Simposio CEA de Robótica, Bioingeniería, Visión por Computador y Automática Marina 2026

Detección local de caídas y aviso visual-auditivo con tecnología accesible

Sánchez-Girón, C.[a,*], García-Gómez, M.[a], Duque-Domingo, J.[a], Gómez-García-Bermejo, J.[a,b], Zalama, E.[a,b]

[a] *DISA-ITAP, Universidad de Valladolid, C/ Dr Mergelina s/n, 47011 Valladolid, España*
[b] *CARTIF Centro Tecnológico 47151, Valladolid, España*

Resumen

El envejecimiento demográfico plantea importantes retos para los sistemas de salud y asistencia. La necesidad de encontrar soluciones que garanticen la seguridad de los mayores sin depender de suscripciones temporales se hace cada vez mayor. Este artículo presenta un sistema integral de detección mediante visión artificial y alerta ante caídas empleando dispositivos cotidianos. A diferencia de soluciones basadas en la nube o wearables, nuestro sistema opera de forma local garantizando la privacidad. Proponemos una solución novedosa y ligera basada en detección de pose y estimación con LSTM que realiza una predicción en tiempo real. El sistema incluye un mecanismo dual de notificaciones de emergencia mediante mensajería (WhatsApp) y asistentes de voz (Alexa). Los resultados demuestran una alta eficiencia computacional, reduciendo drásticamente las falsas alarmas y permitiendo su despliegue en hardware de bajo coste.

Palabras clave: Detección de caídas, Teleasistencia, Internet de las cosas, Sensorización, Sistemas centrados en el ser humano

Local fall detection and audio-visual alerting with accessible technology

Abstract

Demographic aging poses significant challenges for healthcare and assistance systems. There is a growing need to find solutions that ensure the safety of the elderly without relying on recurring subscriptions. This article presents a comprehensive detection and fall-alert system using computer vision and everyday devices. Unlike cloud-based or wearable solutions, our system operates locally, ensuring privacy. We propose a novel, lightweight solution based on pose detection and LSTM estimation that performs real-time predictions. The system includes a dual emergency notification mechanism via messaging (WhatsApp) and voice assistants (Alexa). The results demonstrate high computational efficiency, drastically reducing false alarms and allowing for deployment on low-cost hardware.

Keywords: Fall detection, Tele-medicine, Internet of things, Perception and sensing, Human-centered systems engineering

1. Introducción

España es uno de los países más envejecidos de Europa, con más del 20 % de su población superando los 65 años, cifra que se estima que rondará el 30 % en 2050 (Enfermería, 2024). En este sector demográfico, las caídas representan la primera causa de muerte accidental y generan más de 250.000 urgencias anuales. Dado que el 80 % de las personas mayores prefiere envejecer en su hogar (Rural, 2022), existe una necesidad crítica de sistemas de monitorización discretos y fiables.

La teleasistencia tradicional, basada en pulsadores o dispositivos colgantes, presenta altas tasas de abandono. Además, en situaciones de pérdida de conciencia, el usuario no puede activar la alarma (Ferguson et al., 2021). El mercado demanda soluciones inteligentes autónomas basadas en visión e IA local. En el presente artículo, se propone un sistema accesible y fiable de detección y alerta ante caídas que no requiere *wearables*, no almacena vídeo permanentemente ni lo envía a la nube, preservando la privacidad del usuario en su entorno doméstico. El sistema ha sido diseñado con el objetivo de integrar la domótica cotidiana de un hogar y aplicaciones de comunicación populares, como Whatsapp, para que los familiares y cuidadores sean informados del estado de la persona monitorizada sin necesidad de descargar software o contratar servicios adicionales.

El sistema se enmarca dentro del sector de teleasistencia, domótica avanzada y envejecimiento activo, uno de los merca-

*Autor para correspondencia: celia.sanchez-giron@uva.es

dos de más rápido crecimiento en la Unión Europea debido al envejecimiento demográfico y la presión sobre los servicios de cuidados (Euronews, 2025).

2. Trabajos relacionados

Aproximadamente un tercio de las personas mayores de 65 años sufre al menos una caída al año, cifra que puede aumentar hasta la mitad en mayores de 80 años, reflejando el fuerte incremento del riesgo con la edad (Ministerio de Sanidad, 2025). En este contexto, un gran número de investigaciones en los últimos años han estado orientadas a detectar y alertar sobre caídas en tiempo real, para que las personas mayores que vivan solas en sus hogares puedan tener la seguridad de que se detectará y avisará en el momento del accidente.

En el ámbito de la detección de caídas, la visión por computador se ha consolidado como una línea clave para la monitorización no intrusiva, especialmente cuando no solo se busca reconocer una postura, sino también la evolución temporal del evento. En este contexto, los enfoques espaciotemporales basados en LSTM (Long Short-Term Memory) han adquirido protagonismo, ya que permiten modelar transiciones entre actividades, como caminar, perder el equilibrio, impactar contra el suelo y permanecer inmóvil, reduciendo así los falsos positivos asociados a posturas estáticas o movimientos voluntarios. Un ejemplo representativo combina estimación ligera de pose con un LSTM para capturar esta dinámica temporal, obteniendo altas tasas de acierto en entornos domésticos e inteligentes (Bourenane and Henni, 2025). Asimismo, se han propuesto arquitecturas multimodales en las que variantes como Bi-LSTM procesan señales temporales, como sensores inerciales, y se fusionan con información visual para mejorar la robustez (Shin et al., 2025). No obstante, aunque los sensores *wearable* e inerciales pueden ofrecer buenas prestaciones, dependen de que la persona los lleve consigo de forma continua, lo que limita su aplicabilidad en personas mayores. Por su parte, los sensores radar presentan ventajas frente a oclusiones, cambios de iluminación y privacidad, pero pueden generar falsos positivos ante pequeñas variaciones del entorno (Igual et al., 2013).

Por otro lado, para la gestión y atención de la alerta, las soluciones comerciales actuales dependen de servicios en la nube para ejecutar pesados modelos neuronales, lo que supone una limitación para el usuario al depender de una suscripción temporal para garantizar su seguridad.

Nuestro trabajo aborda estas limitaciones al procesar el flujo de vídeo (RTSP/MJPEG) mediante un nodo embebido local, cumpliendo así los principios del RGPD de privacidad por diseño.

3. Metodología

La arquitectura general del sistema se muestra en la Figura 1, que consiste en la integración de un sistema de captura de imagen (Raspberry Pi3B +webcam), un ordenador o unidad de procesamiento donde se ejecute el modelo de caídas y los elementos de aviso ante alerta (Amazon Echo y aplicación móvil Whatsapp) que hemos elegido por ser tecnologías muy utilizadas hoy en día. Dentro del PC se ejecuta el software principal, que realiza tres funciones encadenadas:

1. **Extracción de Postura**: En lugar de procesar la imagen completa o un recorte, que supondría trabajar con miles de píxeles y, por ende, una alta carga computacional, extraemos los puntos clave (articulaciones) o *keypoints* del usuario en cada frame utilizando un modelo de detección de pose.
2. **Modelo de detección**: Las coordenadas normalizadas de los *keypoints* alimentan una red recurrente (LSTM) que ha sido entrenada con un conjunto de datos. Gracias a la dinámica del movimiento, el modelo realiza una predicción sobre la situación de la persona.
3. **Aviso de emergencia**: si el sistema confirma una caída (según la lógica temporal establecida), se activa el protocolo de alerta.

3.1. Adquisición de Imagen

Una cámara web convencional conectada a una Raspberry Pi genera un flujo en la red local. Esta cámara actúa como una "cámara IP doméstica" sin exposición a Internet. De este modo, el vídeo queda accesible mediante streaming para otros dispositivos de la red. Cada fotograma RGB $I_t \in \mathbb{R}^{H\times W\times 3}$ se procesa mediante el detector YOLO11 Pose, escogido por su balance entre rendimiento y carga computacional (Roboflow, 2025), que estima las articulaciones del esqueleto humano. Para el sujeto principal detectado, se obtiene el conjunto:

$$\mathcal{P}_t = \{(x_{t,i}, y_{t,i})\}_{i=1}^{K}$$

Estas coordenadas se normalizan con respecto a las dimensiones de la imagen, para reducir la dependencia de la resolución de entrada y las coordenadas absolutas de los píxeles, dividiendo por el ancho W y el alto H del fotograma:

$$x_{t,i}^{(n)} = \frac{x_{t,i}}{W}, \qquad y_{t,i}^{(n)} = \frac{y_{t,i}}{H}$$

A partir de aquí se forma un vector por fotograma $V_t \in \mathbb{R}^{2K}$ (en este caso $2K = 34$) concatenando las coordenadas normalizadas:

$$V_t = [x_{t,1}^{(n)}, y_{t,1}^{(n)}, \ldots, x_{t,K}^{(n)}, y_{t,K}^{(n)}]$$

Cuando el detector no devuelve una persona válida en un frame, el sistema introduce un vector nulo $V_t = \mathbf{0} \in \mathbb{R}^{34}$. Esta decisión de diseño permite mantener una secuencia temporal consistente, tolerando pérdidas de detección puntuales sin necesidad de interpolación.

3.2. Arquitectura empleada

La arquitectura diseñada sustituye el procesamiento intensivo de píxeles por un análisis de trayectorias articulares, reduciendo drásticamente la carga computacional. Se compone de dos etapas principales:

En cada fotograma t, el modelo YOLO11 Pose procesa la imagen y extrae un conjunto de 17 puntos clave (*keypoints*) del cuerpo humano. Cada punto i está definido por sus coordenadas bidimensionales (x_i, y_i). Para que el modelo sea invariante a la resolución de la cámara, las coordenadas se normalizan respecto al ancho (W) y alto (H) del fotograma:

$$x_{norm,i} = \frac{x_i}{W}, \quad y_{norm,i} = \frac{y_i}{H}$$

Esto genera un vector de características espaciales $V_t \in \mathbb{R}^{34}$ por cada fotograma.

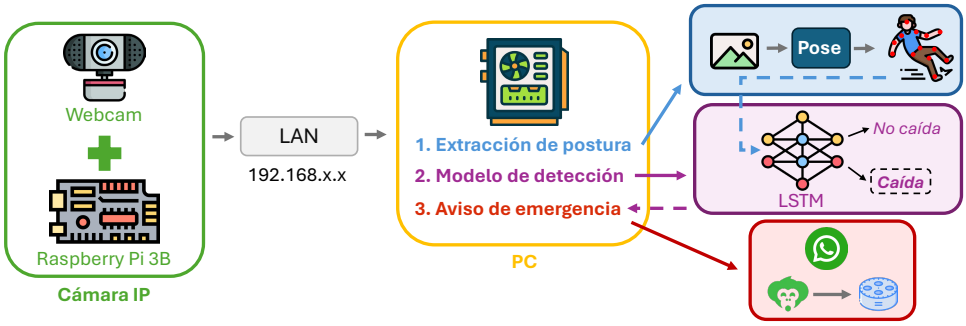


Figura 1: Arquitectura general del sistema detección-alerta

Una caída se caracteriza por su dinámica temporal, por lo que el sistema utiliza una ventana deslizante de longitud $N = 30$ frames. La entrada al modelo en el instante t es:

$$S_t = [V_{t-N+1}, \ldots, V_t] \in \mathbb{R}^{N\times 34}$$

El clasificador temporal se implementa como una red LSTM con 2 capas y tamaño oculto 64. El estado oculto y de celda se inicializa en cero para cada secuencia procesada:

$$h_0 = \mathbf{0} \in \mathbb{R}^{L\times B\times h}, \qquad c_0 = \mathbf{0} \in \mathbb{R}^{L\times B\times h}$$

donde B es el tamaño del batch. La LSTM produce una salida por paso temporal; sin embargo, el sistema toma únicamente la representación del último instante, h_t, como resumen espacio-temporal de la secuencia:

$$\hat{z}_t = \text{LSTM}(S_t) \Rightarrow \hat{z}_t = h_t \in \mathbb{R}^{64}$$

La salida final se obtiene mediante una cabeza totalmente conectada que traduce el embedding temporal ($\mathbb{R}^{64}$) en dos clases: no-caída (0) y caída (1). Siguiendo el código, se emplea:

- una capa densa de 32 neuronas,
- activación ReLU,
- *Dropout* $p = 0{,}3$ como regularización,
- y una última capa lineal a 2 logits.

Formalmente:

$$\mathbf{o}_t = W_2(\text{Dropout}(\text{ReLU}(W_1\hat{z}_t + b_1))) + b_2$$

$$\hat{y}_t = \arg\max_{c\in\{0,1\}} \text{Softmax}(\mathbf{o}_t)_c$$

La Tabla 1 resume las decisiones de arquitectura implementadas.

El sistema propuesto de detección de caídas se ejecutó en un ordenador de sobremesa equipado con un procesador Intel Core i5-12400, 32 GB de memoria RAM, una unidad SSD de 256 GB y gráficos integrados, con un precio aproximado de 649 euros. Esta configuración puede considerarse una plataforma hardware de coste accesible, especialmente en comparación con estaciones de trabajo que requieren una GPU dedicada para tareas de visión artificial en tiempo real. Aunque los experimentos presentados en este trabajo se llevaron a cabo sobre esta máquina, el sistema está concebido para poder desplegarse en otros equipos con especificaciones similares. En particular, la configuración mínima recomendada para un funcionamiento práctico sin GPU dedicada incluiría un procesador actual de gama media, al menos 32 GB de RAM y almacenamiento basado en SSD.

3.3. *Protocolo automático de emergencias*

Para minimizar los falsos positivos se ha establecido como desencadenante del protocolo de emergencias que pasen 10 segundos desde el momento del accidente hasta que se inicia el protocolo. Al confirmarse el episodio, se lanza un mecanismo dual:

- **Alerta Visual:** Envío automático de un mensaje de WhatsApp (vía *pywhatkit*) al cuidador con una instantánea del frame y la confianza del modelo.
- **Alerta Auditiva y Telefónica:** Activación de una rutina mediante un *trigger* que emite un aviso por voz en un dispositivo Amazon Echo (“He detectado que te has caído”) y realiza una llamada manos libres al contacto de emergencia.

Para la alerta por mensaje, es necesario iniciar previamente sesión en WhatsApp Web en el PC donde se ejecuta el sistema. Una vez vinculado el móvil, el protocolo de emergencia abre automáticamente la web y envía un mensaje junto con una captura del posible accidente al número configurado. Por ello, el cuidador debe tener WhatsApp instalado en el teléfono asociado.

Además, se ha implementado una alerta auditiva mediante VoiceMonkey y Alexa. En este caso, VoiceMonkey actúa como *trigger* y activa, a través de una petición HTTP, una rutina de Alexa previamente configurada. Como resultado, el dispositivo Amazon Echo instalado en el domicilio emite un aviso de voz indicando que se ha detectado una caída e inicia una llamada en manos libres con el contacto de emergencia, que debe tener instalada la aplicación Alexa en su móvil.

Una de las principales ventajas de esta solución es que la persona monitorizada no necesita desplazarse hasta un teléfono para pedir ayuda, sino que puede comunicarse directamente a través del altavoz inteligente. Dado que la aplicación Alexa no permite activar rutinas directamente desde software externo, se recurrió a VoiceMonkey como servicio intermediario, ya que permite lanzar dichas rutinas mediante peticiones HTTP desde un script en Python. La Figura 2 muestra la configuración empleada en ambos servicios.

Tabla 1: Configuración de la arquitectura espaciotemporal basada en *keypoints* (según implementación).

Componente	**Configuración**
Extractor de pose	YOLO11 Pose
# *keypoints*	$K = 17$ (34 coordenadas 2D)
Normalización	$x/W,\ y/H$ por fotograma
Ventana temporal	$N = 30$ frames (*sliding window*)
Modelo temporal	LSTM, $L = 2$ capas, $h = 64$
Cabeza de clasificación	FC(64→32) + ReLU + Dropout(0.3) + FC(32→2)
Clases	0: No-caída, 1: Caída
Manejo de frames sin detección	Vector nulo $\mathbf{0} \in \mathbb{R}^{34}$

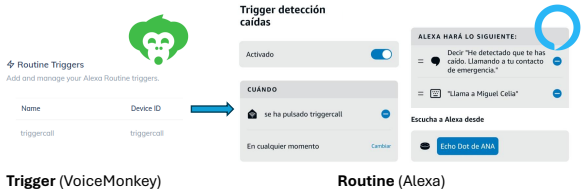


Figura 2: Sincronización servicios de alerta ante caídas (Alexa y VoiceMonkey)

4. Experimentación

4.1. Entrenamiento

A la hora de entrenar nuestro modelo, se ha seleccionado como conjunto de datos el URFall dataset (Kwolek and Kepski, 2014), que contiene 100 vídeos RGB de los cuales 60 tienen eventos de caída y 40 actividades diarias, tomados a 25 frames por segundo. Estos vídeos fueron grabados con dos cámaras Kinetic, junto con datos de acelerómetros y otros sensores inerciales.

El entrenamiento del módulo temporal (LSTM) se realizó a partir de secuencias generadas automáticamente mediante la extracción de *keypoints* en vídeo y su alineación con anotaciones frame-a-frame proporcionadas en un fichero CSV.

Cada frame se procesa con el modelo extractor de pose para obtener el esqueleto humano. Si existe una estimación válida, se construye el vector $V_t \in \mathbb{R}^{34}$ con coordenadas normalizadas. En caso contrario, se asigna el vector nulo. Esta estrategia es útil en escenarios reales donde pueden ocurrir pérdidas de detección (movimiento rápido, oclusiones, desenfoque), evitando romper la continuidad temporal de la ventana.

Con el flujo de vectores por frame $\{V_t\}$ se generan ventanas deslizantes de longitud fija $N = 30$:

$$S_t = [V_{t-29}, \ldots, V_t]$$

La etiqueta objetivo de cada ventana se selecciona como la etiqueta del último frame de la ventana:

$$y_t = \texttt{label}(t)$$

Este diseño es coherente con la detección en tiempo real: el sistema decide en el instante actual utilizando el historial reciente.

En cuanto al tratamiento de clases, el código implementa una conversión de etiquetas a clasificación binaria, de forma que el conjunto final de entrenamiento queda definido como:

$$\mathcal{D} = \{(S_t, y_t)\}, \quad y_t \in \{0, 1\}$$

donde $y_t = 0$ representa no-caída y $y_t = 1$ representa caída.

El modelo fue entrenado con ventanas de 30 frames en datasets a 25 fps, por lo que cada ventana tiene una duración de 1.2 segundos.

La Tabla 2 resume los hiperparámetros implementados durante el entrenamiento.

4.2. Test

Para comprender cómo se comporta el modelo en escenarios desconocidos, se utilizó un nuevo dataset como conjunto de test, en particular el Le2i (Charfi et al., 2013) dataset. Le2i es un dataset de imágenes RGB estándar orientado a la detección de caídas. Consiste en 130 vídeos de escenas de caídas y no-caídas capturados a 25 frames por segundo, interpretados por 17 actores. Para este dataset, se empleó una sola cámara. La duración media de cada vídeo es en torno a 1 minuto, en el cual la persona realiza una actividad cotidiana o bien se cae y permanece en el suelo. Los vídeos han sido grabados en varias estancias domésticas con una elevada varianza entre eventos en los que ha habido eventos de caída y no-caída. Se ha escogido este dataset porque ofrece secuencias de vídeo significativas que pueden ocurrir en un entorno doméstico.

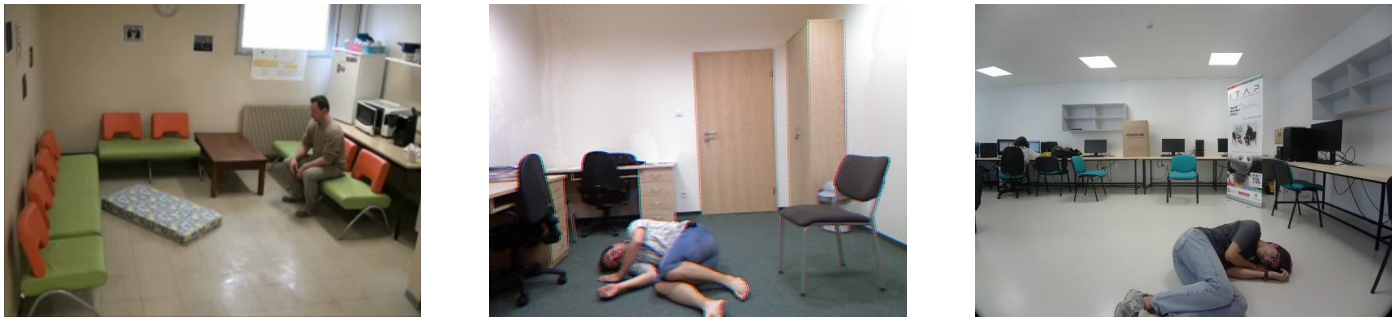
Además de este dataset público, realizamos diferentes pruebas con imágenes tomadas en condiciones controladas en un entorno de laboratorio con sus etiquetas correspondientes; en total, se adquirieron 5084 imágenes para ser evaluadas. Al realizar el test, tuvimos en cuenta que la altura de la cámara empleada fuera de 1.5 metros aproximadamente, altura que se utilizó para la captura de los vídeos durante el entrenamiento. Con esto intentamos que la configuración del entorno sea comparable de forma aproximada a la del dataset empleado.

En tiempo real el sistema funciona con una tasa variable entre 17 y 25 fps, por lo que la duración temporal efectiva de la ventana no siempre coincide con 1.2 segundos. Esto podría introducir una discrepancia temporal entre entrenamiento e inferencia. A pesar de esto durante las pruebas realizadas en tiempo real no se apreció ningun retardo apreciable en la detección ni anomalías considerables, por lo que no se ha utilizado ningún mecanismo explícito de alineación temporal; se tendrá en cuenta para trabajos futuros.

En la Figura 3 se muestra un ejemplo de un frame obtenida de cada conjunto de datos. Como se observa, la altura de la cámara es aproximadamente similar en las tres bases de datos escogidas.

Tabla 2: Hiperparámetros de entrenamiento del clasificador LSTM (según implementación).

Parámetro	**Valor**
Tamaño de entrada	34 (17 *keypoints* 2D)
Batch size	12
Épocas	100
Optimizador	Adam
Learning rate	1×10^{-3}
Función de pérdida	Entropía cruzada (`CrossEntropyLoss`)
Regularización	Dropout $p = 0{,}3$ en cabeza FC



(a) Dataset Le2i (b) Dataset UR Fall (c) Datos propios

Figura 3: Ejemplos de secuencias utilizadas en la evaluación del sistema.

4.3. Resultados

Las métricas obtenidas en ambos conjuntos de evaluación, recogidos en la Tabla 3, reflejan un comportamiento consistente y robusto del modelo ante escenarios distintos. En el dataset Le2i, evaluado sobre 38296 fotogramas, el sistema alcanza una *accuracy* cercana al 90 % y un *F1-score* superior al 80 %, lo que indica un equilibrio adecuado entre detecciones correctas y errores. La *precision* y el *recall* presentan valores relativamente próximos, lo que sugiere que el modelo mantiene una buena capacidad para detectar caídas sin generar un número excesivo de falsas alarmas, incluso en un entorno caracterizado por variabilidad en iluminación, puntos de vista y posibles oclusiones.

Tabla 3: Métricas de rendimiento en conjunto de test (%)

	Dataset Le2i	Datos propios
Accuracy	89.34	95.62
Precision	76.30	87.60
Recall	85.44	98.76
F1-Score	80.61	92.85

En el conjunto de datos que hemos recogido en el laboratorio, en total 5084 muestras, el rendimiento mejora de forma notable. La *accuracy* supera el 95 % y el *F1-score* se sitúa por encima del 90 %, evidenciando una clasificación más estable. Destaca especialmente el *recall*, que se aproxima al 100 %, lo que indica que el sistema apenas deja caídas sin detectar. Al mismo tiempo, la *precision* también aumenta respecto al dataset Le2i, lo que implica una reducción adicional de falsas detecciones.

En conjunto, estos resultados muestran una adecuada capacidad de generalización, junto con una adaptación especialmente eficaz a escenarios operativos reales.

La matriz de confusión normalizada del conjunto de test Le2i (Figura 4) indica que el 90.71 % de los casos no-caída se clasifican correctamente. Para la clase caída, el modelo identifica correctamente el 85.44 % de los eventos, aunque aún 14.56 % se etiquetan como no-caída. En conjunto, el mayor *recall* respecto a la *precision* sugiere una estrategia ligeramente más sensible, priorizando detectar caídas a costa de incrementar moderadamente los falsos positivos.

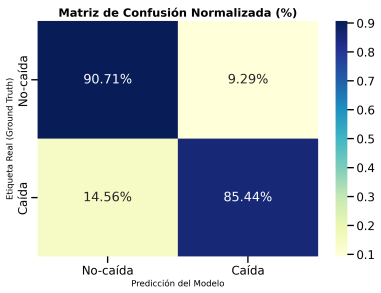


Figura 4: Matriz de confusión normalizada del conjunto de evaluación Le2i

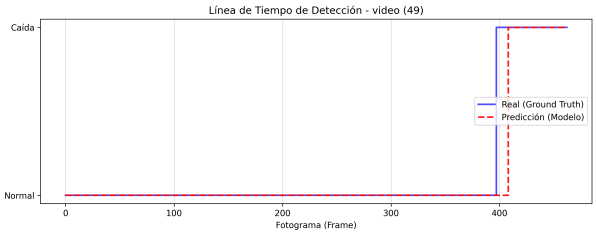


Figura 5: Línea de tiempo de detección. Desfase en el momento de la caída

El análisis de los falsos negativos mostró que estos se concentraban en los instantes inmediatamente anteriores a la caída. En varias secuencias, la detección se producía con unos pocos fotogramas de retraso respecto a la referencia, como se observa en la Figura 5. Esto indica que el sistema presenta cierta ambigüedad durante la transición entre no-caída y caída, una fase cuya etiquetación también puede resultar subjetiva.

En cuanto al 14.56 % de falsos positivos, aproximadamente dos tercios se debieron a errores de perspectiva, especialmente en situaciones como la entrada en escena o al sentarse en el sofá, debido a la mayor altura de la cámara en algunos vídeos. El tercio restante estuvo asociado igualmente a un ligero desfase temporal previo a la caída. Cabe destacar que estos errores no se observaron en nuestro entorno experimental, donde la cámara se situó a una altura similar a la del conjunto de entrenamiento.

En cuanto a la curva ROC, la Figura 6 muestra un comportamiento claramente superior al azar, con una trayectoria situada por encima de la diagonal de referencia y una buena capacidad discriminativa incluso para bajas tasas de falsos positivos.

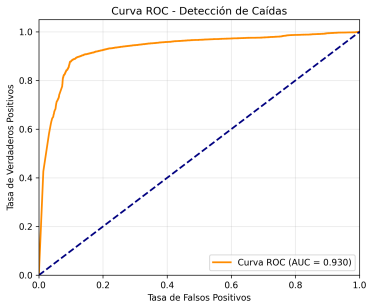


Figura 6: Curva ROC

A pesar de estos resultados, el sistema, basado únicamente en visión artificial, presenta las limitaciones propias de la monitorización mediante cámaras en entornos domésticos reales. Las caídas que ocurran fuera del campo de visión, detrás de muebles o bajo oclusiones prolongadas pueden no detectarse. Además, una cobertura completa del hogar requeriría varias cámaras, aumentando el coste, el tráfico de red y la carga de procesamiento. Por todo esto, es necesario explorar en trabajos futuros estrategias de percepción multimodal y fusión de sensores.

5. Conclusiones

Se ha presentado un sistema integral de detección y alerta ante caídas orientado a entornos domésticos y diseñado bajo criterios de privacidad, bajo coste y procesamiento local. La arquitectura propuesta combina estimación de pose con un modelo temporal LSTM capaz de capturar la dinámica del movimiento humano, evitando el procesamiento intensivo de píxeles. Esta aproximación basada en trayectorias articulares permite su despliegue en hardware accesible, facilitando su integración en escenarios reales.

Los resultados experimentales demuestran que el modelo mantiene un comportamiento sólido tanto en datasets públicos como en un conjunto de datos propio adquirido en un entorno controlado. En ambos casos se observa un equilibrio adecuado entre sensibilidad y precisión, con una elevada capacidad para detectar caídas reales. Además, una parte significativa de los falsos negativos se produce en la transición entre el estado no-caída y caída, momentos ambiguos desde el punto de vista del etiquetado humano, lo que refuerza la robustez del sistema.

Además del módulo de detección, el sistema incorpora un protocolo automático de emergencias que integra notificaciones visuales y auditivas mediante WhatsApp y Alexa, permitiendo una respuesta inmediata sin requerir intervención activa del usuario. Este enfoque resulta especialmente relevante en situaciones de pérdida de conciencia o imposibilidad de alcanzar un dispositivo móvil, aportando un valor añadido frente a soluciones tradicionales basadas en wearables o pulsadores.

En conjunto, la metodología presentada demuestra que es posible desarrollar un sistema de monitorización eficiente y capaz de ofrecer un rendimiento fiable en condiciones reales. Como líneas futuras de trabajo, se plantea ampliar la evaluación en entornos más diversos, incorporar estrategias de fusión multimodal y optimizar aún más la lógica temporal para reducir los errores en fases de transición.

Agradecimientos

Esta investigación ha recibido financiación del proyecto SOROCARE PID2024-157671OB-I00 financiado por MICIU/AEI/10.13039/501100011033/ FEDER, UE.

Referencias

Bourenane, S. A., Henni, S. A., 2025. Audio-visual multimodal fall detection to ensure the safety of elderly people in intelligent buildings: an innovative approach using lstm, cnn, and a shallow neural network. Signal, Image and Video Processing 19 (15), 1271.

Charfi, I., Miteran, J., Dubois, J., Atri, M., Tourki, R., 2013. Optimized spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and adaboost-based classification. Journal of Electronic Imaging 22 (4), 041106–041106.

Enfermería, C. G., 2024. España será uno de los países más envejecidos del mundo en 2050.
URL: https://www.consejogeneralenfermeria.org/actualidad-y-prensa/sala-de-prensa/noticias/item/8089-espana-sera-uno-de-los-paises-mas-envejecidos-del-mundo-en-2050

Euronews, Dec. 2025. El médico al teléfono: la teleconsulta gana terreno en europa.
URL: https://es.euronews.com/salud/2025/12/30/el-medico-al-telefono-la-teleconsulta-gana-terreno-en-europa

Ferguson, C., Hickman, L. D., Turkmani, S., Breen, P., Gargiulo, G., Inglis, S. C., 2021. "wearables only work on patients that wear them": Barriers and facilitators to the adoption of wearable cardiac monitoring technologies. Cardiovascular Digital Health Journal 2 (2), 137–147.

Igual, R., Medrano, C., Plaza, I., 2013. Challenges, issues and trends in fall detection systems. Biomedical engineering online 12 (1), 66.

Kwolek, B., Kepski, M., 2014. Human fall detection on embedded platform using depth maps and wireless accelerometer. Computer methods and programs in biomedicine 117 (3), 489–501.

Ministerio de Sanidad, G. d. E., 2025.
URL: https://estilosdevidasaludable.sanidad.gob.es/seguridad/caidas/mayores/home.htm

Roboflow, 2025. Best pose estimation models 2025.
URL: https://blog.roboflow.com/best-pose-estimation-models/

Rural, R. E. D., Dec. 2022. Me quedo en mi casa y en mi entorno.
URL: https://www.redr.es/es/noticias/la-mayoria-de-las-personas-mayores-en-espana-desean-quedarse-en-sus-casas-y-su-entorno-conclusiones-de-me-quedo-en-mi-casa-y-en-mi-entorno-un-proyecto-de-codinse/#:~:text=Quedarse%20en%20su%20casa%20y%20en%20su,a%C3%B1os%20seg%C3%BAn%20datos%20de%20la%20organizaci%C3%B3n%20de

Shin, J., Miah, A. S. M., Egawa, R., Hassan, N., Hirooka, K., Tomioka, Y., 2025. Multimodal fall detection using spatial–temporal attention and bi-lstm-based feature fusion. Future Internet 17 (4), 173.