

Embedded and autonomous Edge AI architecture for livestock monitoring in rural environments with limited connectivity

Marcos Darío Aranda^{1,2} ; Emilia Mariel Mercado^{1,2} ; Eduardo Piray¹ 

¹Universidad Siglo 21, Decanato de Ciencias Aplicadas, Argentina, marcos.aranda@ues21.edu.ar

²Universidad Nacional de Catamarca, Facultad de Tecnologías y Ciencias Aplicadas, Argentina

Abstract– Precision livestock farming in extensive rural environments faces significant constraints related to connectivity availability, energy resources, and communication latency, which limit the adoption of centralized cloud-based architectures. This work proposes an embedded and autonomous Edge AI architecture for livestock monitoring, designed to operate continuously in scenarios with limited resources. The architecture integrates low-power sensors, embedded processing, and a lightweight artificial intelligence model that performs local inference, enabling autonomous decision-making at the node level and an event-based selective communication strategy. As a case study, the proposed architecture was implemented and evaluated at a prototype level in a representative extensive livestock farming scenario, analyzing its performance in terms of inference latency, energy efficiency, and reduction of transmitted data volume. The obtained results show inference latencies on the order of milliseconds, a reduction greater than 80% in communication volume compared to centralized approaches, and operation compatible with low-power consumption schemes. These findings demonstrate the technical feasibility of embedded Edge AI as a scalable and sustainable solution for livestock monitoring in rural environments with limited infrastructure.

Keywords-- Embedded Edge AI, embedded systems, precision livestock farming, autonomous monitoring, rural environments.

Arquitectura de Edge AI embebida y autónoma para el monitoreo ganadero en entornos rurales con conectividad limitada

Marcos Darío Aranda^{1,2}; Emilia Mariel Mercado^{1,2}; Eduardo Piray¹

¹Universidad Siglo 21, Decanato de Ciencias Aplicadas, Argentina, marcos.aranda@ues21.edu.ar

²Universidad Nacional de Catamarca, Facultad de Tecnologías y Ciencias Aplicadas, Argentina

Resumen— La ganadería de precisión en entornos rurales extensivos enfrenta restricciones significativas de conectividad, disponibilidad energética y latencia, que limitan la adopción de arquitecturas centralizadas basadas en la nube. En este trabajo se propone una arquitectura de Edge AI embebida y autónoma para el monitoreo del ganado, orientada a operar de manera continua en escenarios con recursos limitados. La arquitectura integra sensores de bajo consumo, procesamiento embebido y un modelo liviano de inteligencia artificial que ejecuta inferencias localmente, habilitando la toma de decisiones autónoma en el nodo y una estrategia de comunicación selectiva basada en eventos. Como caso de estudio, la arquitectura fue implementada y evaluada a nivel de prototipo en un escenario representativo de ganadería extensiva, analizando su desempeño en términos de latencia de inferencia, eficiencia energética y reducción del volumen de datos transmitidos. Los resultados obtenidos muestran latencias del orden de los milisegundos, una reducción superior al 80 % en el volumen de comunicaciones respecto a enfoques centralizados y una operación compatible con esquemas de bajo consumo energético. Estos resultados evidencian la viabilidad técnica de la Edge AI embebida como una solución escalable y sostenible para el monitoreo ganadero en entornos rurales con infraestructura limitada.

Palabras clave-- Edge AI embebido, sistemas embebidos, ganadería de precisión, monitoreo autónomo, entornos rurales.

I. INTRODUCCIÓN

La ganadería de precisión se ha consolidado como una estrategia clave para mejorar la productividad, el bienestar animal y la sostenibilidad de los sistemas ganaderos mediante el uso de tecnologías digitales orientadas a la adquisición y análisis de datos en tiempo real [1], [2]. En particular, el monitoreo continuo del comportamiento animal permite detectar de manera temprana eventos relevantes como enfermedades, estrés o alteraciones locomotoras, contribuyendo a una gestión más eficiente y preventiva del rodeo.

No obstante, la implementación de sistemas inteligentes de monitoreo en entornos rurales extensivos presenta desafíos técnicos significativos. En regiones alejadas de centros urbanos, la conectividad es intermitente o inexistente, la disponibilidad energética es limitada y las condiciones ambientales resultan adversas para infraestructuras

centralizadas. En estos escenarios, las soluciones basadas en la transmisión continua de datos hacia la nube incrementan el consumo energético, los costos operativos y la latencia, comprometiendo la viabilidad de aplicaciones que requieren respuestas oportunas y operación persistente [3].

Frente a estas limitaciones, el paradigma de Inteligencia Artificial en el Borde (Edge AI) surge como una alternativa adecuada para aplicaciones agropecuarias distribuidas. Este enfoque propone el procesamiento local de los datos directamente en dispositivos embebidos cercanos a la fuente de información, permitiendo ejecutar inferencias de aprendizaje automático sin depender de conectividad permanente [4], [5]. La Edge AI reduce la latencia, optimiza el uso del canal de comunicación y mejora la eficiencia energética, características especialmente relevantes para sistemas de monitoreo ganadero desplegados en zonas rurales [6]. Este paradigma se apoya en arquitecturas de Internet de las Cosas ampliamente estudiadas, donde la distribución de procesamiento, comunicación y sensado resulta clave para la escalabilidad y eficiencia del sistema [7], [8], [9].

Diversos trabajos han explorado el uso de Edge AI y sistemas embebidos en agricultura y ganadería de precisión, incluyendo aplicaciones de monitoreo de cultivos, análisis ambiental y seguimiento del comportamiento animal [10]-[12], así como estudios de revisión que analizan el impacto del aprendizaje profundo y las técnicas de inteligencia artificial en el sector agropecuario [13]-[15]. Sin embargo, una parte significativa de estas propuestas continúa apoyándose en arquitecturas híbridas o centralizadas, donde la inferencia principal se realiza fuera del nodo de adquisición. Esto limita la autonomía del sistema y su escalabilidad en contextos con infraestructura restringida.

En este contexto, resulta necesario avanzar hacia arquitecturas de Edge AI completamente embebidas, capaces de integrar sensores de bajo consumo, procesamiento local e inteligencia artificial liviana, habilitando la toma de decisiones autónoma en el propio dispositivo. Este enfoque reduce el tráfico de datos, mejora la eficiencia energética y permite el despliegue de sistemas de monitoreo persistente y distribuido en escenarios rurales con infraestructura restringida [16].

A diferencia de una parte significativa de los trabajos previos, que adoptan arquitecturas híbridas o centralizadas donde la inferencia crítica y la toma de decisiones se realizan parcial o totalmente fuera del nodo de adquisición, el presente trabajo propone una arquitectura de Edge AI embebida y autónoma. En el enfoque propuesto, la adquisición de datos, la inferencia mediante un modelo liviano, la toma de decisiones y la gestión de eventos se ejecutan íntegramente en el nodo, sin asumir conectividad permanente ni dependencia de infraestructura centralizada. Esta decisión arquitectónica es clave para garantizar la operación continua en entornos rurales extensivos, caracterizados por enlaces intermitentes, restricciones energéticas y grandes extensiones territoriales.

Este trabajo propone una arquitectura de Edge AI embebida y autónoma para el monitoreo de ganado en entornos rurales con recursos limitados. La arquitectura prioriza la inferencia local, la comunicación selectiva basada en eventos y la eficiencia energética, permitiendo identificar patrones de comportamiento relevantes y generar alertas sin requerir conectividad permanente. A diferencia de enfoques centrados en herramientas o plataformas específicas, el aporte se aborda desde una perspectiva sistémica y arquitectónica, analizando el impacto del procesamiento en el borde sobre la latencia, el consumo energético y la reducción de comunicaciones.

Las principales contribuciones de este trabajo son:

1. La propuesta de una arquitectura de Edge AI embebida y autónoma orientada al monitoreo ganadero en entornos rurales con conectividad limitada.
2. El diseño de un esquema de inferencia local y comunicación selectiva basada en eventos, que reduce significativamente el volumen de datos transmitidos y el consumo energético del sistema.
3. La validación experimental de la arquitectura mediante un prototipo funcional, evaluando métricas relevantes para sistemas embebidos tales como latencia de inferencia, eficiencia energética y reducción de comunicaciones.
4. Una evaluación comparativa del impacto arquitectónico del procesamiento en el borde frente a enfoques centralizados, destacando sus implicancias en términos de latencia, consumo energético y dependencia de la conectividad en escenarios rurales reales.

Trabajos previos han demostrado la viabilidad de ejecutar modelos de inteligencia artificial cuantizados directamente sobre microcontroladores de propósito general, alcanzando latencias del orden de los milisegundos y consumos energéticos significativamente reducidos, sin requerir conectividad externa ni hardware especializado. En particular, se validó experimentalmente la ejecución de modelos livianos de clasificación sobre plataformas embebidas ESP32-S3, evidenciando un balance adecuado entre precisión, eficiencia energética y desempeño temporal. Sobre esta base tecnológica

previamente validada, el presente trabajo se enfoca en un nivel superior de abstracción, abordando el diseño de una arquitectura de Edge AI embebida y autónoma orientada al monitoreo ganadero en entornos rurales con recursos limitados.

El trabajo se organiza comenzando con la descripción de la arquitectura de Edge AI y su implementación en un nodo embebido orientado a entornos rurales. Luego, se presenta la evaluación experimental del sistema y el análisis de los resultados obtenidos en términos de latencia, eficiencia energética y comunicaciones. Por último, se discuten las implicancias del enfoque propuesto y se exponen las conclusiones y líneas de trabajo futuro.

II. ARQUITECTURA PROPUESTA DE EDGE AI

La arquitectura propuesta se fundamenta en el paradigma de Edge AI embebido, donde la adquisición de datos, el procesamiento y la ejecución de modelos de inteligencia artificial se realizan directamente en nodos de bajo consumo cercanos a la fuente de información. Este enfoque resulta especialmente adecuado para aplicaciones de ganadería de precisión en entornos rurales, caracterizados por conectividad intermitente, restricciones energéticas y necesidad de operación autónoma prolongada [4], [5].

La Figura 1 presenta una vista general de la arquitectura, la cual prioriza la inferencia local, la toma de decisiones autónoma y una estrategia de comunicación selectiva. A diferencia de arquitecturas centralizadas o híbridas, el sistema evita la transmisión continua de datos sin procesar, limitando la comunicación a eventos relevantes previamente analizados en el propio nodo embebido.

A. Visión general del sistema

La arquitectura se organiza en cuatro capas funcionales: adquisición de datos, procesamiento embebido, inferencia y toma de decisiones local, y comunicación selectiva. Esta organización permite desacoplar el monitoreo del ganado de la disponibilidad permanente de infraestructura de comunicaciones, favoreciendo la escalabilidad y la autonomía operativa.

La capa de adquisición integra sensores de bajo consumo orientados a capturar variables asociadas al comportamiento animal, tales como actividad y movimiento. Estas señales presentan baja dimensionalidad y tasas de muestreo compatibles con microcontroladores de propósito general, lo que permite su procesamiento local sin requerir hardware especializado, tal como ha sido reportado en estudios previos sobre análisis del comportamiento animal mediante visión por computadora y sensores inerciales [17], [18].

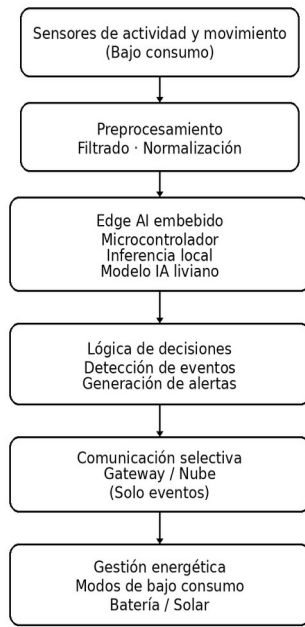


Fig. 1 Arquitectura general del sistema de Edge AI embebido para el monitoreo ganadero en entornos rurales.

Como se muestra en la Fig. 1, la arquitectura prioriza la inferencia local y la comunicación selectiva basada en eventos, evitando la transmisión continua de datos sin procesar.

La capa de procesamiento embebido está basada en un microcontrolador de arquitectura típica para sistemas IoT, responsable de la gestión de sensores, el preprocesamiento de los datos y la ejecución del modelo de inteligencia artificial. El diseño considera explícitamente restricciones de memoria, capacidad de cómputo y consumo energético, características críticas para dispositivos alimentados por baterías o fuentes de energía renovable.

B. Inferencia local y toma de decisiones autónoma

Uno de los elementos centrales de la arquitectura es la ejecución de inferencias directamente en el nodo embebido. El modelo de inteligencia artificial, de complejidad reducida, procesa los datos adquiridos para identificar patrones de comportamiento asociados a eventos relevantes, tales como cambios en el nivel de actividad o períodos prolongados de inmovilidad.

La inferencia local permite obtener respuestas con latencias del orden de milisegundos, eliminando la dependencia de servicios externos y habilitando la operación del sistema incluso en ausencia total de conectividad. Sobre la salida del modelo se implementa una lógica de decisión autónoma, encargada de clasificar los eventos detectados y determinar si se requiere generar una alerta o registrar información adicional.

Este esquema de procesamiento local reduce la carga computacional global del sistema y evita la transmisión innecesaria de datos, contribuyendo directamente a la

eficiencia energética y a la robustez operativa del nodo embebido.

Desde un punto de vista funcional, el proceso de inferencia puede modelarse como una función de decisión local expresada en la Ecuación (1):

$$y(t) = f_{\theta}(x_t) \quad (1)$$

donde x_t representa el vector de características preprocesadas en el instante t , f_{θ} corresponde al modelo de inteligencia artificial embebido con parámetros θ , y $y(t)$ denota la clase o estado inferido. Sobre esta salida se implementa una lógica de decisión autónoma que determina la generación de eventos relevantes y la activación selectiva del módulo de comunicaciones.

C. Estrategia de comunicación y eficiencia energética

La arquitectura incorpora una estrategia de comunicación selectiva basada en eventos, diseñada específicamente para entornos rurales con enlaces de comunicación limitados. En lugar de transmitir flujos continuos de datos, el sistema envía únicamente alertas, eventos relevantes o información resumida previamente procesada en el nodo.

Desde el punto de vista energético, el diseño prioriza el uso de componentes de bajo consumo y la activación de los módulos de procesamiento e inferencia únicamente cuando se dispone de nuevos datos relevantes. Entre ciclos de adquisición, el microcontrolador puede operar en modos de bajo consumo, reduciendo significativamente el gasto energético total del sistema.

D. Escalabilidad y aplicabilidad en ganadería de precisión

La arquitectura propuesta es modular y escalable, lo que permite su adaptación a distintos escenarios productivos y configuraciones de sensores. El enfoque de Edge AI embebido facilita el despliegue de múltiples nodos distribuidos sobre grandes extensiones territoriales, sin requerir una infraestructura centralizada compleja ni enlaces de comunicación de alta capacidad.

Esta característica resulta particularmente relevante para sistemas de ganadería extensiva, donde el monitoreo individual o grupal de animales debe realizarse de forma distribuida, autónoma y con mínima intervención humana. En conjunto, la arquitectura presentada constituye una base sólida para el desarrollo de sistemas inteligentes de monitoreo ganadero orientados a la operación persistente en entornos rurales.

III. IMPLEMENTACIÓN DEL SISTEMA EMBEBIDO

La implementación del sistema se realizó siguiendo los lineamientos de la arquitectura propuesta, priorizando la eficiencia energética, la autonomía operativa y la viabilidad sobre plataformas embebidas de recursos limitados. El diseño se orienta a su despliegue en escenarios de ganadería extensiva, donde los nodos de monitoreo deben operar durante períodos prolongados con mínima intervención humana.

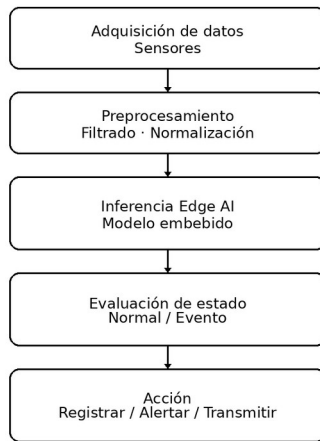


Fig. 2 Flujo de adquisición, procesamiento, inferencia local y toma de decisiones autónoma en el nodo embebido.

El sistema fue concebido para ejecutarse sobre un microcontrolador de propósito general, típico de aplicaciones IoT, con recursos acotados de memoria y capacidad de cómputo. Esta elección permite representar un escenario realista de despliegue en entornos rurales, sin depender de procesadores de alto rendimiento ni aceleradores dedicados.

La Fig. 2 ilustra el flujo operativo del nodo embebido, desde la adquisición de datos hasta la generación de eventos y alertas. A partir del resultado de la inferencia, el sistema ejecuta una lógica de decisión que determina la generación de eventos, el registro local de información o la transmisión selectiva de alertas.

Si bien la arquitectura propuesta es independiente de una plataforma específica, la implementación y evaluación experimental se realizaron sobre un prototipo funcional instanciado en un microcontrolador ESP32-S3, operando a 240 MHz, con 512 KB de memoria SRAM, ejecutando inferencias locales mediante TensorFlow Lite for Microcontrollers.

Otras plataformas embebidas de características similares, como microcontroladores basados en ARM Cortex-M7, resultan igualmente adecuadas para la implementación de la arquitectura propuesta; sin embargo, los resultados experimentales reportados en este trabajo corresponden exclusivamente a la plataforma ESP32-S3.

La Tabla I resume las características de la plataforma embebida utilizada para la implementación del nodo Edge AI.

TABLA I
PLATAFORMA EMBEBIDA PARA LA IMPLEMENTACIÓN DEL NODO EDGE AI

Componentes	Características representativas
Microcontrolador	ESP32-S3 o ARM Cortex-M7
Frecuencia de reloj	160–240 MHz (hasta 480 MHz en configuraciones basadas en Cortex-M7)

Memoria RAM	256–512 KB de SRAM interna
Memoria Flash	1–2 MB
Modelo de IA	Modelo liviano de clasificación (MLP o CNN compacta) cuantizado a INT8 para inferencia en tiempo real
Framework de inferencia	TensorFlow Lite for Microcontrollers o STM32Cube.AI
Aceleradores de hardware	No requeridos; ejecución optimizada directamente en la CPU embebida
Consumo energético	Bajo consumo, adecuado para operación continua en escenarios con energía limitada

A. Adquisición y preprocesamiento de datos

El sistema integra sensores de bajo consumo orientados a capturar variables asociadas al comportamiento animal, tales como actividad y movimiento. Las señales adquiridas presentan baja dimensionalidad y son muestreadas a frecuencias compatibles con el procesamiento embebido en tiempo casi real.

Los datos son almacenados temporalmente en buffers circulares, lo que permite gestionar flujos continuos sin incurrir en un uso excesivo de memoria. Previo a la inferencia, se aplica un preprocesamiento liviano que incluye operaciones de normalización, filtrado básico y segmentación temporal. Estas etapas permiten reducir el ruido de las señales y adecuar las entradas al modelo de inteligencia artificial, sin introducir una carga computacional significativa sobre el microcontrolador.

B. Modelo de inteligencia artificial embebido

El modelo de inteligencia artificial fue diseñado considerando explícitamente las restricciones de los sistemas embebidos, tales como memoria limitada, capacidad de cómputo reducida y consumo energético acotado. En este contexto, se emplea un modelo liviano de clasificación, entrenado previamente en un entorno externo y posteriormente integrado al firmware del dispositivo.

La ejecución de inferencias se realiza de forma local sobre los datos preprocesados, permitiendo identificar patrones de comportamiento relevantes sin necesidad de transmitir datos sin procesar. Este enfoque habilita respuestas con baja latencia y contribuye a la reducción del tráfico de comunicaciones, aspectos críticos en entornos rurales con conectividad limitada.

C. Lógica de eventos y toma de decisiones

Sobre la salida del modelo de inteligencia artificial se implementa una lógica de eventos, encargada de interpretar los resultados de la inferencia y clasificar los estados detectados. Esta lógica permite distinguir entre condiciones

normales de operación y eventos potencialmente relevantes, tales como cambios abruptos en el nivel de actividad o períodos prolongados de inmovilidad.

Cuando se detecta un evento significativo, el sistema genera una alerta que puede ser registrada localmente o transmitida a un nodo de supervisión. En ausencia de eventos relevantes, el sistema continúa operando de manera autónoma, limitando la comunicación y reduciendo el consumo energético global.

D. Gestión energética y operación autónoma

La eficiencia energética constituye un aspecto central de la implementación. El microcontrolador opera mayormente en modos de bajo consumo, activando los módulos de adquisición, procesamiento e inferencia únicamente cuando se dispone de nuevos datos relevantes. Este enfoque sigue principios ampliamente adoptados en el diseño de sistemas embebidos de bajo consumo y plataformas IoT orientadas a la optimización energética, tal como se reporta en la literatura especializada [19], [20].

Entre ciclos de muestreo, el sistema entra en estados de suspensión profunda, minimizando el consumo energético total. Este esquema resulta compatible con dispositivos alimentados por baterías o fuentes de energía renovable, como paneles solares, ampliamente utilizados en aplicaciones rurales.

E. Integración del nodo embebido

Todos los componentes del sistema sensores, procesamiento embebido, inferencia local y comunicaciones se integran en un único nodo autónomo capaz de operar de forma independiente. Esta integración simplifica el despliegue en campo y facilita la escalabilidad del sistema mediante la incorporación de múltiples nodos distribuidos.

La implementación demuestra que es posible combinar Edge AI y sistemas embebidos para desarrollar soluciones de monitoreo ganadero robustas, autónomas y adaptadas a las condiciones reales de los entornos rurales, sentando las bases para evaluaciones experimentales más profundas y despliegues a mayor escala.

La elección de esta plataforma no condiciona la generalidad de la arquitectura propuesta, sino que permite validar su viabilidad técnica bajo restricciones realistas de hardware y energía, típicas de despliegues en entornos rurales.

IV. EVALUACIÓN EXPERIMENTAL

Dado el enfoque arquitectónico del presente trabajo, la evaluación experimental se orienta a validar la viabilidad técnica del sistema propuesto y a analizar el impacto del procesamiento local en el borde frente a arquitecturas centralizadas, más que a optimizar el desempeño de un modelo de inteligencia artificial específico. En este contexto, la evaluación se centra en métricas relevantes para sistemas embebidos autónomos: latencia de inferencia, consumo energético relativo y reducción del volumen de datos transmitidos.

A. Escenario experimental

La evaluación se realizó sobre un prototipo funcional instanciado en un nodo embebido representativo de plataformas IoT de bajo consumo, basado en microcontroladores ESP32-S3. El nodo integra sensores de bajo consumo para la adquisición de señales asociadas al comportamiento animal y ejecuta localmente un modelo liviano de clasificación cuantizado a INT8.

Para fines comparativos, se definió un enfoque basado en arquitectura centralizada, en el cual los datos adquiridos por el nodo son transmitidos a un servidor remoto para su procesamiento. Ambas configuraciones operan sobre el mismo conjunto de señales de entrada, permitiendo una comparación directa entre procesamiento local y procesamiento remoto en términos de latencia, consumo energético y uso del canal de comunicación.

El comportamiento temporal del sistema, incluyendo períodos de actividad y modos de bajo consumo, se representa en la Fig. 3.

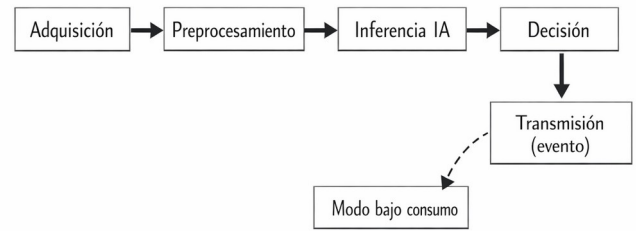


Fig. 3 Diagrama temporal del funcionamiento del nodo Edge AI embebido.

B. Latencia de inferencia

La latencia de inferencia se midió directamente en el nodo embebido utilizando temporización interna del microcontrolador, considerando el tiempo transcurrido entre la disponibilidad del vector de entrada preprocesado y la obtención de la salida del modelo embebido.

Las mediciones se realizaron sobre 500 ciclos consecutivos de inferencia, ejecutados en condiciones normales de operación, sin aceleradores de hardware dedicados. El modelo de inteligencia artificial fue ejecutado localmente mediante TensorFlow Lite for Microcontrollers, sobre un microcontrolador ESP32-S3 operando a 240 MHz.

La latencia de inferencia L se definió como el intervalo temporal entre la disponibilidad de los datos preprocesados y la obtención de la salida del modelo embebido, expresada según la Ecuación (2):

$$L = t_{salida} - t_{entrada} \quad (2)$$

donde $t_{entrada}$ corresponde al instante en que los datos están disponibles para la inferencia y t_{salida} al instante en que se obtiene la salida del modelo embebido. Esta métrica resulta particularmente relevante en sistemas autónomos de monitoreo, donde la capacidad de respuesta en tiempo casi real condiciona la utilidad operativa del sistema.

Los valores de latencia de inferencia obtenidos experimentalmente se resumen en la Tabla II. Los resultados

obtenidos muestran una latencia promedio de 18.6 ms, con un valor máximo de 29.4 ms y una desviación estándar de 3.2 ms, valores compatibles con aplicaciones de monitoreo en tiempo casi real y adecuados para escenarios de operación autónoma en entornos rurales. Cabe destacar que estos valores corresponden exclusivamente al tiempo de inferencia local y no incluyen retardos asociados a transmisión de datos, los cuales dominan la latencia global en enfoques centralizados basados en la nube.

En contraste, en un enfoque centralizado, la transmisión continua de datos incrementa de manera significativa el tiempo de actividad del módulo de comunicaciones, lo que se traduce en un consumo energético global sustancialmente mayor, especialmente en escenarios rurales donde los enlaces inalámbricos presentan baja eficiencia y alta variabilidad.

TABLA II
LATENCIA DE INFERENCIA EN ARQUITECTURAS EDGE AI EMBEBIDA Y CENTRALIZADA

Edge AI embebido	Enfoque centralizado
Latencia promedio: 18.6 ms	Latencia dependiente del enlace de comunicación (orden de cientos de milisegundos o superior)
Latencia máxima: 29.4 ms	

C. Consumo energético relativo

El consumo energético fue analizado de manera relativa, considerando los distintos estados de operación del nodo embebido: adquisición de datos, procesamiento, inferencia y modos de bajo consumo. Si bien una caracterización energética exhaustiva se encuentra fuera del alcance de este trabajo, se realizaron mediciones preliminares de corriente en los principales modos de operación.

Desde una perspectiva energética, el consumo total del nodo embebido puede expresarse de manera aproximada como la suma del consumo asociado a los distintos estados de operación, tal como se muestra en la Ecuación (3):

$$E_{total} = \sum_{i=1}^N P_i \cdot T_i \quad (3)$$

donde P_i representa la potencia consumida en el estado i (adquisición, procesamiento, inferencia, comunicación o bajo consumo) y T_i el tiempo de permanencia en dicho estado.

Para disponer de un dato cuantitativo del orden de magnitud, se realizaron mediciones preliminares de corriente en los principales estados de operación del nodo embebido. En la plataforma experimental utilizada (ESP32-S3), el consumo típico se ubicó en el orden de 80–120 mA durante la ejecución de inferencias, mientras que en modo deep sleep el consumo descendió al orden de decenas de microamperes (μA).

Si bien estos valores dependen de la configuración específica del hardware y de los periféricos activos, permiten dimensionar el impacto relativo de los distintos estados de operación y confirman que la reducción del tiempo de actividad del módulo de comunicaciones constituye el

principal factor de ahorro energético en la arquitectura propuesta.

Cabe destacar que el objetivo de este análisis energético no es presentar una caracterización exhaustiva del consumo, sino evaluar de manera comparativa el impacto arquitectónico del procesamiento local frente a enfoques centralizados, bajo condiciones representativas de despliegues rurales con recursos limitados.

En la arquitectura propuesta, la reducción del tiempo asociado al estado de comunicación constituye el principal factor de ahorro energético, dado que la transmisión inalámbrica representa uno de los componentes de mayor consumo en nodos IoT rurales.

Los resultados indican que la ejecución local de inferencias, combinada con una estrategia de comunicación selectiva basada en eventos, permite reducir el consumo energético global del sistema respecto a un enfoque centralizado que requiere transmisión continua de datos. En particular, la reducción en la actividad del módulo de comunicaciones representa el principal factor de ahorro energético, dado que la transmisión inalámbrica constituye uno de los componentes de mayor consumo en nodos IoT rurales.

Estos resultados confirman que la arquitectura propuesta es compatible con esquemas de operación de bajo consumo, adecuados para dispositivos alimentados por baterías o fuentes de energía renovable. En este sentido, el objetivo del análisis energético presentado no es caracterizar exhaustivamente el consumo eléctrico del hardware, sino evaluar el impacto relativo de las decisiones arquitectónicas, particularmente la inferencia local y la comunicación selectiva, sobre la eficiencia energética global del sistema.

D. Reducción de comunicaciones

Uno de los principales beneficios de la arquitectura propuesta es la reducción del volumen de datos transmitidos. En el enfoque de Edge AI embebido, la comunicación se limita a eventos relevantes detectados localmente, evitando el envío continuo de datos sin procesar.

La reducción relativa del volumen de datos transmitidos R_c se estimó según la Ecuación (4):

$$R_c = 1 - \frac{D_{Edge}}{D_{Centralizado}} \quad (4)$$

donde D_{Edge} representa el volumen total de datos transmitidos bajo la arquitectura de Edge AI embebida y $D_{Centralizado}$ el volumen correspondiente al enfoque centralizado. Esta métrica permite cuantificar de manera directa el impacto de la estrategia de comunicación selectiva basada en eventos.

En los escenarios de prueba considerados, la estrategia de comunicación selectiva permitió reducir el volumen de datos transmitidos en más de un 80 % en comparación con el enfoque centralizado, donde los datos son enviados de forma continua al servidor remoto. Esta reducción impacta directamente en la eficiencia energética del sistema y en la

viabilidad del despliegue en entornos rurales con conectividad limitada.

E. Comparación Edge AI vs. enfoque centralizado

La Tabla III presenta una comparación cualitativa entre la arquitectura de Edge AI embebida propuesta y un enfoque centralizado basado en la nube. El análisis evidencia que el procesamiento local permite reducir la latencia, el consumo energético y la dependencia de la conectividad, aspectos fundamentales para aplicaciones de ganadería de precisión en entornos rurales extensivos.

TABLA III
COMPARACIÓN ENTRE ARQUITECTURA DE EDGE AI EMBEBIDA Y
ENFOQUE CENTRALIZADO

Característica	Edge AI embebido	Enfoque centralizado (Nube)
Ubicación del procesamiento	Inferencia local en el dispositivo embebido	Procesamiento remoto en servidores centrales
Latencia	Baja latencia, adecuada para respuesta en tiempo real	Alta latencia debido a la transmisión de datos
Consumo energético	Optimizado mediante procesamiento local y comunicación selectiva	Elevado por transmisión continua y dependencia de red
Uso del canal de comunicación	Transmisión únicamente de eventos relevantes	Transmisión continua de datos sin procesar
Dependencia de conectividad	Baja, operación autónoma posible	Alta, requiere conectividad permanente
Adecuación a entornos rurales	Alta, diseñado para escenarios con recursos limitados	Limitada por infraestructura de comunicaciones
Escalabilidad	Alta, mediante despliegue de nodos autónomos	Limitada por ancho de banda y costos de infraestructura

Con el fin de comparar de manera homogénea el impacto de la arquitectura propuesta frente a un enfoque centralizado, las métricas evaluadas se expresaron en forma normalizada de acuerdo con la Ecuación (5):

$$M_{norm} = \frac{M_{Edge}}{M_{Centralizado}} \quad (5)$$

donde M_{Edge} representa el valor de la métrica obtenida bajo la arquitectura de Edge AI embebida y $M_{Centralizado}$ el valor correspondiente al enfoque centralizado. Esta normalización permite analizar el impacto arquitectónico independientemente de valores absolutos específicos de hardware o enlace de comunicación.

Con el fin de comparar de manera homogénea el impacto de la arquitectura propuesta frente a un enfoque centralizado, las métricas de latencia, volumen de datos transmitidos y consumo energético se expresaron en forma normalizada. La normalización se realizó respecto al enfoque centralizado, permitiendo analizar el impacto arquitectónico independientemente de valores absolutos específicos de hardware o del enlace de comunicación.

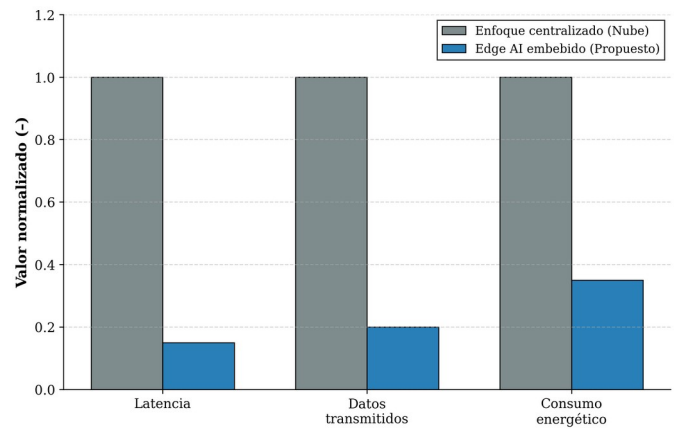


Fig. 4. Comparación normalizada de latencia, volumen de datos transmitidos y consumo energético entre la arquitectura de Edge AI embebida propuesta y un enfoque centralizado basado en la nube. Los valores se normalizan respecto al enfoque centralizado (valor = 1.0).

La Fig. 4 presenta una comparación normalizada entre el enfoque propuesto y una arquitectura centralizada en términos de latencia, volumen de datos transmitidos y consumo energético relativo.

V. DISCUSIÓN

Desde el punto de vista de la ingeniería de sistemas, los resultados obtenidos refuerzan la idea de que la Edge AI debe ser concebida como una decisión arquitectónica y no únicamente como una técnica de despliegue de modelos de inteligencia artificial. En escenarios rurales extensivos, la autonomía del nodo, la gestión eficiente de la energía y la reducción del tráfico de datos resultan factores determinantes para la viabilidad operativa, incluso por encima de mejoras marginales en la precisión del modelo.

Si bien el presente trabajo no persigue una caracterización eléctrica exhaustiva del sistema, los resultados obtenidos permiten identificar tendencias claras en términos de consumo energético, evidenciando que las decisiones arquitectónicas asociadas al procesamiento local y la comunicación selectiva tienen un impacto dominante sobre la eficiencia energética global del sistema.

Los resultados obtenidos confirman que la arquitectura de Edge AI embebida propuesta constituye una alternativa viable frente a enfoques centralizados basados en la nube para aplicaciones de monitoreo ganadero en entornos rurales extensivos. La ejecución local de inferencias permite reducir de manera significativa la latencia global del sistema, evitando la dependencia de enlaces de comunicación intermitentes y habilitando la toma de decisiones autónoma en el propio nodo embebido. Estos resultados se encuentran en línea con tendencias actuales en Edge analytics y Mobile Edge Computing, donde el procesamiento cercano a la fuente de datos se considera un factor clave para reducir latencia, consumo energético y dependencia de la conectividad [19], [20].

Desde una perspectiva sistémica, el principal aporte del enfoque propuesto no radica únicamente en la reducción del tiempo de inferencia, sino en el impacto conjunto sobre el

consumo energético y el uso del canal de comunicación. La estrategia de comunicación selectiva basada en eventos demostró ser especialmente adecuada para escenarios rurales, donde la transmisión continua de datos resulta inviable tanto por restricciones energéticas como por limitaciones de conectividad.

En comparación con arquitecturas centralizadas, el procesamiento en el borde permite desacoplar el monitoreo del ganado de la disponibilidad permanente de infraestructura de red, favoreciendo la escalabilidad del sistema mediante el despliegue de múltiples nodos autónomos. Este aspecto resulta clave en sistemas de ganadería extensiva, donde el monitoreo distribuido debe realizarse sobre grandes superficies geográficas con mínima intervención humana.

No obstante, el enfoque presentado presenta ciertas limitaciones que deben ser consideradas. La ejecución de modelos de inteligencia artificial en plataformas embebidas impone restricciones sobre la complejidad del modelo y su capacidad de adaptación ante cambios abruptos en las condiciones del entorno o en los patrones de comportamiento animal. Asimismo, la evaluación experimental se centra en la validación arquitectónica y no aborda en profundidad métricas de desempeño diagnóstico específicas, las cuales deberán ser consideradas en estudios futuros.

A pesar de estas limitaciones, la arquitectura propuesta sienta una base sólida para el desarrollo de sistemas inteligentes de monitoreo ganadero orientados a la operación autónoma y de bajo consumo. El equilibrio alcanzado entre latencia, eficiencia energética y reducción de comunicaciones posiciona al enfoque de Edge AI embebido como una solución prometedora para aplicaciones agropecuarias en contextos con recursos limitados.

En este sentido, los resultados obtenidos sugieren que, en escenarios rurales con recursos limitados, las decisiones arquitectónicas a nivel de sistema tienen un impacto mayor en la viabilidad operativa que incrementos marginales en la complejidad o precisión del modelo de inteligencia artificial. Este trabajo evidencia que un diseño orientado a la autonomía, la eficiencia energética y la reducción de comunicaciones constituye un factor determinante para la adopción real de soluciones de Edge AI en la ganadería extensiva.

Desde la perspectiva de los sistemas en tiempo real, la arquitectura propuesta implementa un pipeline distribuido de procesamiento de flujos en el que los datos transitan continuamente desde la etapa de adquisición hasta la inferencia local. Dicho diseño se alinea con los paradigmas de procesamiento complejo de eventos (CEP) y procesamiento de flujos en el borde (edge stream processing), distribuyendo las tareas computacionales con el fin de minimizar la latencia [21], [22]. En este contexto, el nodo embebido actúa como un procesador de flujos de primer nivel, encargado del filtrado y la extracción de características en tiempo real.

Para trabajos futuros podrían incorporar esquemas de agregación jerárquica, permitiendo analítica multinivel sin perder la autonomía local. Extensiones futuras de esta arquitectura podrían incorporar modelos jerárquicos de procesamiento de flujos, donde múltiples nodos de borde

alimenten nodos de procesamiento intermedio, permitiendo analítica multinivel y procesamiento de eventos más complejo, sin perder la autonomía local ni la eficiencia energética.

CONCLUSIONES Y TRABAJOS FUTUROS

En este trabajo se presentó una arquitectura de Edge AI embebida orientada al monitoreo del ganado en entornos rurales con recursos limitados, abordando de manera integrada los desafíos asociados a la conectividad intermitente, la eficiencia energética y la necesidad de operación autónoma. El enfoque propuesto prioriza el procesamiento local de los datos y la toma de decisiones autónoma en el nodo embebido, reduciendo la dependencia de infraestructuras centralizadas basadas en la nube.

Desde una perspectiva ingenieril, la arquitectura demuestra ser técnicamente viable para su implementación sobre plataformas embebidas de bajo consumo, permitiendo latencias compatibles con aplicaciones de monitoreo en tiempo casi real y una reducción significativa del volumen de datos transmitidos. Estos resultados confirman que el uso de Edge AI embebido constituye una alternativa adecuada para sistemas de ganadería de precisión desplegados en escenarios rurales extensivos.

Si bien la evaluación experimental se centró en la validación arquitectónica del sistema, los resultados obtenidos sientan una base sólida para el desarrollo de soluciones más avanzadas orientadas al monitoreo persistente del comportamiento animal. El trabajo no persigue en esta etapa la validación diagnóstica de eventos específicos, sino la demostración de la factibilidad técnica y operativa del enfoque propuesto.

Como trabajos futuros, se prevé ampliar la evaluación experimental mediante despliegues prolongados en entornos productivos reales, incorporando métricas cuantitativas más detalladas de consumo energético, latencia y reducción de comunicaciones. Asimismo, se explorará la integración de modelos de inteligencia artificial adaptativos, estrategias de actualización remota del modelo y la incorporación de nuevas fuentes de datos, con el objetivo de avanzar hacia sistemas de monitoreo ganadero inteligentes, escalables y transferibles a contextos reales de producción.

En conjunto, las contribuciones presentadas permiten avanzar desde enfoques experimentales hacia una solución de ingeniería aplicada, validada a nivel de prototipo y orientada a su despliegue en condiciones reales de ganadería extensiva. La arquitectura propuesta demuestra que el uso de Edge AI embebido y autónomo constituye una alternativa técnica viable, escalable y energéticamente eficiente para el monitoreo ganadero en entornos rurales con infraestructura limitada, sentando las bases para su transferencia tecnológica y adopción en sistemas productivos reales.

AGRADECIMIENTOS

Los autores agradecen a la Universidad Siglo 21 por el soporte institucional y por su financiamiento.

REFERENCIAS

- [1] I. Halachmi, M. Guarino, M. Bewley, and M. Pastell, "Smart animal agriculture: Application of real-time sensors to improve animal well-being and production," *Annual Review of Animal Biosciences*, vol. 7, pp. 403–425, 2019.
- [2] D. Berckmans, "Precision livestock farming technologies for welfare management in intensive livestock systems," *Rev. Sci. Tech. Off. Int. Epiz.*, vol. 33, no. 1, pp. 189–196, 2014.
- [3] M. Wolfert, L. Ge, C. Verdouw, and M. Bogaardt, "Big data in smart farming – A review," *Agricultural Systems*, vol. 153, pp. 69–80, 2017.
- [4] W. Shi et al., "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, Oct. 2016.
- [5] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30–39, Jan. 2017.
- [6] F. Conti, D. Rossi, A. Pullini, and L. Benini, "Energy-efficient deep learning inference on edge devices," *ACM Transactions on Embedded Computing Systems*, vol. 20, no. 5s, pp. 1–26, 2021.
- [7] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70–90, 2018.
- [8] J. Gubbi et al., "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.
- [9] R. Want et al., "Enabling the Internet of Things," *Computer*, vol. 48, no. 1, pp. 28–35, Jan. 2015.
- [10] S. Li, L. Da Xu, and S. Zhao, "The internet of things: A survey," *Information Systems Frontiers*, vol. 17, no. 2, pp. 243–259, 2015.
- [11] A. D. Porto et al., "Automatic detection of lameness in dairy cattle using computer vision techniques," *Computers and Electronics in Agriculture*, vol. 119, pp. 17–27, 2015.
- [12] J. Nasirahmadi et al., "Deep learning for detection of cattle using depth images," *Computers and Electronics in Agriculture*, vol. 173, 2020.
- [13] A. Kamilaris, A. Kartakoullis, and F. X. Prenafeta-Boldú, "A review on the practice of big data analysis in agriculture," *Computers and Electronics in Agriculture*, vol. 143, pp. 23–37, 2017.
- [14] J. Nasirahmadi, S. Edwards, and B. Sturm, "Implementation of machine vision for detecting behaviour of cattle and pigs," *Livestock Science*, vol. 202, pp. 25–38, 2017.
- [15] M. Pastell et al., "Monitoring cow behaviour using accelerometers," *Computers and Electronics in Agriculture*, vol. 100, pp. 69–76, 2014.
- [16] N. Sklavos and X. Zhang, "Low power embedded systems: Architecture, design and applications," *Springer*, 2013.
- [17] F. Popescu et al., "Energy consumption optimization in IoT systems using edge computing," *Sensors*, vol. 19, no. 8, 2019.
- [18] M. Satyanarayanan et al., "Edge analytics in the Internet of Things," *IEEE Pervasive Computing*, vol. 14, no. 2, pp. 24–31, 2015.
- [19] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.
- [20] T. Akidau et al., "The Dataflow Model: A Practical Approach to Balancing Correctness, Latency, and Cost in Massive-Scale, Unbounded, Out-of-Order Data Processing," *VLDB*, 2015.
- [21] T. Rausch, S. Nastic, y S. Dustdar, "EdgeCEP: Fully-Distributed Complex Event Processing on IoT Edges," en *2017 13th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2017, pp. 121-129.
- [22] J. Chen, L. Lan, R. Shi, y B. Wang, "A Universal Complex Event Processing Mechanism Based on Edge Computing for Internet of Things Real-Time Monitoring," *IEEE Access*, vol. 7, pp. 101865-101878, 2019.