Optimized Edge AI Framework for Wearable Seizure Monitoring on Microcontrollers

J. Navarro-Lázaro[1 [0009−0008−7132−6338]], M. Lupión[2 [0000−0001−7697−8062]], V. González-Ruiz[1 [0000−0001−6495−4856]], J.F. Sanjuan[1 [0000−0002−2874−0903]], and P.M. Ortigosa[1 [0000−0001−6514−6543]]

Universidad de Almería, 04120 Almería, Spain
jnl941@inlumine.ual.es marcoslupion@ual.es vruiz@ual.es jsanjuan@ual.es ortigosa@ual.es

Abstract. La monitorización fiable de las crisis en el mundo real sigue siendo un reto importante en la atención de la epilepsia, especialmente fuera de los entornos clínicos, donde la electroencefalografía (EEG) convencional resulta poco práctica. Las soluciones wearables existentes ofrecen un soporte parcial mediante modalidades de sensores limitadas y, a menudo, dependen de teléfonos inteligentes para el procesamiento, lo que limita la autonomía y la capacidad de respuesta en tiempo real. Además, las plataformas actuales carecen de soporte para la anotación clínica sincronizada y la generación de conjuntos de datos estructurados, ambos elementos esenciales para el entrenamiento y la validación de algoritmos de detección de crisis en las primeras fases de desarrollo. Para abordar estas limitaciones, presentamos un "framework" completo, de bajo consumo y de extremo a extremo basado en el Internet de las cosas (IoT) para la detección de crisis mediante dispositivos wearables. El sistema integra la detección fisiológica multimodal (actividad electrodérmica (EDA), electromiografía (EMG) y temperatura (TEMP)) con una arquitectura modular y en contenedores para la ingesta de datos en tiempo real, el almacenamiento histórico y la anotación por expertos.

A diferencia de soluciones anteriores, que son específicas para cada aplicación o están vinculadas a hardware propietario, nuestra plataforma permite una implementación flexible y admite flujos de trabajo de anotación desacoplados, lo que permite tanto a los equipos clínicos como a los investigadores desarrollar conjuntamente conjuntos de datos de alta calidad a partir de datos reales de pacientes. Partiendo de esta base, demostramos la implementación de un modelo de red neuronal convolucional (CNN) cuantificada capaz de realizar inferencias en menos de 15 milisegundos directamente en microcontroladores ESP32, lo que permite una detección autónoma y en tiempo real sin necesidad de procesamiento externo. Además, se introduce un método de posprocesamiento para reducir los falsos negativos durante episodios de convulsiones prolongados. Validado con datos clínicos del Hospital Universitario Regional de Málaga, el "framework" propuesto ilustra una vía escalable y práctica hacia sistemas de detección de convulsiones portátiles de última generación que son a la vez inteligentes y autónomos.

Keywords: Edge AI · TinyML · Neural Network Quantization · Seizure Detection

1 Introducción

Uno de los principales retos en el tratamiento de la epilepsia es la falta de herramientas fiables para la monitorización continua de las crisis en el entorno real. Aunque la electroencefalografía (EEG) es el método de referencia debido a su alta sensibilidad y especificidad, su uso fuera del ámbito clínico se ve limitado por su elevado coste, la necesidad de supervisión y la aceptabilidad social. Para abordar estas cuestiones, se están desarrollando dispositivos portátiles de detección de crisis que permitan una monitorización autónoma, faciliten una intervención oportuna y mejoren los resultados del tratamiento.

Los dispositivos portátiles se clasifican en cinco fases de validación que reflejan niveles crecientes de madurez clínica y complejidad [4]. La fase 0 utiliza datos simulados; la fase 1 introduce pacientes reales con registros anotados; la fase 2 requiere pruebas en más de 10 usuarios con garantías de seguridad; la fase 3 exige alertas en tiempo real validadas en entornos clínicos con más de 30 usuarios; y la fase 4 se centra en la usabilidad sin vídeo de referencia ni EEG.

Según esta escala, nuestro estudio corresponde a la fase 2, ya que se basa en datos de pacientes reales, anotados clínicamente y recopilados en un entorno hospitalario, e incluye una validación con múltiples usuarios (12 pacientes en total, con un análisis detallado de tres de ellos). Sin embargo, aún no implica el despliegue a gran escala y en tiempo real con más de 30 usuarios que caracteriza a la fase 3.

Según la revisión sistemática de [14], de los 16 dispositivos de detección de convulsiones disponibles en el mercado, solo dos, *Empatica* [12] y *EpiCare* [9], han obtenido la certificación de la Fase 4, mientras que otros tres, *Nightwatch* [1], *SPEAC* [6] y *EDDi* [3], permanecen en la Fase 3. Estos dispositivos cumplen los criterios para la generación de alertas en tiempo real y han sido validados en entornos clínicos o ambulatorios con más de 30 usuarios y eventos convulsivos verificados.

Los tipos de crisis epilépticas difieren en sus manifestaciones fisiológicas, lo que determina la selección de los sensores. Las crisis tónicas implican contracciones musculares sostenidas y suelen monitorizarse mediante electromiografía (EMG), actividad electrodérmica (EDA), temperatura cutánea (TEMP) y electrocardiogramas (ECG). Las crisis clónicas y mioclónicas, que implican espasmos musculares rítmicos o breves, se detectan mediante combinaciones similares, que a menudo incluyen el EEG. Las crisis atónicas, caracterizadas por una pérdida repentina del tono muscular, se monitorizan mediante EEG y ECG, mientras que los tipos complejos, como las crisis discognitivas focales o las crisis autonómicas, requieren una detección multimodal (p. ej., EEG, ECG, EDA, espectroscopia de infrarrojo cercano (NIRS)). Comprender estas relaciones entre los sensores y las crisis es clave para desarrollar sistemas de detección eficaces y aplicables en la vida real.

La mayoría de los dispositivos comerciales se centran en la acelerometría (ACM), a menudo con un uso limitado de modalidades complementarias. Por ejemplo, *Empatica* combina ACM, EDA, TEMP y fotopletismografía (PPG); *Nightwatch* utiliza ACM con PPG; y tanto *SPEAC* como *EDDi* se basan exclusivamente en EMG. Sin embargo, la sinergia entre EMG, EDA y TEMP sigue sin explorarse lo suficiente, a pesar de su potencial para detectar una gama más amplia de tipos de crisis.

Por otra parte, a pesar de los avances logrados, los dispositivos actuales suelen carecer de capacidad de inferencia autónoma integrada debido a las limitaciones computacionales de los microcontroladores. La mayoría de los sistemas delegan el procesamiento a teléfonos inteligentes o "gateways", lo que limita la capacidad de respuesta en tiempo real y la autonomía. Este cuello de botella se ve agravado por la necesidad de modelos de detección multimodal y de aprendizaje profundo, como las CNN y las redes neuronales recurrentes (RNN), que requieren una gran capacidad computacional. Para abordar esto se requieren estrategias de inferencia optimizadas para plataformas de bajo consumo y con memoria limitada. En este trabajo, proponemos y evaluamos dichas técnicas para permitir la detección de convulsiones en tiempo real y en el propio dispositivo en el borde.

En los estudios de fase 1, se requieren datos anotados que combinen señales de EEG y de sensores para entrenar los modelos de detección. Esto se suele hacer utilizando dispositivos como Shimmer [13] o Biosignals Plux [5], que carecen de funcionalidades de base de datos personalizadas para la anotación sincronizada y la generación de conjuntos de datos.

Se han propuesto varios "frameworks" de IoT para la monitorización y anotación de señales biológicas. Awais et al. [2] presentan una configuración de IoT inalámbrica optimizada para la clasificación de emociones mediante protocolos MAC de bajo consumo y hardware personalizado, pero que carece robustez para ser desplegado en entornos hospitalarios. Ham et al. [7] describen un sistema basado en Android construido sobre el middleware oneM2M, estrechamente acoplado a plataformas específicas sin modularidad ni soporte para contenedores. Murhe et al. [10] utilizan hardware basado en Arduino, pero omiten la integración en la nube y el soporte de datos históricos. Nada et al. [11] integran datos de EEG y de dispositivos wearables en un "framework" que cumple con el RGPD y se centra en la precisión de la detección, pero carece de ingestión modular o gestión de anotaciones. Estos sistemas suelen ser específicos para cada aplicación, no están contenedorizados y están vinculados a "stacks" fijos de hardware y software. Ninguno separa las anotaciones humanas de los datos de los sensores, admite el registro de datos históricos ni proporciona canalizaciones de ingestión modulares y flexibles.

Estas limitaciones ponen de manifiesto la necesidad de contar con sistemas de bajo coste, eficientes desde el punto de vista energético y autónomos que permitan una detección más amplia y el procesamiento en tiempo real en el borde. Este trabajo presenta un "framework" integral para la detección de convulsiones mediante dispositivos wearables que cubre estas carencias. Incluye: (1) la capacidad de detectar diversos tipos de convulsiones mediante señales multimodales

(EMG, EDA, TEMP); (2) modelos de aprendizaje profundo optimizados para la inferencia en tiempo real en microcontroladores; y (3) una plataforma modular y en contenedores para la adquisición y visualización de datos con anotación sincronizada y almacenamiento histórico. Esto contribuye al avance de los sistemas portátiles de próxima generación para la monitorización de convulsiones en el mundo real.

Para validar el "framework" propuesto, adoptamos una metodología sistemática que abarca el diseño de hardware, la adquisición de datos, la optimización de modelos y la implementación. Las siguientes secciones detallan este proceso de desarrollo. La sección 2 describe la arquitectura del sistema, el protocolo de recopilación de datos clínicos, la estrategia de diseño de modelos y los métodos de posprocesamiento. La sección 3 presenta los resultados experimentales sobre los datos de los pacientes, incluyendo pruebas de rendimiento de las redes neuronales optimizadas en dispositivos integrados. Por último, la sección 4 resume los hallazgos clave y el trabajo futuro para mejorar la integración clínica y las capacidades de análisis en tiempo real.

2 Metodología

En esta sección se presenta la metodología seguida para desarrollar el "framework" propuesto basado en el IoT y sus modelos de IA asociados para la detección de crisis epilépticas. Se comienza con una descripción de la arquitectura del sistema, que incluye un dispositivo portátil que captura señales fisiológicas multimodales y las transmite a una plataforma de visualización basada en la nube. A continuación, se detalla el proceso de adquisición de datos, incluida la colaboración con el Hospital Universitario Regional de Málaga para obtener registros de crisis sincronizados y anotados clínicamente. A continuación, se esboza la arquitectura para el procesamiento de datos y la visualización en tiempo real, seguida de la estrategia de aumento de datos que mejora la diversidad de las muestras de crisis mediante ventanas deslizantes temporales. Posteriormente, se describe el diseño y la optimización de los modelos de aprendizaje profundo, pasando de una red neuronal convolucional (CNN) con memoria a corto y largo plazo (LSTM) a una CNN ligera compatible con microcontroladores ESP32. Por último, se introduce una técnica de posprocesamiento para mejorar la robustez de la clasificación mediante la reducción de los falsos negativos durante episodios de crisis prolongados.

2.1 Arquitectura del sistema

La arquitectura general del sistema se basa en una solución distribuida basada en el IoT para la monitorización continua de las crisis epilépticas. La figura 1 muestra el flujo de datos desde la adquisición de los sensores hasta la inferencia en tiempo real.

Incluye sensores multimodales (EDA, EMG, TEMP y ACM) utilizados para extraer señales fisiológicas de los pacientes en una unidad de video-EEG. Estas

señales se transmiten mediante Bluetooth Low Energy (BLE) a una aplicación móvil personalizada, que utiliza Message Queuing Telemetry Transport (MQTT) para almacenar los datos en una base de datos temporal. A continuación, el personal médico puede anotar los datos utilizando el panel de control de Grafana con todas las señales. Al mismo tiempo, se sigue un enfoque de ventana deslizante para generar los conjuntos de datos, entrenando modelos de aprendizaje profundo con el conjunto de datos anotado. Finalmente, estos modelos se optimizan y se ejecutan en un dispositivo integrado con capacidades informáticas limitadas.

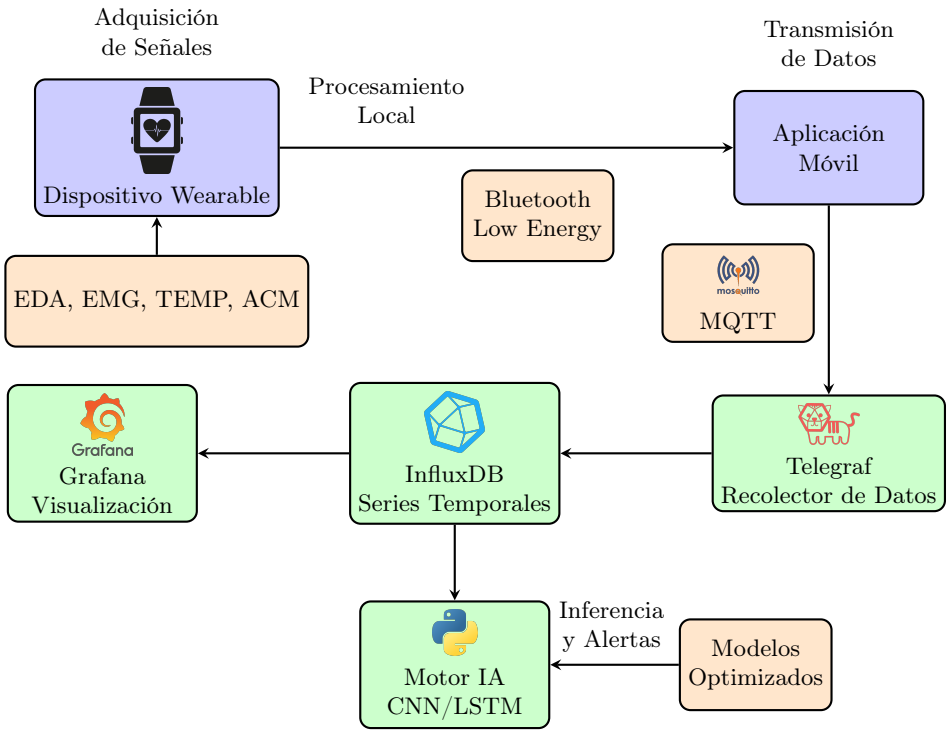

Fig. 1: Arquitectura inicial del Wearable con procesamiento IA centralizado.

2.2 Adquisición de datos

Dado que ningún conjunto de datos público incluye las señales fisiológicas multimodales necesarias, llevamos a cabo una fase de registro en el Hospital Universitario Regional de Málaga. Los pacientes fueron sometidos a una monitorización de vídeo-EEG de 24 horas, el método de referencia clínica para el registro de las crisis epilépticas. Además de los sistemas de EEG y ECG del hospital, integramos sensores externos (Biosignals Plux para EMG, EDA y TEMP; Polar

Verity Sense para ACM). Todas las secuencias se sincronizaron con las anotaciones clínicas de las crisis del hospital, lo que garantizó la alineación de los datos fisiológicos con los eventos de crisis verificados, tal y como se muestra en la figura 2.

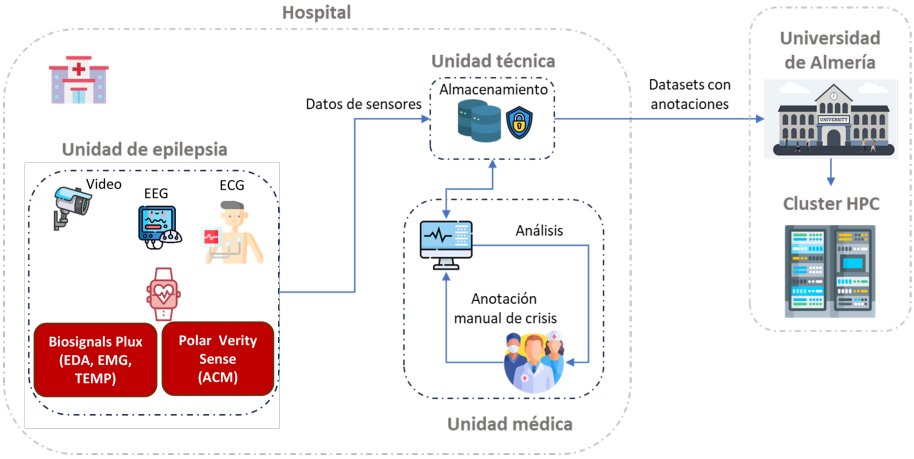

Fig. 2: Proceso de adquisición de datos fisiológicos en la unidad de Video-EEG.

2.3 Sistema IoT para la visualización de datos

La arquitectura del sistema de IoT que se muestra en la figura 1 admite un flujo de datos completo, desde la adquisición de los sensores hasta la visualización en tiempo real. Un dispositivo wearable recoge señales fisiológicas (EDA, EMG, TEMP, ACM) y las envía a través de BLE a una aplicación móvil, la cual reenvía los conjuntos de datos a un broker MQTT Mosquitto para una transmisión eficiente de alto rendimiento. Un agente Telegraf se suscribe al "topic" MQTT, almacena los datos en búfer y los escribe por lotes en una base de datos de series temporales InfluxDB, optimizando el rendimiento de la red y de la base de datos.

Grafana se conecta a InfluxDB para proporcionar paneles con vistas tanto superpuestas como individuales de las señales sin procesar, junto con métricas estadísticas (media, mínimo/máximo, desviación estándar) para cada señal. Las anotaciones clínicas se guardan junto con los datos del electroencefalograma en archivos de formato europeo de datos (EDF), se analizan mediante un script de Python y se cargan a través de MQTT en una base de datos MariaDB. Otro script sincroniza estas anotaciones con Grafana a través de su interfaz de programación de aplicaciones (API), integrando las crisis clínicas en los paneles de control.

La figura 3 muestra un ejemplo de la visualización de múltiples crisis del Paciente 8, donde la sección izquierda muestra tres señales fisiológicas sin proce-

sar superpuestas, y la sección derecha las muestra individualmente con el fin de realizar estudios específicos. Se muestran cinco crisis en áreas rectangulares verticales de color azul.

Fig. 3: Visualización de las crisis del Paciente 8 a través de Grafana

2.4 Aumento de datos

Para hacer frente al desequilibrio inherente entre clases en los conjuntos de datos de detección de crisis epilépticas, se implementó una estrategia de aumento de datos basada en ventanas deslizantes de una determinada longitud con desplazamientos temporales inferiores a dicha longitud. Esta técnica extrae múltiples segmentos superpuestos de la misma señal original, utilizando desplazamientos temporales. Solo se conservan para el entrenamiento los segmentos que contienen más del 50% de la duración de la crisis, a menos que formen parte de los datos sin desplazamiento, ya que necesitamos incluir los datos base sin crisis. Este enfoque aumenta significativamente la diversidad y la frecuencia de las muestras de crisis sin requerir una recopilación de datos adicional, mejorando así la robustez y la capacidad de generalización del modelo de clasificación.

La figura 4 ilustra este proceso de aumento de datos. La señal EDA original (arriba) contiene dos episodios convulsivos marcados en rojo mediante ventanas deslizantes de 30 segundos de duración. Mediante desplazamientos temporales de 7,5 s y 15 s, se generan múltiples segmentos superpuestos a partir de los mismos datos de origen. Solo las ventanas de los desplazamientos temporales que contienen más del 50% de contenido de crisis se conservan como muestras de entrenamiento válidas (mostradas en ventanas de color verde), mientras que las ventanas con contenido insuficiente de crisis se descartan (mostradas en gris).

Los datos sin desplazar, tanto de los eventos sin crisis como de las crisis, no se descartan.

En este contexto, puede surgir un posible problema de fuga de datos, ya que los desplazamientos temporales en ventanas consecutivas podrían capturar involuntariamente partes del conjunto de validación. Para mitigar esto, el punto de división entre los datos de entrenamiento y los de validación se eligió cuidadosamente para garantizar que ningún evento convulsivo del conjunto de validación apareciera en ninguna de las ventanas desplazadas utilizadas para el entrenamiento, lo que da como resultado que no se añadan puntos de datos aumentados del conjunto de validación.

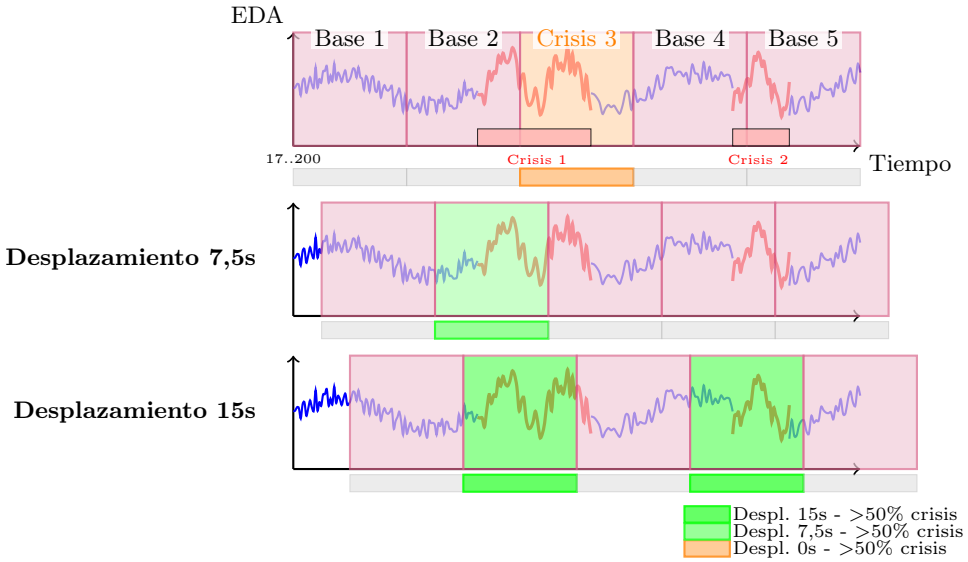

Fig. 4: Estrategia de aumento de datos mediante ventanas deslizantes con desplazamientos temporales.

2.5 Arquitectura y optimización del modelo IA

El desarrollo de un sistema de inferencia en el borde requirió importantes modificaciones arquitectónicas para adaptarse a las limitaciones de recursos de los microcontroladores con recursos limitados. Inicialmente, se diseñó una arquitectura híbrida CNN-LSTM por sus capacidades de aprendizaje temporal. Sin embargo, debido a las limitaciones de hardware y a la incompatibilidad de las capas LSTM con el "framework" TensorFlow Lite en los microcontroladores, la arquitectura se rediseñó como una CNN más sencilla con tres capas Conv1D.

La tabla 1 muestra una comparación de ambas arquitecturas, en la que el modelo CNN+LSTM procesa 7 características de entrada incorporando 4 atrib-

utos temporales (día, hora, minuto, día de la semana) junto con las 3 señales fisiológicas (EDA, EMG, TEMP), mientras que el modelo CNN simplificado conserva únicamente las señales fisiológicas.

A pesar de la simplificación, la arquitectura CNN optimizada tiene un mayor número de parámetros (31 457) que el híbrido CNN+LSTM original (21 282), debido al aumento de la profundidad y el tamaño de las capas convolucionales y densas.

Se eliminaron las capas de normalización por lotes, ya que los datos están preescalados, lo que reduce la sobrecarga computacional y de memoria. Entre las optimizaciones críticas se incluyen la cuantización de `FLOAT32` a `INT8` y la compatibilidad con la biblioteca ESP-NN, lo que permite un ahorro de memoria de 4× y una inferencia más rápida. Se consideró la "poda", pero a menudo da lugar a una dispersión irregular con una aceleración mínima en los microcontroladores, mientras que la "destilación del conocimiento" sigue siendo una estrategia prometedora para el futuro que permite equilibrar aún más la precisión y la eficiencia [8].

Capa	**Forma de Salida**	**Parámetros**
InputLayer	(10, 7)	0
Conv1D	(10, 32)	1,152
MaxPooling1D	(5, 32)	0
BatchNormalization	(5, 32)	128
Dropout	(5, 32)	0
Conv1D	(5, 64)	6,208
MaxPooling1D	(2, 64)	0
BatchNormalization	(2, 64)	256
Dropout	(2, 64)	0
LSTM	(32)	12,416
Dropout	(32)	0
Dense	(32)	1,056
Dropout	(32)	0
Dense	(2)	66
Parámetros Totales		**21,282**

(a) Arquitectura CNN+LSTM para clasificación de señales multivariantes.

Capa	**Forma de Salida**	**Parámetros**
InputLayer	(10, 3)	0
Conv1D	(10, 32)	416
MaxPooling1D	(5, 32)	0
Conv1D	(5, 64)	6,208
MaxPooling1D	(2, 64)	0
Conv1D	(2, 128)	16,512
MaxPooling1D	(1, 128)	0
Flatten	(128)	0
Dense	(64)	8,256
Dense	(1)	65
Parámetros Totales		**31,457**

(b) Arquitectura CNN optimizada para despliegue en ESP32 sin capas LSTM.

Table 1: Comparación de arquitecturas de Deep Learning: CNN+LSTM completa vs. CNN ligera para sistemas embebidos.

2.6 Posprocesamiento

Para reducir los falsos negativos, se aplicó una estrategia de posprocesamiento temporal. Como se muestra en la figura 5, las predicciones se agrupan en intervalos de 30 segundos. Si más del 50% de las predicciones de un intervalo indican una crisis, todo el intervalo se reclasifica como crisis. Esta estrategia mitiga los falsos negativos aislados debidos a predicciones inexactas en períodos de crisis naturalmente continuos y mejora la "sensibilidad" global.

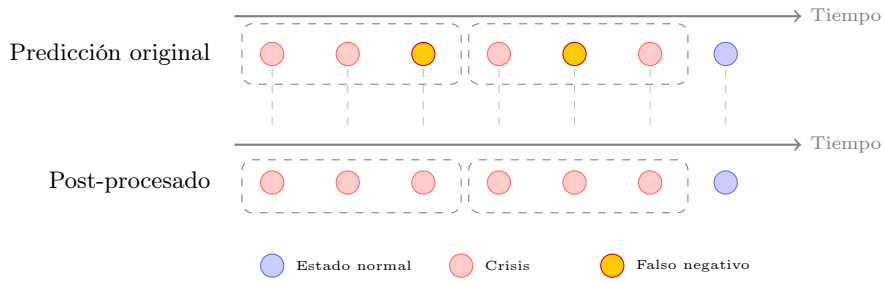

Fig. 5: Corrección de falsos negativos mediante post-procesamiento con ventana temporal.

3 Resultados

En esta sección se presentan los resultados obtenidos tras aplicar el "framework" propuesto a los datos fisiológicos recopilados en el Hospital Universitario Regional de Málaga. Se comienza detallando las características del conjunto de datos, incluidos los retos que plantea la escasez de datos validados de pacientes, y se explica la estrategia de división temporal de los datos utilizada para garantizar un entrenamiento y una evaluación del modelo sin sesgos. A continuación, destacamos el buen rendimiento del modelo de aprendizaje profundo CNN+LSTM tanto en contextos generales como específicos de cada paciente. Sin embargo, debido a las limitaciones de hardware, solo se pudo implementar el modelo CNN en el objetivo integrado final, el microcontrolador Arduino ESP32. Por lo tanto, también presentamos los resultados de la cuantización y validación de este modelo CNN optimizado, demostrando su eficacia para la inferencia en tiempo real con recursos limitados.

3.1 Validación del modelo y dataset

El "framework" y los modelos de IoT se validaron utilizando un subconjunto de datos de 12 pacientes, que fueron monitorizados en el Hospital Universitario Regional de Málaga. El conjunto de datos de cada paciente incluye varias crisis

epilépticas anotadas y períodos normales prolongados. Todas las señales fisiológicas (EDA, EMG, TEMP, ACM) se sincronizaron con los datos de EEG y se organizaron en carpetas específicas para cada paciente dentro del flujo de datos en contenedores.

La validación se centró en los pacientes 5, 8 y 9, ya que eran los únicos casos con datos EDA+EMG+TEMP médicamente validados y totalmente sincronizados. Se implementó una estrategia de separación temporal para garantizar que los datos de entrenamiento y de prueba permanecieran separados cronológicamente, preservando la secuencia natural de los eventos y evitando la fuga de datos, algo esencial para las aplicaciones sanitarias en tiempo real.

La validación del modelo siguió dos estrategias: un modelo general y modelos específicos para cada paciente. El paciente P5, con una sola crisis epiléptica anotada, contribuyó exclusivamente al entrenamiento del modelo general. Por el contrario, los pacientes P8 y P9, que presentaban múltiples episodios de crisis, se utilizaron para desarrollar y probar modelos individuales específicos para cada paciente, validando la escalabilidad y la adaptabilidad del sistema a perfiles de datos de pacientes variables.

3.2 Resultados de Deep Learning

La arquitectura CNN+LSTM muestra un rendimiento sólido y consistente con la combinación de sensores EDA+EMG+TEMP en el modelo general, tal y como se muestra en la tabla 3. Esta combinación alcanza una precisión de 0,9800 para la clase sin convulsiones (C0) y de 1,0000 para la clase con convulsiones (C1), con valores de "sensibilidad" de 1,0000 y 0,7500 respectivamente, lo que demuestra una capacidad de detección equilibrada.

Los modelos específicos para cada paciente resaltan aún más la variabilidad en el rendimiento de la detección con diferentes combinaciones de sensores. El modelo del paciente 8, que utiliza solo TEMP, alcanza una alta precisión y una "sensibilidad" similares a los del modelo general, mientras que el modelo del paciente 9, que utiliza EDA+EMG, muestra una excelente precisión para la detección de eventos no convulsivos, pero tiene dificultades con la precisión en la detección de convulsiones, lo que indica posibles falsos positivos. Estos resultados enfatizan la importancia de la calibración personalizada y la selección de sensores para mejorar la fiabilidad de la detección.

Modelo	**Sensores**	**Precisión C0**	**Precisión C1**	**Recall C0**	**Recall C1**
CNN+LSTM (General)	EDA+EMG+TEMP	0.9800	1.0000	1.0000	0.7500
CNN+LSTM (Paciente 8)	TEMP	0.9444	1.0000	1.0000	0.7500
CNN+LSTM (Paciente 9)	EDA+EMG	1.0000	0.0281	0.6581	1.0000

Table 2: Mejores resultados del modelo general y los modelos específicos por paciente para los modelos de Deep Learning.

Modelo	Sensores	Precisión C0	Precisión C1	Recall C0	Recall C1
CNN+LSTM (General)	EDA+EMG+TEMP	0.9800	1.0000	1.0000	0.7500
CNN (Paciente 8)	TEMP	0.9255	1.0000	1.0000	0.7500
CNN+LSTM (Paciente 9)	EDA+EMG	1.0000	0.0281	0.6581	1.0000

Table 3: Mejores resultados del modelo general y los modelos específicos por paciente para los modelos de Deep Learning.

0.0833 0.8878 0.1250

3.3 Despliegue de la CNN embebida

La CNN optimizada se implementó con éxito en microcontroladores ESP32, lo que puso de manifiesto la capacidad del "framework" para el procesamiento autónomo en el borde. A modo de demostración, el modelo se entrenó con el paciente 8, el sujeto con el mayor número de crisis epilépticas, y los modelos de CNN con mejor rendimiento.

La cuantización redujo el tamaño del modelo de 284,3 KB a 45,5 KB, lo que permitió su implementación en dispositivos ESP32 con recursos limitados sin perder precisión. Se codificaron de forma fija seis filas del conjunto de datos del Paciente 8 en el código de inferencia del ESP32 para realizar pruebas comparativas, lo que confirmó la coherencia con la implementación en Python; sin embargo, el modelo implementado puede procesar señales arbitrarias en tiempo real.

La validación que se muestra en la tabla 4 reveló desviaciones inferiores al 0,78% entre los resultados de ESP32 y Python, con tiempos de inferencia inferiores a 15 ms. Estos resultados confirman tanto la fiabilidad del modelo cuantizado como la eficiencia de las estrategias de optimización dentro del "framework" del IoT.

Grupo	ESP32	Python	Desviación	Valor Real	Predicción
1	0.459	0.460	0.0039	Normal	Normal
2	0.461	0.461	0.0034	Normal	Normal
3	0.459	0.459	0.0025	Normal	Normal
4	0.567	0.574	0.0059	Crisis	Crisis
5	0.573	0.578	0.0078	Crisis	Crisis
6	0.556	0.556	0.0000	Crisis	Crisis
Total	–	–	<0.78%	**Exactitud: 100%**	

Table 4: Validación del despliegue: consistencia entre la implementación en ESP32 y Python.

4 Conclusiones y trabajos futuros

Este trabajo presenta un "framework" modular de IoT e IA para la monitorización de la salud mediante dispositivos wearables en casos de crisis epilépticas, que abarca todo el proceso, desde la adquisición de datos de los sensores hasta la inferencia en el borde. La arquitectura en contenedores integra MQTT, InfluxDB y Grafana para la monitorización en tiempo real y del historial, tendiendo un puente entre el manejo de datos clínicos y la implementación de IA integrada.

La validación con datos multimodales sincronizados de pacientes confirma la gestión fiable de conjuntos de datos sensibles con integridad temporal. Las redes neuronales optimizadas en microcontroladores ESP32 lograron tiempos de inferencia inferiores a 15 ms y una pérdida mínima de precisión tras la cuantización, lo que demuestra la viabilidad de la monitorización en el borde en tiempo real y de bajo consumo. La modularidad, la escalabilidad y la compatibilidad con los estándares de EEG del "framework" ponen de relieve su potencial para la asistencia sanitaria inteligente.

El estudio se ve limitado por el reducido tamaño del dataset, compuestao por 12 pacientes epilépticos, debido a las restricciones en la recopilación de datos sobre las crisis, aunque los resultados siguen siendo coherentes con las validaciones de las fases iniciales. Los trabajos futuros incluyen la incoporación de nuevos pacientes para lograr una mayor solidez estadística y avanzar hacia la validación clínica de las fases 3 y 4.

El desarrollo posterior mejorará los análisis con visualizaciones avanzadas de Grafana, ampliará las modalidades de los sensores y perfeccionará la integración clínica mediante interfaces intuitivas para el registro de crisis y la sincronización en tiempo real. Estas medidas tienen como objetivo facilitar una monitorización escalable, autónoma y de múltiples pacientes con IA de vanguardia.

Por último, se explorará la "destilación del conocimiento" junto con la "cuantización" para diseñar modelos compactos que mantengan el rendimiento predictivo al tiempo que reducen las exigencias computacionales en dispositivos con recursos limitados.

Acknowledgments. Este trabajo ha sido financiado por los proyectos de I+D+i PID2021-123278OB-I00 y PDC2022-133370-I00, con el número de referencia MCIN/AEI/10.13039/501100011033/, y con fondos del FEDER; así como por el Departamento de Informática de la Universidad de Almería.

References

1. Arends, J., Thijs, R.D., Gutter, T., Ungureanu, C., Cluitmans, P., van Dijk, J., van Andel, J., Tan, F., de Weerd, A., Vledder, B., Hofstra, W., Lazeron, R., van Thiel, G., Roes, K.C.B., Leijten, F., Consortium, D.T.E.: Multimodal nocturnal seizure detection in a residential care setting: A long-term prospective trial. Neurology **91**(21), e2010–e2019 (2018). https://doi.org/10.1212/WNL.0000000000006545
2. Awais, M., Raza, M., Singh, N., Bashir, K., Manzoor, U., Ul Islam, S., Rodrigues, J.J.P.C.: Lstm-based emotion detection using physiological signals: Iot framework

for healthcare and distance learning in covid-19. IEEE Internet of Things Journal **8**(23), 16863–16871 (2021). https://doi.org/10.1109/JIOT.2020.3044031
3. Beniczky, S., Conradsen, I., Henning, O., Fabricius, M., Wolf, P.: Automated real-time detection of tonic-clonic seizures using a wearable emg device. Neurology **90**(5), e428–e434 (2018). https://doi.org/10.1212/WNL.0000000000004893
4. Beniczky, S., Ryvlin, P.: Standards for testing and clinical validation of seizure detection devices. Epilepsia **59**(suppl 1), 9–13 (2018). https://doi.org/10.1111/epi.14049
5. Biosignals, P.W.: biosignalsplux: Wireless biometric signal acquisition (2025), https://www.pluxbiosignals.com/
6. Halford, J.J., Sperling, M.R., Nair, D.R., Dlugos, D.J., Tatum, W.O., Harvey, J., French, J.A., Pollard, J.R., Faught, E., Noe, K.H., Henry, T.R., Jetter, G.M., Lie, O.V., Morgan, L.C., Girouard, M.R., Cardenas, D.P., Whitmire, L.E., Cavazos, J.E.: Detection of generalized tonic-clonic seizures using surface electromyographic monitoring. Epilepsia **58**(11), 1861–1869 (2017). https://doi.org/10.1111/epi.13897
7. Ham, S.M., Lee, H.M., Lim, J.H., Seo, J.: A negative emotion recognition system with internet of things-based multimodal biosignal data. Electronics **12**(20), 4321 (2023). https://doi.org/10.3390/electronics12204321
8. Lê, M.T., Wolinski, P., Arbel, J.: Efficient neural networks for tiny machine learning: A comprehensive review. ArXiv **abs/2311.11883** (2023), https://api.semanticscholar.org/CorpusID:265295114
9. Meritam, P., Ryvlin, P., Beniczky, S.: User-based evaluation of applicability and usability of a wearable accelerometer device for detecting bilateral tonic-clonic seizures: a field study. Epilepsia **59**(suppl 1), 48–52 (2018). https://doi.org/10.1111/epi.14051
10. Murhe, V., Nagpure, S., Bhanudas, D.A.: Iot-convnet + lamb: A deep learning based emotion recognition framework using smart iot systems. International Journal of Information Technology **17**(5), 2871–2876 (2025). https://doi.org/10.1007/s41870-025-02482-4
11. Rashad, Z., et al.: An iot framework for emotion detection and behavior analysis with gdpr/hipaa compliance. Americas PG **3**(3), 144–163 (2025). https://doi.org/10.54216/FPA.190113
12. Regalia, G., Onorati, F., Lai, M., Caborni, C., Picard, R.W.: Multimodal wrist-worn devices for seizure detection and advancing research: Focus on the empatica wristbands. Epilepsy Res **153**, 79–82 (2019). https://doi.org/10.1016/j.eplepsyres.2019.02.007
13. Sensing, S.: Shimmer wearable wireless sensors (2025), https://www.shimmersensing.com/
14. Verdru, J., Van Paesschen, W.: Wearable seizure detection devices in refractory epilepsy. Acta Neurol Belg **120**, 1271–1281 (2020). https://doi.org/10.1007/s13760-020-01417-z