep convolutional framework for abnormal behavior detection in a smart surveillance system. *Engineering Applications of Artificial Intelligence*, 67, 226–234. <https://doi.org/10.1016/j.engappai.2017.10.001>
- Kochenderfer, M. J., & Wheeler, T. A. (2025). Algorithms for optimization (2nd ed.). MIT Press.
- Köck, M., & Paramythis, A. (2011). Activity sequence modelling and dynamic clustering for personalized e-learning. *User Modeling and User-Adapted Interaction*, 21(1–2), 51–97. <https://doi.org/10.1007/s11257-010-9087-z>
- Krause, B., Lu, L., Murray, I., & Renals, S. (2016). Multiplicative LSTM for sequence modelling. *arXiv*. <https://doi.org/10.48550/arXiv.1609.07959>
- Kumaran, S. K., Dogra, D. P., Roy, P. P., & Mitra, A. (2018). Video trajectory classification and anomaly detection using hybrid CNN-VAE. *arXiv preprint arXiv:1812.07203*. <https://doi.org/10.48550/arXiv.1812.07203>
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. <https://doi.org/10.1109/5.726791>
- Lee, W.-K., Leong, C.-F., Lai, W.-K., Leow, L.-K., & Yap, T.-H. (2018). ArchCam: Real time expert system for suspicious behaviour detection in ATM site. *Expert Systems with Applications*, 109, 12–24. <https://doi.org/10.1016/j.eswa.2018.05.014>
- Lee, Y., Cho, S., & Choi, J. (2019). App usage prediction for dual display device via two-phase sequence modeling. *Pervasive and Mobile Computing*, 58, 101025. <https://doi.org/10.1016/j.pmcj.2019.05.006>
- Lee, Y.-J., Park, N.-E., & Lee, I.-G. (2023). Lightweight anomaly detection mechanism based on machine learning using low-cost surveillance cameras.

- En J.-S. No & D. Hong (Eds.), *Information security and cryptology – ICISC 2023: 26th International Conference on Information Security and Cryptology, Seoul, South Korea, November 29 – December 1, 2023, Revised selected papers, Part II* (pp. 81–92). Springer.
https://doi.org/10.1007/978-981-97-1238-0_5
- Li, C., Han, Z., Ye, Q., & Jiao, J. (2013). Visual abnormal behavior detection based on trajectory sparse reconstruction analysis. *Neurocomputing*, 119, 94–100.
<https://doi.org/10.1016/j.neucom.2012.03.040>
- Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12), 6999–7019.
<https://doi.org/10.1109/TNNLS.2021.3084827>
- Ma, L., Zhang, F., Sun, J., Xue, M., Li, B., Juefei-Xu, F., Xie, C., Li, L., Liu, Y., Zhao, J., & Wang, Y. (2018). DeepMutation: Mutation testing of deep learning systems. In *2018 IEEE 29th International Symposium on Software Reliability Engineering (ISSRE)* (pp. 100–111). IEEE.
<https://doi.org/10.1109/ISSRE.2018.00021>
- Mahmood, S. A., Abid, A. M., & Lafta, S. H. (2021). Anomaly event detection and localization of video clips using global and local outliers. *Indonesian Journal of Electrical Engineering and Computer Science*, 24(2), 1063–1073.
<https://doi.org/10.11591/ijeecs.v24.i2.pp1063-1073>
- Maleki, F., Muthukrishnan, N., Ovens, K., Reinhold, C., & Forghani, R. (2020). Machine learning algorithm validation: From essentials to advanced applications and implications for regulatory certification and deployment. *Neuroimaging Clinics of North America*, 30(4), 433–445.
<https://doi.org/10.1016/j.nic.2020.08.004>
- Martinez Torres, D., Loaiza Correa, H., & Caicedo Bravo, E. (2019). Online learning of contexts for detecting suspicious behaviors in surveillance videos. *Image*

and Vision Computing, 89, 197–210.
<https://doi.org/10.1016/j.imavis.2019.07.006>

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*.
<https://arxiv.org/abs/2005.02142>

Mehmood, A. (2021). LightAnomalyNet: A lightweight framework for efficient abnormal behavior detection. *Sensors*, 21(24), 8501.
<https://doi.org/10.3390/s21248501>

Ministerio de Seguridad de la Nación. (s. f.). Estadísticas criminales de la República Argentina. *Argentina.gob.ar*. Consultado el 23 de mayo de 2024.
<https://www.argentina.gob.ar/seguridad/estadisticascriminales>

Ministerio del Interior del Perú, Viceministerio de Seguridad Pública, Dirección General de Información para la Seguridad, Dirección de Canales de Atención y Denuncias. (2024, enero). *Reporte de las comunicaciones recibidas por la Central Única de Denuncias del Ministerio del Interior: Enero 2024*. Ministerio del Interior del Perú. Consultado el 1 de abril de 2024.
<https://www.policia.gob.pe/siepol/documentos/reporte-2024/enero/reporte-enero-2024.pdf>

Mohan, P. S. (2021). Violent crime and firm performance: Evidence from the Caribbean. *International Journal of the Economics of Business*, 28(2), 309–327. <https://doi.org/10.1080/13571516.2021.1896926>

Murugesan, M., & Thilagamani, S. (2022). Bayesian feed forward neural network-based efficient anomaly detection from surveillance videos. *Intelligent Automation & Soft Computing*, 34(1), 389–405.
<https://doi.org/10.32604/iasc.2022.024641>

Nassauer, A. (2018). How robberies succeed or fail: Analyzing crime caught on

- CCTV. *Journal of Research in Crime and Delinquency*, 55(1), 125–154.
<https://doi.org/10.1177/0022427817715754>
- Nawaratne, R., Alahakoon, D., De Silva, D., & Yu, X. (2020). Spatiotemporal anomaly detection using deep learning for real-time video surveillance. *IEEE Transactions on Industrial Informatics*, 16(1), 393–402.
<https://doi.org/10.1109/TII.2019.2938527>
- Oriazowanlan, A. O., & Erah, D. O. (2019). Effect of insecurity on micro, small and medium enterprises (MSMEs) development in Benin-City, Edo State. *International Journal of Business Management*, 4(3), 75-86.
<https://sciarena.com/storage/models/article/5BRQ0S82IVl8cXamiJQ4WNuygXhrtLQK6Anh9nzdOec0typZtStVL0uASING/effect-of-insecurity-on-micro-small-and-medium-enterprises-msmes-development-in-benin-city-edo-sta.pdf>
- Otiniano Otiniano, C. J. (2020). *Gestión administrativa y seguridad ciudadana en la Municipalidad Distrital de Víctor Larco Herrera, Trujillo, año 2019* [Tesis de maestría, Universidad César Vallejo]. Repositorio Institucional UCV.
<https://hdl.handle.net/20.500.12692/47570>
- Patel, A. S., Merlino, G., Bruneo, D., Puliafito, A., Vyas, O. P., & Ojha, M. (2020). Video representation and suspicious event detection using semantic technologies. *Semantic Web*, 12(3), 467–491.
<https://doi.org/10.3233/SW-200393>
- Paul, M., Haque, S. M. E., & Chakraborty, S. (2013). Human detection in surveillance videos and its applications: A review. *EURASIP Journal on Advances in Signal Processing*, 2013, 176. <https://doi.org/10.1186/1687-6180-2013-176>
- Pawar, K., & Attar, V. (2019). Deep learning approaches for video-based anomalous activity detection. *World Wide Web*, 22(2), 571–601.
<https://doi.org/10.1007/s11280-018-0582-1>
- Pawar, K., & Attar, V. (2020). Deep learning-based intelligent surveillance model for

- detection of anomalous activities from videos. *International Journal of Computational Vision and Robotics*, 10(4), 301–317. <https://doi.org/10.1504/IJCVR.2020.108152>
- Philip, F. M., Jayakrishnan, V., Ajesh, F., & Haseena, P. (2022). Video anomaly detection using the optimization-enabled deep convolutional neural network. *The Computer Journal*, 65(5), 1272–1292. <https://doi.org/10.1093/comjnl/bxaa177>
- Pinazo-Dallenbach, P. (2021). Los efectos de la inseguridad ciudadana sobre el emprendimiento: Un freno al desarrollo de Latinoamérica. *Estudios de Economía Aplicada*, 39(3). <https://doi.org/10.25115/eea.v39i2.3876>
- Policía Nacional del Perú, Dirección de Tecnología de la Información y Comunicaciones, División de Estadística. (2023). *Anuario PNP 2023*. Policía Nacional del Perú. Consultado el 1 de abril de 2024. <https://www.policia.gob.pe/siepol/documentos/anuario-2023/2.%20ANUARIO%20PNP%202023.pdf>
- Policía Nacional del Perú, Dirección de Tecnología de la Información y Comunicaciones, División de Estadística. (2024). *Boletín estadístico policial: I trimestre 2024*. Policía Nacional del Perú. Consultado el 1 de abril de 2024. <https://www.policia.gob.pe/siepol/documentos/boletin-2024/Boletin%20I%20Trimestre%202024.pdf>
- Pouyan, S., Charmi, M., Azarpeyvand, A., & Hassanpoor, H. (2023). Propounding first artificial intelligence approach for predicting robbery behavior potential in an indoor security camera. *IEEE Access*, 11, 60471–60489. <https://doi.org/10.1109/ACCESS.2023.3284472>
- Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, 2(1), 37–63. <https://doi.org/10.48550/arXiv.2010.16061>

- Quadri, S. A., & Katakdhond, K. S. (2022). Suspicious activity detection using convolution neural network. *Journal of Pharmaceutical Negative Results*, 13(S01), 1235–1245. <https://doi.org/10.47750/pnr.2022.13.S01.151>
- Rai, H., Kolekar, M. H., Keshav, N., & Mukherjee, J. K. (2015). Trajectory based unusual human movement identification for video surveillance system. In H. Selvaraj, D. Zydek, & G. Chmaj (Eds.), *Progress in systems engineering* (Advances in Intelligent Systems and Computing, Vol. 366, pp. 759–764). Springer. https://doi.org/10.1007/978-3-319-08422-0_114
- Raj, P., Soundarabai, P. B., & Augustine, P. (Eds.). (2023). *Machine intelligence: Computer vision and natural language processing* (1st ed.). Auerbach Publications. <https://doi.org/10.1201/9781003424550>
- Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., & Sebe, N. (2017). Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)* (pp. 1577–1581). IEEE. <https://doi.org/10.1109/ICIP.2017.8296547>
- Rodrigues, R., Bhargava, N., Velmurugan, R., & Chaudhuri, S. (2020). Multi-timescale trajectory prediction for abnormal human activity detection. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 2615–2623). IEEE. <https://doi.org/10.1109/WACV45572.2020.9093633>
- Ronceros Leyva, E. G. (2020). *Estrategias de prevención contra la inseguridad ciudadana y su impacto en el incremento de los delitos en el distrito de Surquillo en los años 2015-2019* [Tesis de maestría, Universidad Tecnológica del Perú]. Repositorio Institucional UTP. <https://hdl.handle.net/20.500.12867/4842>
- Roy, D., & Krishna Mohan, C. (2018). Snatch theft detection in unconstrained surveillance videos using action attribute modelling. *Pattern Recognition Letters*, 108, 56–61. <https://doi.org/10.1016/j.patrec.2018.03.004>

- Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Pearson.
- Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Proceedings of Interspeech 2014*, 338–342. <https://doi.org/10.21437/Interspeech.2014-80>
- Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology*, 52(1), 471–499. <https://doi.org/10.1146/annurev.psych.52.1.471>
- Salem, F. M. (2022). *Recurrent neural networks: From simple to gated architectures*. Springer. <https://doi.org/10.1007/978-3-030-89929-5>
- Sánchez Agreda, Z. (2024). Inteligencia artificial y sus subcampos de estudio: Una revisión histórica. *Revista Peruana de Computación y Sistemas*, 6(2), 95–110. <https://doi.org/10.15381/rpcs.v6i2.28857>
- Sarkis, A. (2023). *Training data for machine learning: Human supervision from annotation to data science*. O'Reilly Media.
- Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004)* (Vol. 3, pp. 32–36). IEEE. <https://doi.org/10.1109/ICPR.2004.1334462>
- Sharma, M. K., Sheet, D., & Biswas, P. K. (2020). Spatiotemporal deep networks for detecting abnormality in videos. *Multimedia Tools and Applications*, 79(15–16), 11237–11268. <https://doi.org/10.1007/s11042-020-08786-w>
- Shen, M., Jiang, X., & Sun, T. (2018). Anomaly detection based on nearest neighbor search with locality-sensitive B-tree. *Neurocomputing*, 289, 55–67. <https://doi.org/10.1016/j.neucom.2018.02.012>

- Shen, M., Liu, D.-R., & Shann, S.-H. (2015). Outlier detection from vehicle trajectories to discover roaming events. *Information Sciences*, 294, 242–254. <https://doi.org/10.1016/j.ins.2014.09.037>
- Sheu, R.-K., Pardeshi, M., Chen, L.-C., & Yuan, S.-M. (2019). STAM-CCF: Suspicious tracking across multiple camera based on correlation filters. *Sensors*, 19(13), 3016. <https://doi.org/10.3390/s19133016>
- Singleton, C. R., Winata, F., Adams, A. M., Appelhans, B. M., Odoms-Young, A., & Powell, L. M. (2022). County-level associations between food retailer availability and violent crime rate. *BMC Public Health*, 22, 2002. <https://doi.org/10.1186/s12889-022-14415-y>
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- Szeliski, R. (2022). *Computer vision: Algorithms and applications* (2nd ed.). Springer. <https://doi.org/10.1007/978-3-030-34372-9>
- Tax, N., Teinemaa, I., & van Zelst, S. J. (2020). An interdisciplinary comparison of sequence modeling methods for next-element prediction. *Software and Systems Modeling*, 19(6), 1345–1365. <https://doi.org/10.1007/s10270-020-00789-3>
- Thurtell, S. C. (2021). When shopping turns scary: Twitter conversations about violent crime in Johannesburg malls. *Critical Arts*, 35(1), 85–96. <https://doi.org/10.1080/02560046.2021.1891445>
- Townsley, M., & Hutchins, B. (2024). Framing retail crime through an environmental criminological lens: Insights from Australia and New Zealand. *Crime Prevention and Community Safety*, 26(2), 157–179. <https://doi.org/10.1057/s41300-024-00208-5>

- Tripathi, R. K., Jalal, A. S., & Agrawal, S. C. (2018). Suspicious human activity recognition: A review. *Artificial Intelligence Review*, 50(2), 283–339. <https://doi.org/10.1007/s10462-017-9545-7>
- Ullah, W., Ullah, A., Hussain, T., Khan, Z. A., & Baik, S. W. (2021). An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos. *Sensors*, 21(8), 2811. <https://doi.org/10.3390/s21082811>
- Umbaugh, S. E. (2023). *Digital image processing and analysis: Computer vision and image analysis* (4th ed.). CRC Press. <https://doi.org/10.1201/9781003221135>
- Van Rijsbergen, C. J. (1979). *Information retrieval* (2.^a ed.). Butterworths.
- Vina, A. (2024, December 5). *The benefits of Ultralytics YOLO11 being an anchor-free detector*. Ultralytics. Consultado el 7 de enero de 2024. <https://www.ultralytics.com/blog/benefits-ultralytics-yolo11-being-anchor-free-detector>
- Wang, B., Yang, C., & Chen, Y. (2022). Detection anomaly in video based on deep support vector data description. *Computational Intelligence and Neuroscience*, 2022, Article 5362093. <https://doi.org/10.1155/2022/5362093>
- Wang, L., Tan, H., Zhou, F., Zuo, W., & Sun, P. (2022). Unsupervised anomaly video detection via a double-flow ConvLSTM variational autoencoder. *IEEE Access*, 10, 44278–44289. <https://doi.org/10.1109/ACCESS.2022.3165977>
- Wang, T., Qiao, M., Deng, Y., Zhou, Y., Wang, H., Lyu, Q., & Snoussi, H. (2018). Abnormal event detection based on analysis of movement information of video sequence. *Optik*, 152, 50–60. <https://doi.org/10.1016/j.ijleo.2017.07.064>
- Wang, T., Qiao, M., Snoussi, H., Li, Y., Liu, B., & Gao, Y. (2019). Generative neural networks for anomaly detection in crowded scenes. *IEEE Transactions on Information Forensics and Security*, 14(5), 1390–1399. <https://doi.org/10.1109/TIFS.2018.2878538>

- Wang, Y., Wu, J., & Li, H. (2020). Human detection based on improved Mask R-CNN. *Journal of Physics: Conference Series*, 1575(1), 012067. IOP Publishing. <https://doi.org/10.1088/1742-6596/1575/1/012067>
- Waqas, M., & Wannasingha Humphries, U. (2024). A critical review of RNN and LSTM variants in hydrological time series predictions. *MethodsX*, 13, 102946. <https://doi.org/10.1016/j.mex.2024.102946>
- Wiliem, A., Madasu, V., Boles, W., & Yarlagadda, P. (2008). Detecting uncommon trajectories. In *Proceedings of the 2008 Digital Image Computing: Techniques and Applications (DICTA)* (pp. 398–404). IEEE. <https://doi.org/10.1109/DICTA.2008.45>
- Yadav, K., Pandit, V., Pabbaraju, C., & Bahl, L. (2020). *Probabilistic graphical models, Spring 2020: 14. Deep sequence models* [Lecture notes]. Carnegie Mellon University. Consultado el 28 de febrero de 2025. https://www.cs.cmu.edu/~epxing/Class/10708-20/scribe/lec14_scribe.pdf
- Yan, R., Peng, L., Xiao, S., Wang, J., & Wang, Y. (2019). Dynamic temporal residual network for sequence modeling. *International Journal on Document Analysis and Recognition (IJ DAR)*, 22(3), 235–246. <https://doi.org/10.1007/s10032-019-00328-x>
- Yang, Z., Liu, J., & Wu, P. (2021). Bidirectional retrospective generation adversarial network for anomaly detection in videos. *IEEE Access*, 9, 107842–107857. <https://doi.org/10.1109/ACCESS.2021.3100678>
- Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A survey. *ACM Computing Surveys*, 38(4), 13–es. <https://doi.org/10.1145/1177352.1177355>
- Zargar, S. A. (2021). *Introduction to sequence learning models: RNN, LSTM, GRU*. ResearchGate. <https://doi.org/10.13140/RG.2.2.36370.99522>

- Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). *Dive into deep learning*. arXiv. <https://doi.org/10.48550/arXiv.2106.11342>
- Zhang, X., Sugano, Y., & Bulling, A. (2018). Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (ETRA '18, Article 12, pp. 1–9). Association for Computing Machinery. <https://doi.org/10.1145/3204493.3204548>
- Zhang, X., Wang, R., & Ding, J. (2018). Abnormal event detection by learning spatiotemporal features in videos. In Y. Wang, Z. Jiang, & Y. Peng (Eds.), *Image and graphics technologies and applications (IGTA 2018)* (Communications in computer and information science, Vol. 875, pp. 421–431). Springer. https://doi.org/10.1007/978-981-13-1702-6_42
- Zhao, Z.-Q., Zheng, P., Xu, S.-T., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11), 3212–3232. <https://doi.org/10.1109/TNNLS.2018.2876865>
- Zhou, K., Hui, B., Wang, J., Wang, C., & Wu, T. (2021). A study on attention-based LSTM for abnormal behavior recognition with variable pooling. *Image and Vision Computing*, 108, 104120. <https://doi.org/10.1016/j.imavis.2021.104120>

ANEXOS

ANEXO A: Clasificación de papers según fuentes

Se presenta a continuación el archivo digital con la clasificación detallada de las publicaciones científicas revisadas en el desarrollo de la presente investigación, organizada según tipo de fuente, relevancia, año de publicación y aportes principales.

Archivo: Clasificación de Papers Según Fuentes.xlsx

Formato: Hoja de cálculo (Google Sheets)

Enlace de acceso:

<https://docs.google.com/spreadsheets/d/1XT3oemSfDAT-okMWCn37Lq86itPTrqvZ/edit?usp=sharing&ouid=108733148890533056270&rtpof=true&sd=true>

ANEXO B: Pseudocódigo del seguimiento de objetos con YOLO

A continuación se incluye el pseudocódigo del procedimiento de tracking básico recomendado por Ultralytics, que utiliza la opción `persist=True` para mantener la consistencia de los identificadores entre frames.

Algorithm 1 Seguimiento básico de objetos utilizando YOLO con persistencia de tracks (basado en la documentación oficial de Ultralytics)

Require: Video de entrada, modelo YOLO preentrenado

Ensure: Video anotado con seguimiento de objetos

```
1: Importar bibliotecas: cv2, ultralytics.YOLO
2: model ← YOLO("yolo11n.pt")                                ▷ o yolo26n.pt, etc.
3: video_path ← "ruta/del/video.mp4"
4: cap ← VIDEOCAPTURE(video_path)
5: while CAP.ISOPENED do
6:     success, frame ← cap.read()
7:     if success then
8:         results ← model.track(frame, persist = true) ▷ Mantiene IDs entre
           frames
9:         annotated_frame ← results[0].plot() ▷ Dibuja cajas, IDs y máscaras
           si corresponde
10:        Mostrar ventana: "YOLO Tracking", annotated_frame
11:        if CV2.WAITKEY(1) ∧ tecla presionada = 'q' then
12:            break
13:        end if
14:    else
15:        break                                                ▷ Fin del video
16:    end if
17: end while
18: cap.release()
19: CV2.DESTROYALLWINDOWS
```

ANEXO C: Parámetros de entrenamiento del modelo LSTM

El modelo recurrente tipo LSTM desarrollado para la predicción de posiciones futuras (coordenadas x, y y tiempo) en las trayectorias analizadas se construyó con la siguiente arquitectura base y parámetros de entrenamiento generales:

Arquitectura del modelo:

- Capa LSTM inicial: 64 unidades, función de activación ReLU, con retorno de secuencias activado
- Capa LSTM secundaria: 32 unidades, función de activación ReLU
- Capa densa de salida: una unidad por cada variable a predecir (3 en total: x, y, time)
- Optimizador: Adam
- Función de pérdida: Error Cuadrático Medio (MSE)

,Parámetros de entrenamiento principales:

- Longitud de la secuencia de entrada (sequence_length): * (detallado en Capítulo 4, sección de preparación de datos)
- Número de épocas (epochs): * (detallado en el Capítulo 3)
- Tamaño del lote (batch_size): 32

ANEXO D: Ejemplo de ejecución del módulo de detección de trayectorias sospechosas

A continuación se presenta una captura representativa de la consola durante el procesamiento y clasificación del conjunto de prueba compuesto por 12 trayectorias diseñadas como actividades sospechosas.

Se observa:

- El procesamiento secuencial de cada trayectoria (longitud total y porcentaje de puntos clasificados como sospechosos).
- Todas las trayectorias fueron etiquetadas como SOSPECHOSAS.
- Resumen final: 100% de detección (12/12), superando ampliamente el criterio H1 (tasa de detección $\geq 80\%$).

```
Analizando 12 trayectorias anómalas o sospechosas...
Usando 10% como Umbral para trayectorias SOSPECHOSAS
Procesando Trayectoria 1/12 (longitud: 1009)
  Trayectoria 1: SOSPECHOSA (170/999 puntos = 17.0%)
Procesando Trayectoria 2/12 (longitud: 44)
  Trayectoria 2: SOSPECHOSA (8/34 puntos = 23.5%)
Procesando Trayectoria 3/12 (longitud: 2009)
  Trayectoria 3: SOSPECHOSA (946/1999 puntos = 47.3%)
Procesando Trayectoria 4/12 (longitud: 1225)
  Trayectoria 4: SOSPECHOSA (271/1215 puntos = 22.3%)
Procesando Trayectoria 5/12 (longitud: 1229)
  Trayectoria 5: SOSPECHOSA (336/1219 puntos = 27.6%)
Procesando Trayectoria 6/12 (longitud: 1160)
  Trayectoria 6: SOSPECHOSA (241/1150 puntos = 21.0%)
Procesando Trayectoria 7/12 (longitud: 847)
  Trayectoria 7: SOSPECHOSA (165/837 puntos = 19.7%)
Procesando Trayectoria 8/12 (longitud: 1975)
  Trayectoria 8: SOSPECHOSA (1124/1965 puntos = 57.2%)
Procesando Trayectoria 9/12 (longitud: 2212)
  Trayectoria 9: SOSPECHOSA (1418/2202 puntos = 64.4%)
Procesando Trayectoria 10/12 (longitud: 772)
  Trayectoria 10: SOSPECHOSA (762/762 puntos = 100.0%)
Procesando Trayectoria 11/12 (longitud: 1027)
  Trayectoria 11: SOSPECHOSA (861/1017 puntos = 84.7%)
Procesando Trayectoria 12/12 (longitud: 1204)
  Trayectoria 12: SOSPECHOSA (141/1194 puntos = 11.8%)

=== Resultados de la Detección ===
Trayectorias analizadas en total: 12
Detectadas como anómalas (>10% puntos): 12
Porcentaje de detección: 100.00%
H1 criterio (detection rate  $\geq 80\%$ ):  PASS
```