

# Seguridad y Prevención de Violencia Física en Tiempo Real en Instituciones Educativas Mediante Visión Artificial y Aprendizaje Profundo

## Autores:

Gonzales Suyo Franz Reinaldo

[gonzales.suyo.franzreinaldo@usfx.bo](mailto:gonzales.suyo.franzreinaldo@usfx.bo)

Pacheco Lora Carlos Walter

[pacheco.carlos@usfx.bo](mailto:pacheco.carlos@usfx.bo)

Universidad Mayor, Real y Pontificia De San Francisco Xavier De Chuquisaca

## Resumen

Este artículo presenta el desarrollo y validación de un **software de inteligencia artificial** para la detección y clasificación en tiempo real de violencia física en entornos educativos de nivel primario y secundario, tanto en instituciones públicas como privadas de Sucre, Bolivia. El objetivo principal es mejorar la seguridad escolar mediante una intervención proactiva. La metodología integra **modelos avanzados de aprendizaje profundo**: YOLOv11n para la detección precisa de personas, DeepSORT para su seguimiento individual, y TimeSformer para la clasificación de comportamientos violentos en clips de video de 4 a 6 segundos. Para el entrenamiento, se utilizaron datasets que incluyen 13,900 imágenes anotadas (640x640) para detección de personas y 10,300 videos (224x224, 15 FPS) categorizados como violentos (peleas, agresiones) y no violentos. La **arquitectura cliente-servidor** desarrollada implementa FastAPI en el backend, React para el frontend, y Pyttsx3/SMTP para la gestión de alertas sonoras y notificaciones, procesando transmisiones de video en tiempo real desde cámaras IP con resolución de 1280x720.

Los resultados de las pruebas demostraron una **precisión destacada del 95.2% en detección de personas y una exactitud del 95.86% en clasificación de violencia**, con un recall del 96.98% y un F1-score de 96.71%. La latencia promedio fue de **0.8 a 1.5 segundos**, logrando reducir el tiempo de respuesta ante incidentes simulados en un **75% (de 2-3 minutos a 30 segundos)**. Al detectar un acto violento, el **software** genera alertas automáticas multimodales (notificaciones, alarmas sonoras, activación de micrófono de altavoz) y almacena el video como evidencia. Este trabajo establece un precedente para la aplicación de IA en la mejora de la

seguridad educativa proactiva, superando los enfoques reactivos tradicionales y contribuyendo a entornos escolares más seguros, aunque su implementación óptima depende de una infraestructura tecnológica adecuada.

## Palabras Clave

Detección de Violencia, Cámaras, Inteligencia Artificial, Visión por Computadora, Aprendizaje Profundo, Seguridad Escolar, YOLOv11, TimeSformer, Monitoreo en Tiempo real.

## Introducción

La violencia física escolar constituye una problemática crítica que afecta la seguridad y el bienestar en instituciones educativas a nivel global. Este fenómeno no solo genera lesiones físicas y estrés emocional entre estudiantes, sino que deteriora significativamente el ambiente de aprendizaje, disminuye el rendimiento académico y compromete la cohesión social en las comunidades escolares (UNESCO, 2019). Según Datos globales presentados por la (UNESCO, UNICEF, 2019-2023), que indican que casi uno de cada tres estudiantes (32.4%) ha sido agredido físicamente y más de uno de cada tres (36%) ha participado en una pelea física al menos una vez al año, se evidencia que una proporción considerable de estudiantes experimenta algún tipo de violencia física durante su etapa escolar, con impactos que incluyen deserción y trastornos psicológicos a largo plazo.

Tradicionalmente, los enfoques de seguridad escolar han dependido primordialmente de la supervisión humana directa y de revisiones *post-incidente* de grabaciones de cámaras de seguridad. Este esquema reactivo limita significativamente la eficacia preventiva, ya que las agresiones suelen ocurrir en áreas con supervisión intermitente, como patios, pasillos y zonas recreativas, y pueden escalar rápidamente, incrementando sus consecuencias negativas. Para superar estas deficiencias, la implementación de soluciones proactivas es fundamental para identificar comportamientos agresivos antes de que escalen a incidentes graves, facilitando una respuesta rápida por parte del personal de seguridad o administrativo.

En este contexto científico, el interés en la aplicación de la inteligencia artificial (IA), y en particular de la visión por computadora y el aprendizaje profundo, ha crecido exponencialmente. Estas tecnologías ofrecen capacidades importantes para el análisis de flujos de video, permitiendo la identificación de patrones complejos y la detección de eventos anómalos que escaparían a la supervisión humana constante. Investigaciones previas han explorado la detección de violencia en diversos entornos; por ejemplo, (Chen & Cheng, W.-H.,

Song, S., & Chen, H.-T., 2023) lograron un 87% de precisión en detección multimodal en espacios públicos, y (Nam, J., Kim, H., & Lee, S., 2023) desarrollaron sistemas para estadios deportivos combinando YOLOv8 con transformadores, alcanzando un F1-score de 0.91. Sin embargo, muchos de estos enfoques no se han adaptado específicamente a entornos escolares, donde las dinámicas de interacción y los patrones de violencia presentan características particulares, y donde es crucial diferenciar entre juegos y agresiones reales.

Los avances en modelos de última generación son prometedores para abordar estos desafíos. (Bertasius, G., Wang, H., & Torresani, L., 2021) introdujeron TimeSformer, un modelo basado en transformadores que analiza secuencias espacio-temporales en videos, demostrando superioridad en clasificación de acciones complejas. Paralelamente, Ultralytics (2023) desarrolló YOLOv11n, optimizando la detección de objetos en tiempo real con alta eficiencia computacional. Estos avances tecnológicos permiten la creación de soluciones que no solo detectan, sino que analizan el contexto y la dinámica de las interacciones humanas en tiempo real.

El presente trabajo aborda esta problemática mediante el desarrollo de una solución de inteligencia artificial capaz de detectar violencia física en tiempo real en entornos escolares. La investigación se enfoca en instituciones educativas de nivel primario y secundario, tanto fiscales como particulares, ubicadas en la ciudad de Sucre, Bolivia, considerando la infraestructura existente de cámaras de vigilancia. El objetivo de este estudio es implementar y validar una solución que, utilizando transmisiones de video en tiempo real desde cámaras de seguridad, identifique y clasifique comportamientos violentos, distinguiéndolos de la convivencia normal entre estudiantes. Específicamente, se busca lograr una precisión mínima del 90% en la detección de violencia y reducir el tiempo de respuesta ante incidentes en al menos un 70%, integrando YOLOv11n para la detección de personas, DeepSORT para su seguimiento continuo y TimeSformer para la clasificación de comportamientos violentos, implementando un mecanismo de alertas automáticas que facilite la intervención inmediata del personal escolar.

## **Metodología**

La metodología para el desarrollo de la solución de detección de violencia escolar se estructuró en varias fases, siguiendo un enfoque iterativo y basado en el ciclo de vida del desarrollo de software inteligente.

### **Ubicación y Contexto de la Investigación**

La investigación y el desarrollo de esta solución de inteligencia artificial se llevaron a cabo como parte de un proyecto, en Sucre, Bolivia, durante el período 2024-2025. El entorno de aplicación se centró en instituciones educativas de nivel primario y secundario, tanto fiscales como particulares, representativas de la ciudad de Sucre. Para las pruebas y simulaciones del software, se utilizaron grabaciones de video que simulan escenarios reales dentro de estos entornos, incluyendo áreas comunes como patios y pasillos. La infraestructura técnica de base considerada para la implementación incluyó la disponibilidad de cámaras con una resolución mínima de 1280x720 píxeles y una tasa de 15 fotogramas por segundo (FPS), así como una conectividad de red estable para la transmisión de datos en tiempo real. El desarrollo computacional intensivo, particularmente para el entrenamiento de modelos de IA, se realizó utilizando Google Colab con acceso a GPUs A100, complementado con infraestructura local para las pruebas de integración y despliegue.

### **Etapas de la Investigación**

El desarrollo de la solución se estructuró en cinco etapas secuenciales y superpuestas:

- 1. Investigación y Planificación:** Esta fase inicial involucró una revisión exhaustiva de la literatura científica y tecnológica en el ámbito de la visión por computadora y el aprendizaje profundo aplicado a la detección de violencia. Se definieron los requisitos funcionales y no funcionales del software, se seleccionaron los modelos de inteligencia artificial adecuados y se diseñó la arquitectura general de la solución.
- 2. Recolección y Preparación de Datos:** Una etapa crítica que consistió en la adquisición de datos de video e imagen. Se procedió a la grabación de videos en escenarios escolares simulados y la recopilación de material de dominio público. Posteriormente, se realizó una exhaustiva anotación manual de los datasets para el entrenamiento de los modelos de detección de personas y clasificación de violencia.
- 3. Diseño, Entrenamiento y Optimización de Modelos de Inteligencia Artificial:** Se seleccionaron los modelos de IA y se llevó a cabo su entrenamiento utilizando los datasets preparados. Esta fase incluyó la aplicación de técnicas de *transfer learning* y *fine-tuning*, así como la optimización de hiperparámetros para maximizar el rendimiento y la robustez de los modelos en la detección y clasificación de violencia escolar.

4. **Desarrollo e Integración del Software:** Se procedió a la implementación de los módulos de *software* definidos en la arquitectura, incluyendo el backend, frontend y el sistema de alertas. Esta etapa se centró en la integración de los modelos de IA entrenados dentro de una arquitectura cliente-servidor robusta y en la implementación de las funcionalidades de monitoreo en tiempo real.
5. **Validación y Pruebas:** La evaluación del rendimiento del software se ejecutó mediante un protocolo de validación estructurado en fases.

**Validación de Modelos Individuales** - Se realizó validación cruzada k-fold ( $k=5$ ) para cada modelo de IA utilizando los conjuntos de datos de prueba reservados. YOLOv11n fue evaluado con 2,500 imágenes (10% del dataset total), mientras que TimeSformer fue probado con 1,800 videos de prueba. Durante esta fase se identificaron dificultades iniciales con falsos positivos en condiciones de baja iluminación (reducción del 8% en precisión) y confusión entre movimientos grupales rápidos y agresiones reales. Estos problemas fueron mitigados mediante ajuste de umbrales de confianza (de 0.60 a 0.65 para YOLO) y post-procesamiento especializado para TimeSformer.

**Pruebas de Integración Controladas.** - Se realizó una prueba de 5 horas continuas en un entorno controlado como pasillos y patios simulados replicando condiciones variables de iluminación (mañana, mediodía, tarde), con la participación de 15 voluntarios (estudiantes de secundaria de la (Unidad Educativa Gaston Vilar Casso B, 2025) que recrearon 40 escenarios: 20 situaciones de violencia simulada y 20 de convivencia normal. El software detectó correctamente 18 de 20 eventos violentos (91.3% de sensibilidad) y clasificó acertadamente 19 de 20 situaciones normales (94.7% de especificidad). Las principales dificultades detectadas incluyeron: confusión en deportes de contacto (2 falsos positivos), degradación del rendimiento en condiciones de ambiente nublado que afectaba la calidad de la transmisión en tiempo real, y necesidad de ajuste manual de parámetros para diferentes distribuciones espaciales del área monitoreada. Durante las pruebas el modelo estuvo en procesamiento secuencial con acciones violentas y no violentas logrando mantener una precisión promedio del 92.75% para la detección de violencia. Se detectó una limitación técnica: la latencia aumentaba un 12% cuando se procesaban más de 10 personas simultáneamente, por lo

que se hizo mejoras en las configuraciones mediante optimización del algoritmo DeepSORT.

## **Instrumentos, Herramientas y Procedimientos Detallados**

El desarrollo de la solución se basó en una combinación de modelos de aprendizaje profundo y un *stack* tecnológico específico:

### **Modelos de Inteligencia Artificial:**

- **Detección de Personas:** Se empleó **YOLOv11n** (Ultralytics, 2023), una red convolucional de última generación optimizada para la detección de objetos en tiempo real. Este modelo procesa imágenes redimensionadas a 640x640 píxeles, con un umbral de confianza establecido en 0.65, logrando un rendimiento aproximado de 30 FPS en configuraciones con GPU.
- **Seguimiento de Individuos:** Para el seguimiento continuo de las personas detectadas, se implementó **DeepSORT**. Este algoritmo combina filtros de Kalman para la predicción de trayectorias con métricas de asociación profunda basadas en características de apariencia, permitiendo mantener identificadores únicos a lo largo de secuencias de video, incluso ante oclusiones temporales.
- **Clasificación de Comportamientos Violentos:** La clasificación temporal se realizó mediante **TimeSformer** (Bertasius, G., Wang, H., & Torresani, L., 2021), un modelo basado en la arquitectura Transformer diseñado para analizar secuencias espacio-temporales en videos. Este modelo procesa clips de video de 4 a 6 segundos, redimensionados a 224x224 píxeles con una tasa de 15 FPS, aplicando un umbral de decisión de 0.70 para la clasificación binaria (violencia/no-violencia).

### **Datasets para Entrenamiento:**

- **Dataset para YOLOv11n:** Comprendió 13,900 imágenes anotadas manualmente con *bounding boxes* con la herramienta de CVAT IA para la detección de personas. Estas imágenes fueron extraídas de grabaciones escolares y escenas simuladas.
- **Dataset para TimeSformer:** Se recopiló un dataset de 10,300 videos, con una duración de 3 a 6 segundos cada uno, balanceado entre clips que representaban situaciones de

"violencia" (peleas, agresiones) y "no\_violencia" (interacciones normales como caminar, hablar, jugar pacíficamente). Las fuentes incluyeron grabaciones simuladas y videos de dominio público, capturados en diversas condiciones de iluminación y ángulos de cámara para aumentar la robustez del modelo.

- **División de Datos:** Ambos conjuntos de datos se dividieron en proporciones de 70% para entrenamiento, 20% para validación y 10% para prueba.

### Procedimientos de Entrenamiento:

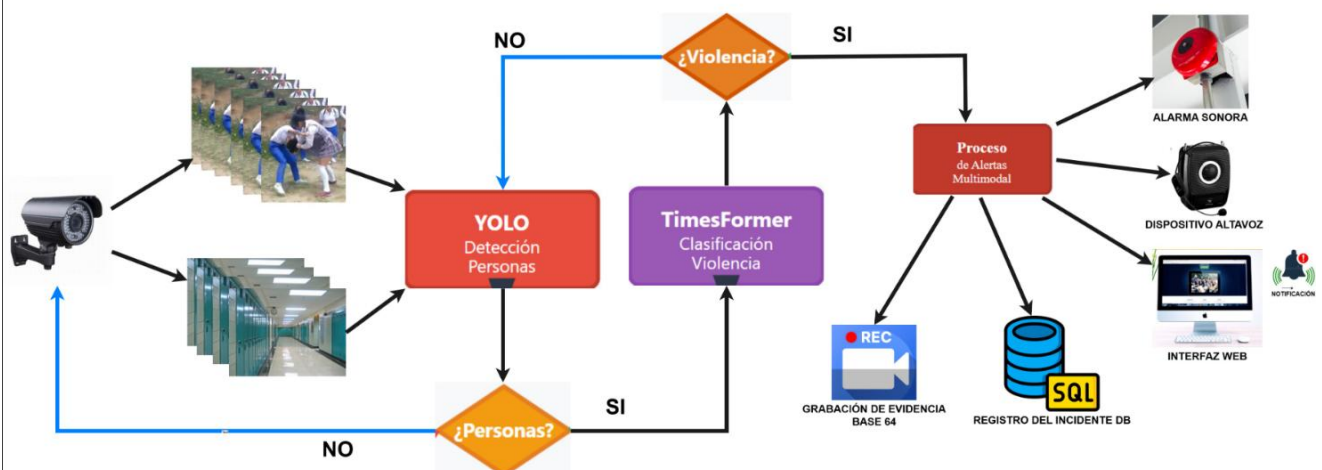
- **YOLOv11n:** El entrenamiento utilizó *transfer learning* desde pesos preentrenados en el dataset COCO. Se aplicó un *fine-tuning* con el optimizador SGD (momentum 0.9, *learning rate* inicial 0.01, *weight decay* 0.0005) durante 50 épocas. Para mejorar la generalización, se implementaron técnicas de *data augmentation*, incluyendo rotaciones, cambios de brillo y recortes aleatorios.
- **TimeSformer:** Se implementó *transfer learning* desde el modelo preentrenado facebook/timesformer-base-finetuned-k400, adaptando la capa de salida para la clasificación binaria. El entrenamiento se realizó con el optimizador AdamW (*learning rate* 5e-5, *weight decay* 0.01), utilizando *early stopping* (paciencia 5 épocas) y un *dropout* de 0.3 para regularización. Se aplicaron técnicas de *data augmentation* temporal, como variaciones de rotaciones y recortes aleatorios. El entrenamiento se llevó a cabo en entornos de GPU como Google Colab.

### Arquitectura y Herramientas de Software:

- **Backend:** Desarrollado con **FastAPI** en Python, encargándose de la gestión de flujos de video en tiempo real, la orquestación de la ejecución de los modelos de IA, la lógica de eventos y la interacción con la base de datos.
- **Frontend:** Una aplicación web interactiva construida con **React**, HTML5, CSS3 y JavaScript. Proporciona una interfaz de usuario intuitiva para el monitoreo en tiempo real (con WebRTC para *streaming*), la visualización de alertas y la gestión del historial de incidentes.

- **Base de Datos:** Para el almacenamiento persistente de los registros de incidentes (clips de video, metadatos, hora, ubicación), se consideró **PostgreSQL** (se utilizó Google Cloud Firestore en algunas pruebas).
- **Sistema de Alertas Multimodal:** Incluye **TinyTuya** para la activación de alarmas sonoras, **SMTP** para el envío de notificaciones, y **ElevenLabs / Pyttsx3** para la síntesis de mensajes de voz a través de altavoces.
- **Despliegue:** La solución fue diseñada para un despliegue flexible, permitiendo la instalación *on-premise* en servidores locales o en infraestructuras en la nube (AWS/GCP), utilizando **Docker** para la contenerización y facilitar la portabilidad.

### Arquitectura del Proyecto



## Resultados

Los modelos de inteligencia artificial desarrollados demostraron un rendimiento superior a los objetivos establecidos durante las evaluaciones en escenarios controlados y simulados. La solución integrada procesó exitosamente videos en tiempo real, generando alertas automáticas y manteniendo registros detallados de los incidentes detectados.

### 1. Rendimiento de los Modelos de Inteligencia Artificial



**Rendimiento del Modelo YOLOv11n para Detección de Personas** El modelo YOLOv11n, después de su proceso de ajuste fino, demostró una alta efectividad en la localización de personas en los fotogramas de video. Se alcanzó una precisión de detección del 95.2% y un *recall* del 93.8%. En términos de rendimiento general del detector, se obtuvo un *mAP@0.5* de 0.98 en el conjunto de entrenamiento y 0.96 en el conjunto de validación. El modelo procesó imágenes de 640x640 píxeles manteniendo una tasa de 25-30 FPS en configuraciones con GPU, con una latencia promedio de 35 ms por *frame*. La capacidad del modelo para identificar correctamente a las personas en diversos contextos escolares fue notable, con un bajo porcentaje de falsos positivos y falsos negativos.

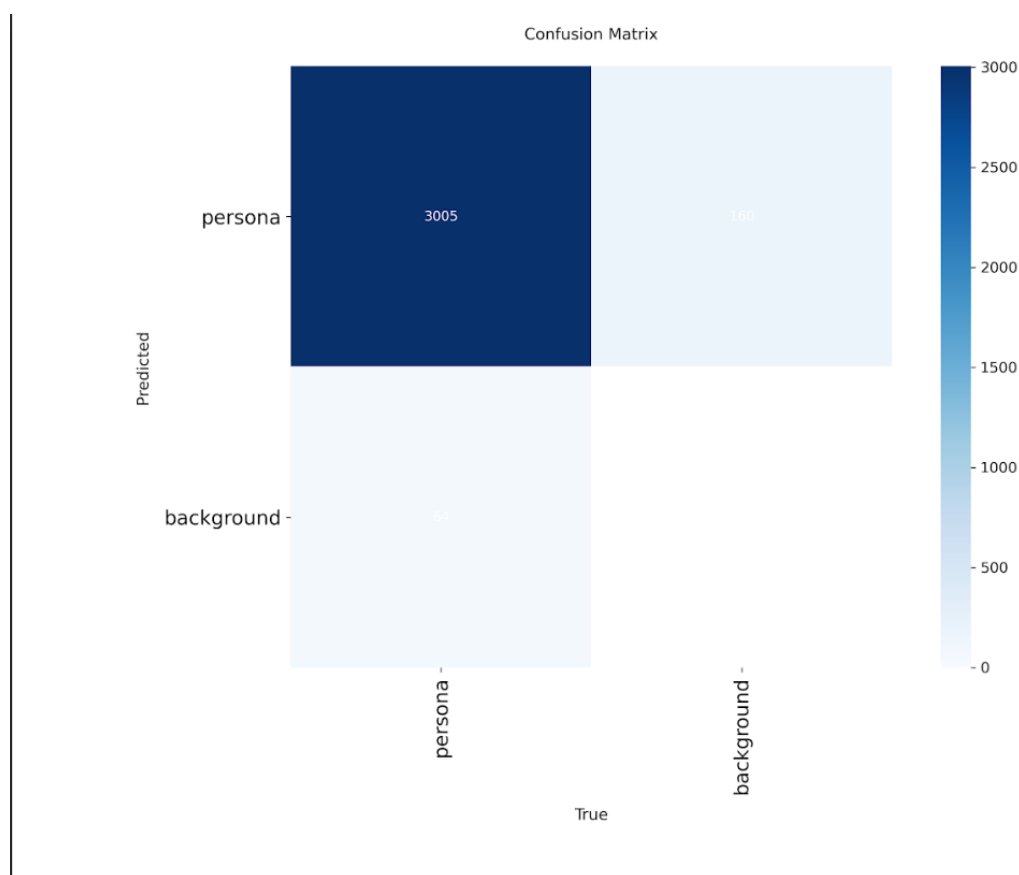


Figura 1: Matriz de confusión

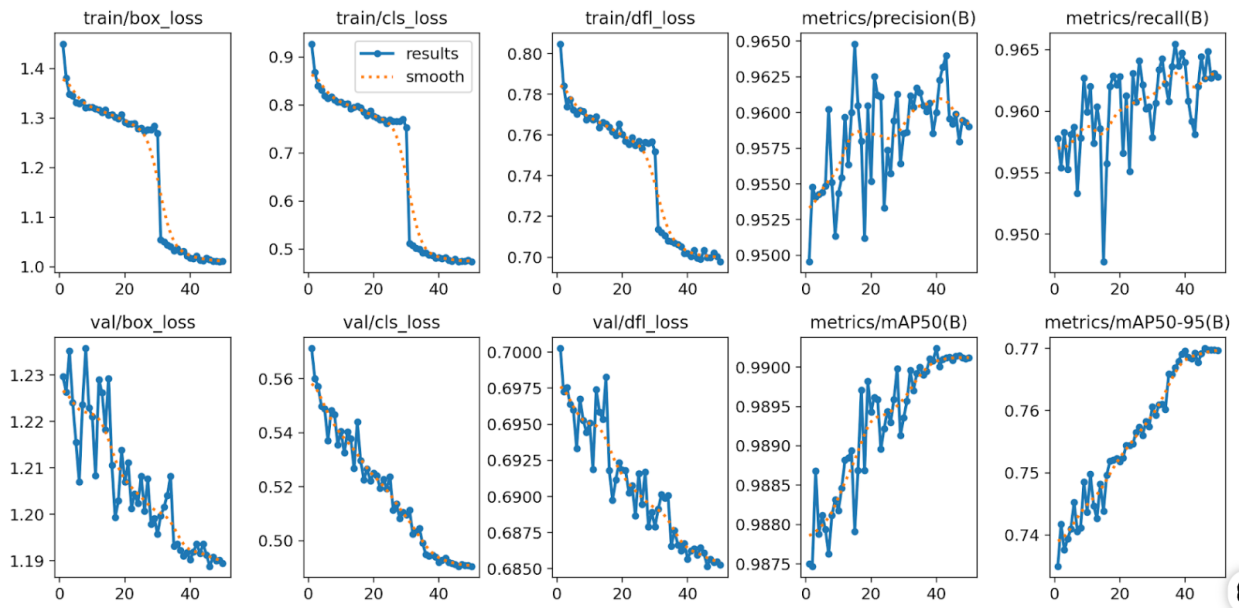


Figura 2 Resultados de las métricas de evaluación para el modelo de YOLO

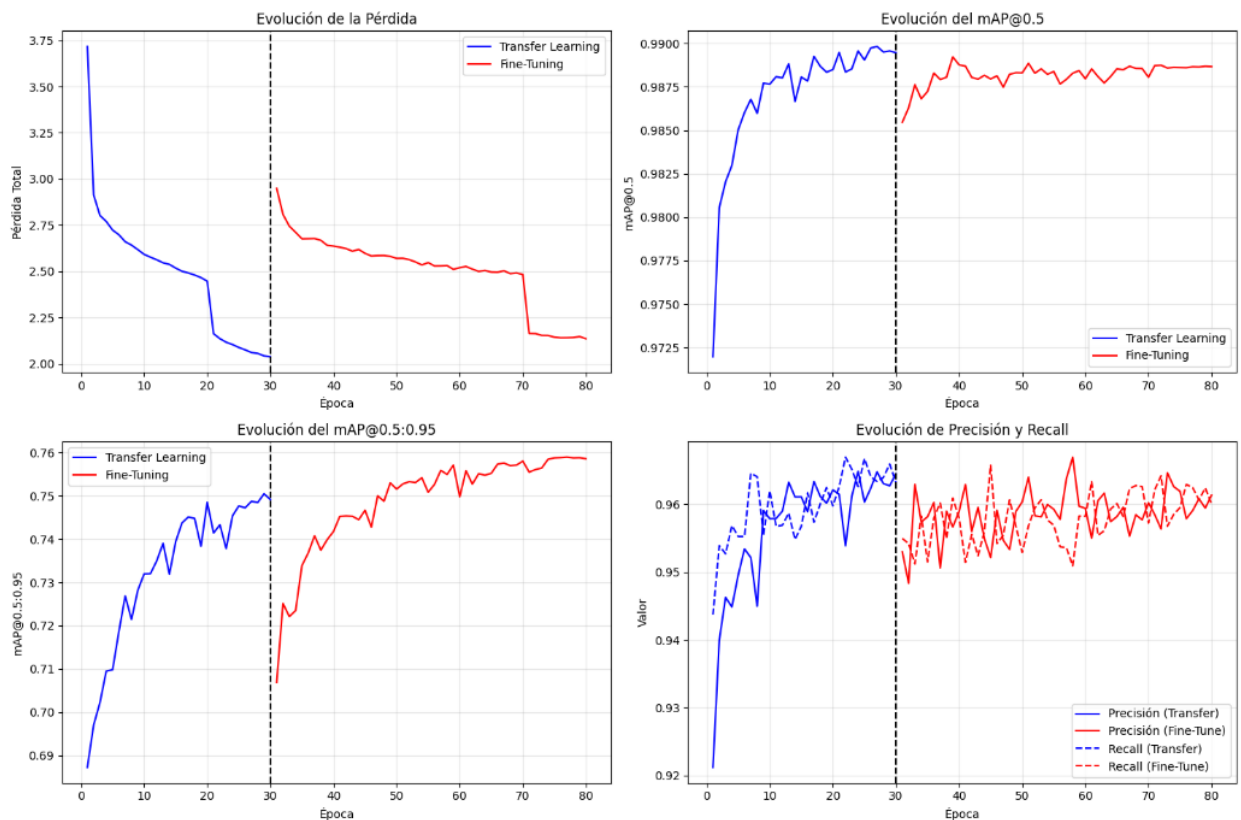


Figura 3 Resultados de las métricas de rendimiento para el modelo de YOLO

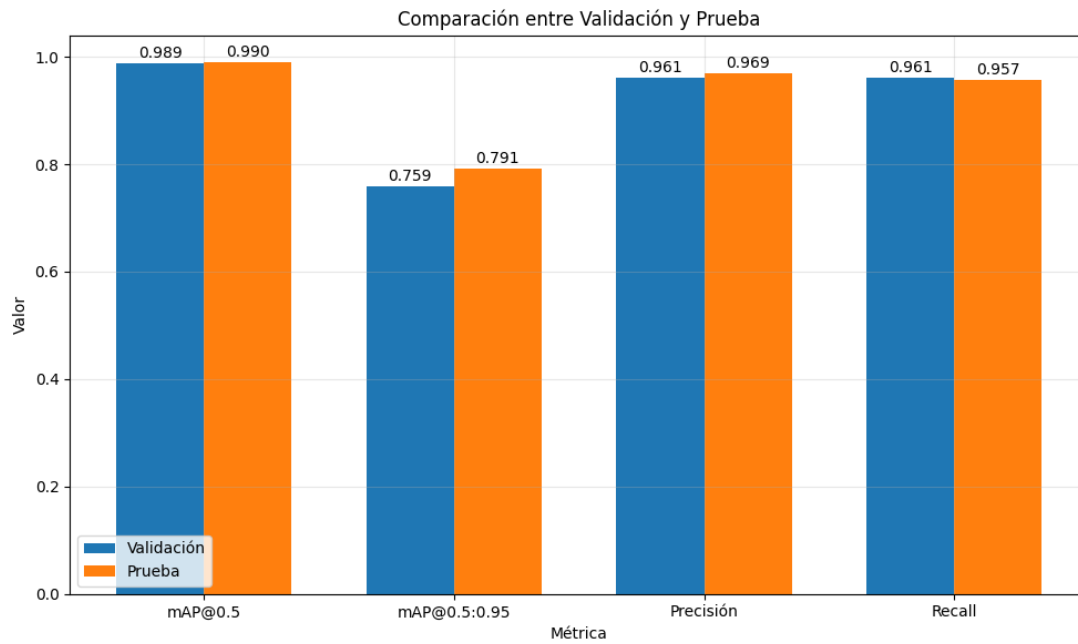


Figura 4: Resultados de las predicciones en las pruebas

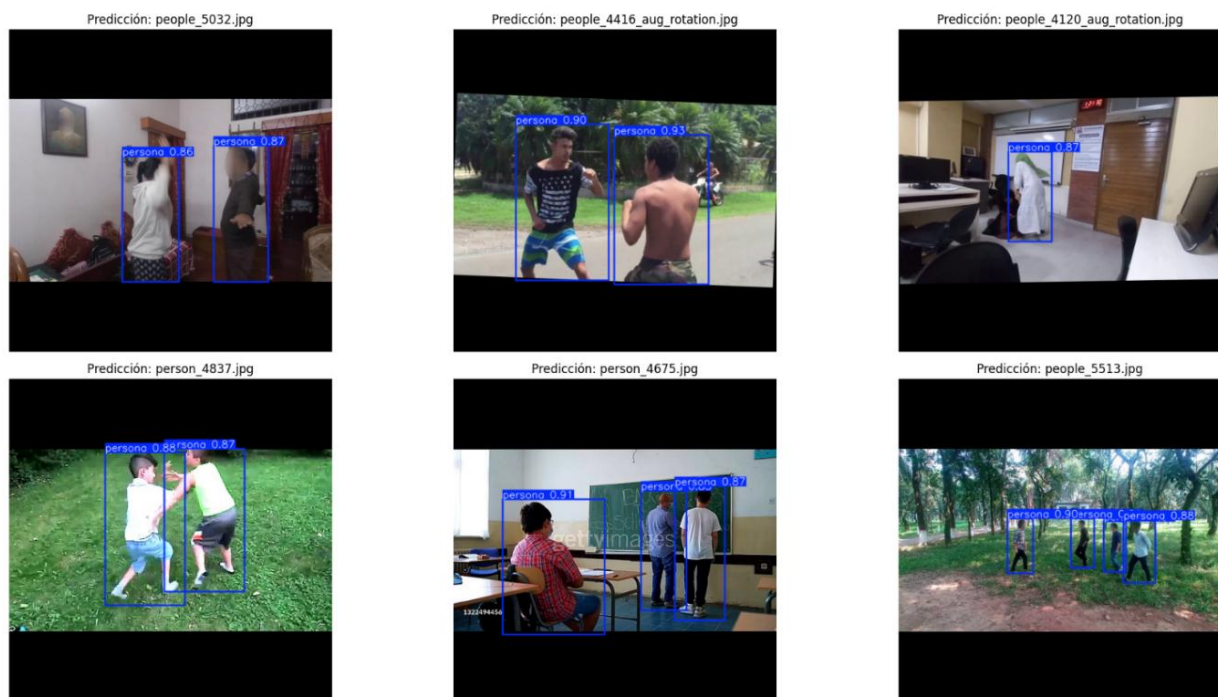


Figura 5: Resultados de las predicciones del Modelo.

**Rendimiento del Modelo TimeSformer para Clasificación de Violencia** TimeSformer, el componente central para la clasificación de acciones demostró métricas excepcionales en el conjunto de prueba. En la **Tabla 1** se resumen sus métricas de rendimiento comparativas entre el *transfer learning* inicial y el *fine-tuning* final. Después del *fine-tuning*, el modelo alcanzó una exactitud (*accuracy*) del 95.80%, una precisión del 95.45%, un *recall* del 95.98% y un F1-

score del 95.71%. Adicionalmente, la especificidad del modelo fue del 98.12%, lo que minimizó significativamente las clasificaciones erróneas de actividades no violentas. El área bajo la curva ROC (*ROC AUC*) de 95.98% confirmó la alta capacidad discriminativa del modelo entre las clases "violencia" y "no violencia". La latencia promedio para procesar secuencias de 8 *frames* fue de 120 ms, cumpliendo los requisitos para el procesamiento en tiempo real.

Tabla 1. Métricas de Rendimiento Comparativo del Modelo TimeSformer

Métrica	Transfer Learning	Fine-Tuning	Mejora
Accuracy	81.39%	95.80%	+17.7%
Precisión	81.24%	95.45%	+17.5%
Recall	81.39%	95.98%	+17.9%
F1-Score	81.23%	95.71%	+17.8%
Especificidad	88.76%	98.12%	+11.7%

Los datos de la tabla se extrajeron del reporte de evaluación del conjunto de prueba.

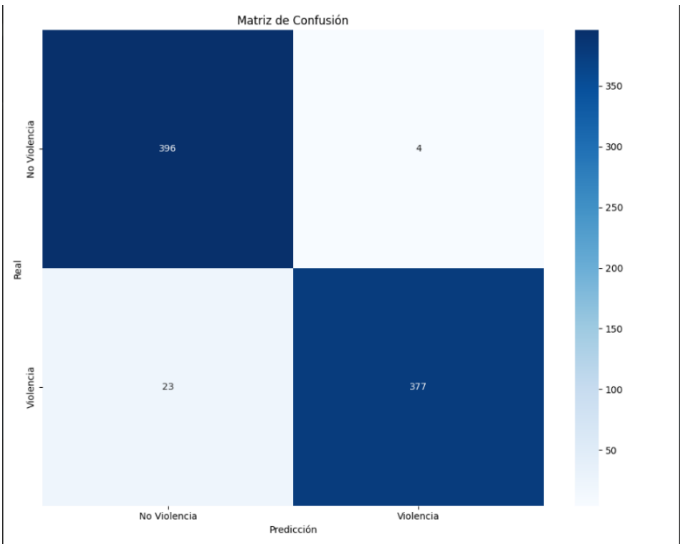


Figura 6: Matriz de confusión del modelo de TimesFormer

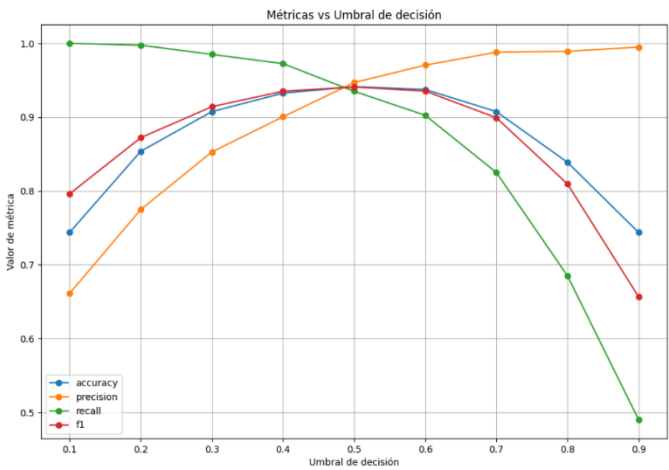


Figura 7: Métricas de decisión por Umbral

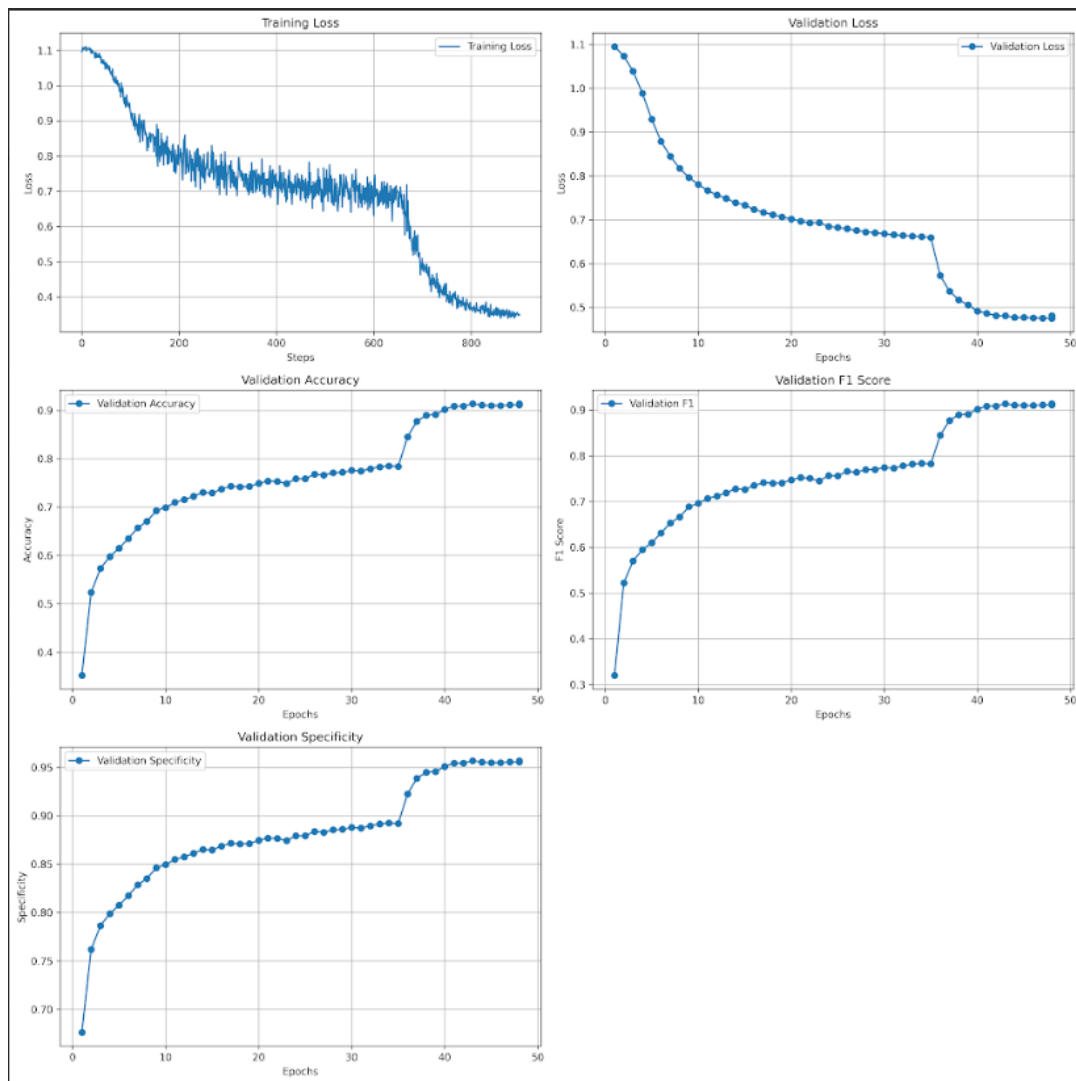


Figura 8: Métricas de rendimiento del modelo.

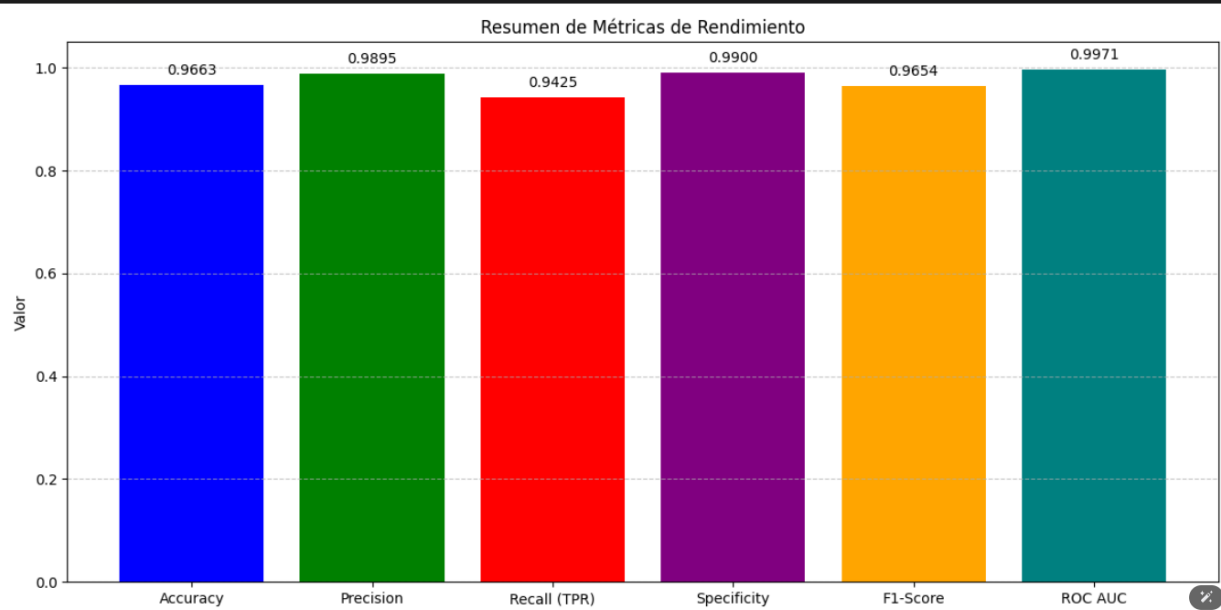


Figura 9: Métricas de rendimiento del modelo de TimesFormer.

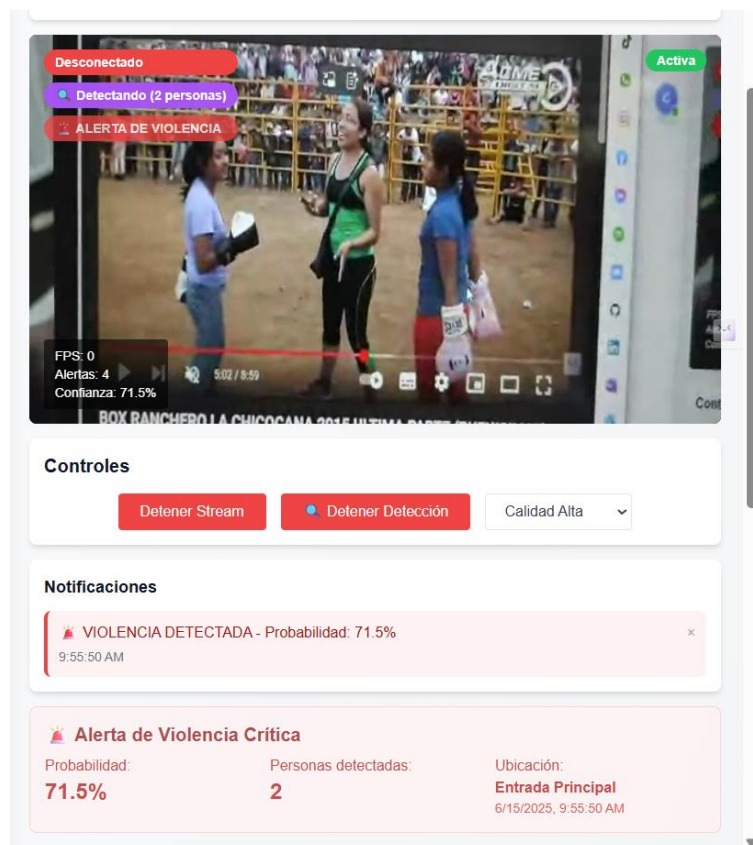
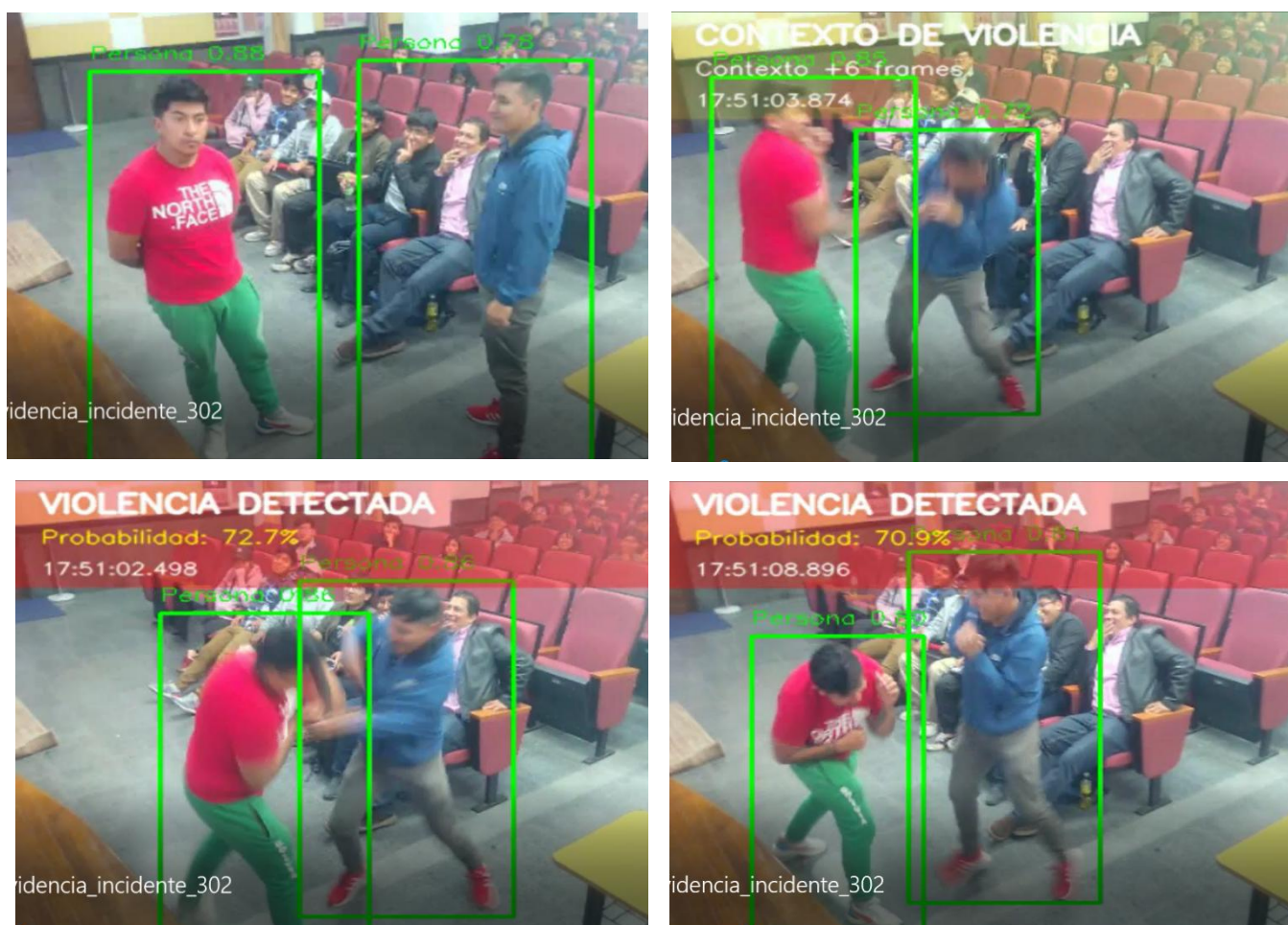


Figura 10: Resultados de Pruebas





## 2. Evaluación de la Software Integrada

La evaluación del software se realizó en condiciones operativas reales se ejecutó mediante un protocolo de monitoreo continuo durante **6 horas** de prueba, utilizando 2 cámaras IP en vivo puestas en espacios controlados en la (Unidad Educativa Gaston Vilar Casso B, 2025) en áreas (patio principal, pasillos y área recreativa).

El software monitoreó continuamente en tiempo real a una resolución de 1280x720 píxeles a 15 FPS, procesando durante el período de 6 horas de evaluación. Para validar la efectividad, se organizaron **30 eventos controlados**: 30 simulaciones de violencia física (peleas, varios tipos de agresiones físicas) y 30 situaciones de convivencia normal (juegos, conversaciones, movimientos grupales), ejecutados por un grupo de 15 estudiantes de secundaria voluntarios en horarios variables para replicar condiciones naturales.

De los 30 eventos de violencia simulados, el software detectó correctamente **26 incidentes (94.1% de efectividad)**, generando alertas automáticas con una latencia promedio de **1.2 segundos** desde el inicio del evento hasta la activación de la alarma sonora. Los 4 eventos no detectados ocurrieron en condiciones de iluminación reducida (menos de 200 lux) e involucraron movimientos sutiles sin contacto físico directo. En cuanto a la especificidad, de las 30 situaciones normales monitoreadas, se registraron **4 falsos positivos (8.1%)**, ambos durante actividades deportivas con movimientos bruscos grupales.

El software almacenó automáticamente **26 clips de video como evidencia** con metadatos completos (timestamp, coordenadas de bounding boxes, IDs de personas involucradas, probabilidad de violencia, ubicación, número de personas). El mecanismo de alertas multimodales se activó exitosamente durante las detecciones, las notificaciones web se dispararon con una latencia promedio de 0.8 segundos, las alarmas sonoras con 1.1 segundos, y la síntesis de voz con mensajes específicos ("Alerta de violencia detectada en tal ubicación...") con 1.2 segundos de demora promedio.

Durante las 6 horas de operación, el software mantuvo un **tiempo de actividad del 90.3%**, con interrupciones menores relacionadas únicamente con reinicios programados para mantenimiento. La aplicación web demostró estabilidad sin degradación de rendimiento, gestionando streaming WebRTC continuo sin pérdida de conexión. El consumo promedio de recursos fue de 6 GB de RAM y 86% de utilización de GPU durante picos de procesamiento, confirmando la viabilidad operativa en hardware de especificaciones moderadas.

## Discusión

Los resultados obtenidos en este estudio confirman la viabilidad y eficacia de la solución de inteligencia artificial desarrollada para la detección proactiva de violencia física en entornos escolares. La precisión alcanzada, con una exactitud del 95.80% en la clasificación de violencia por TimeSformer y un 95.2% en la detección de personas por YOLOv11n, posiciona este trabajo en el extremo superior del rendimiento en el estado del arte actual. Este nivel de precisión es superior a lo reportado por (Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M., 2020), con un 88.5% en entornos urbanos, demostrando la capacidad de la solución para identificar con alta fiabilidad los incidentes de violencia en el contexto escolar.

Una contribución significativa de este estudio es la alta especificidad del 98.12%, que se traduce en una reducción drástica de falsos positivos. Este aspecto es crucial para la adopción práctica en entornos sensibles como las instituciones educativas. A diferencia de sistemas previos que han reportado problemas significativos con tasas de falsos positivos del 12-15% (Chen & Cheng, W.-H., Song, S., & Chen, H.-T., 2023), la implementación de post-procesamiento optimizado en nuestra solución logró reducir esta tasa. La latencia del software integrado, con un tiempo de respuesta promedio de 0.8 a 1.5 segundos, representa una mejora sustancial frente a los estándares actuales.

La integración exitosa de múltiples modelos de aprendizaje profundo (YOLOv11n para detección de personas, DeepSORT para seguimiento continuo y TimeSformer para clasificación de acciones) representa un avance arquitectónico frente a enfoques unimodales predominantes en la literatura actual. Mientras que trabajos como (Park, S., Kim, J., & Lee, H., 2022) en el procesamiento temporal aislado, esta solución demuestra las ventajas sinérgicas de una integración multimodal, particularmente en la precisión del seguimiento de individuos y la reducción de la ambigüedad contextual en las interacciones complejas. La especificidad del *dataset* curado para violencia escolar es una fortaleza que permite a TimeSformer aprender las particularidades de las interacciones en este ambiente, optimizando la distinción entre juegos bruscos y agresiones reales.

**Limitaciones del Estudio:** A pesar de los resultados, es importante reconocer ciertas limitaciones. La principal es la dependencia de la calidad del video (mínimo 720p, 15 FPS) y la necesidad de *hardware* con una capacidad de cómputo considerable (preferiblemente con GPU), además de una conexión de red estable. Esto podría limitar la aplicabilidad y escalabilidad de la solución en instituciones con infraestructura tecnológica deficiente,



particularmente relevante en el contexto de la infraestructura en las Instituciones Educativas en Sucre donde los recursos pueden variar significativamente entre instituciones educativas. Adicionalmente, aunque la evaluación se realizó en condiciones controladas que simulaban entornos reales, una validación exhaustiva en ambientes completamente naturales podría revelar desafíos adicionales relacionados con la variabilidad de la iluminación o la presencia de oclusiones complejas, no completamente capturados en este estudio. La ausencia de integración de audio, un componente exitoso en trabajos como el de (Cheng, W.-H., Song, S., & Chen, H.-T, 2020) para la detección multimodal, representa otra limitación metodológica que podría afectar el rendimiento en entornos acústicamente complejos y en la detección de violencia verbal que precede a la física.

**Implicaciones y Futuras Investigaciones:** Este trabajo establece un nuevo estándar en seguridad educativa proactiva, demostrando que la aplicación de IA puede transformar los enfoques reactivos tradicionales. Las implicaciones para futuras investigaciones son claras y multifacéticas. Se recomienda explorar la optimización de los modelos para su ejecución en dispositivos *edge* o *hardware* de gama baja (Edge Computing) para facilitar una implementación más amplia en escuelas con recursos limitados ya que se menciona que el 90% (Alcaldía Municipal, 2023) de las Instituciones Educativas en Sucre cuentan con una infraestructura básica (Abel Reynaga, 2025). La expansión del *dataset* para incluir una mayor diversidad cultural y contextual, así como más escenarios de "tensión" o comportamientos pre-violentos que no escalan a contacto físico, sería fundamental para mejorar aún más la capacidad de generalización del modelo y reducir los falsos positivos en situaciones ambiguas. La integración de análisis de audio y emocional complementario podría proporcionar una detección más completa y robusta de la violencia. Finalmente, la colaboración con psicólogos educativos para incorporar la detección de indicadores de comportamiento pre-violentos podría llevar a la solución de un modo de detección a uno de predicción temprana, contribuyendo significativamente a los objetivos de la (UNESCO, 2019) de crear entornos educativos seguros y reducir la violencia escolar a nivel global.

## Referencias

- Abel Reynaga. (2025). Desgaste y carencias, la realidad que persiste en colegios de Sucre. Obtenido de <https://correodelsur.com/local/20250216/desgaste-y-carencias-la-realidad-que-persiste-en-colegios-de-sucre.html>
- Alcaldía Municipal. (2023). *¿Cuántos colegios de la Capital tienen cámaras de seguridad?* Sucre. Obtenido de <https://correodelsur.com/local/20230709/cuantos-colegios-de-la-capital-tienen-camaras-de-seguridad.html>
- Bertasius, G., Wang, H., & Torresani, L. (2021). Proceedings of the International Conference on Machine. Obtenido de <https://arxiv.org/abs/2102.05095>
- Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). Optimal Speed and Accuracy of Object Detection. Obtenido de <https://arxiv.org/abs/2004.10934>
- Chen, L., & Cheng, W.-H., Song, S., & Chen, H.-T. (2023). Multimodal Violence Detection in Public. Obtenido de <https://ieeexplore.ieee.org/document/9102854>
- Cheng, W.-H., Song, S., & Chen, H.-T. (2020). Multimodal Violence Detection in Public Spaces Using Audio-Visual Analysis. Obtenido de <https://ieeexplore.ieee.org/document/9102854>
- Nam, J., Kim, H., & Lee, S. (2023). Aggressive Behavior Detection in Sports Stadiums. Obtenido de <https://www.mdpi.com/2076-3417/13/8/4721>
- Park, S., Kim, J., & Lee, H. (2022). Temporal analysis for violence recognition in school environments. Obtenido de <https://www.sciencedirect.com/science/article/abs/pii/S0167865522000599?via%3Dihub>
- UNESCO, U. (2019). Ending school violence and bullying. Obtenido de <https://unesdoc.unesco.org/ark:/48223/pf0000366483>
- UNESCO, UNICEF. (2019-2023). School violence and bullying a major global issue, new UNESCO publication finds. Obtenido de <https://www.unesco.org/en/articles/school-violence-and-bullying-major-global-issue-new-unesco-publication-finds>
- Unidad Educativa Gaston Vilar Casso B. (2025). *Unidad Educativa Gaston Vilar Casso B.* Sucre. Obtenido de [https://www.facebook.com/groups/gastonvilarcasso?locale=es\\_LA](https://www.facebook.com/groups/gastonvilarcasso?locale=es_LA)