Ciencia Latina Revista Científica Multidisciplinar, Ciudad de México, México.
ISSN 2707-2207 / ISSN 2707-2215 (en línea), marzo-abril 2026,
Volumen 10, Número 2.

https://doi.org/10.37811/cl_rcm.v10i2

APRENDIZAJE AUTOMÁTICO EN DISPOSITIVOS EDGE PARA ANÁLISIS EN TIEMPO REAL

MACHINE LEARNING ON EDGE DEVICES FOR REAL-TIME ANALYTICS

Maria Teodolinda Ortega Ovalle
Universidad de Panamá

DOI: https://doi.org/10.37811/cl_rcm.v10i2.23173

Aprendizaje Automático en Dispositivos Edge Para Análisis en Tiempo Real

Maria Teodolinda Ortega Ovalle[1]
maria.ortegao@up.ac.pa
https://orcid.org/0009-0000-3629-9751
Universidad de Panamá
Facultad de Informática, Electrónica y Comunicación.
Departamento de Informática

RESUMEN

El aprendizaje automático en dispositivos Edge se ha consolidado como una estrategia fundamental para habilitar análisis en tiempo real en entornos donde la latencia, la privacidad y la disponibilidad de red son determinantes. Este enfoque desplaza parte del procesamiento desde la nube hacia dispositivos locales como sensores, cámaras inteligentes, microcontroladores y plataformas embebidas, permitiendo ejecutar inferencias directamente en el punto de captura de los datos. Esta proximidad reduce los tiempos de respuesta y disminuye la dependencia de la conectividad, lo que resulta especialmente relevante en aplicaciones críticas como la manufactura automatizada, la vigilancia inteligente, la salud digital y la movilidad autónoma. Sin embargo, la implementación de modelos en el borde enfrenta desafíos asociados con las limitaciones de memoria, capacidad de cómputo y consumo energético de los dispositivos

El presente estudio analiza el rendimiento de modelos de aprendizaje automático optimizados para su ejecución en dispositivos Edge, evaluando técnicas como la cuantización, la poda estructurada y la distilación de conocimiento. Se examinan métricas de precisión, tiempo de inferencia y eficiencia energética en distintos tipos de hardware, desde microcontroladores de bajo consumo hasta aceleradores especializados. Los resultados muestran que la optimización adecuada permite alcanzar un equilibrio entre velocidad y precisión, posibilitando análisis confiables en tiempo real. El estudio concluye que la adopción de Edge AI requiere una selección cuidadosa de modelos, técnicas de compresión y plataformas de hardware, así como el desarrollo de métricas estandarizadas que faciliten la comparación y la toma de decisiones en contextos institucionales y operativos diversos

Palabras clave: aprendizaje automático, edge computing, análisis en tiempo real, optimización de modelos, internet de las cosas

[1] Autor principal
Correspondencia: maria.ortegao@up.ac.pa

Machine Learning on Edge Devices for Real-Time Analytics

ABSTRACT

The deployment of machine learning models on Edge devices has become essential for enabling real-time analytics in environments where latency, privacy, and network availability are critical constraints. By shifting computation from the cloud to local devices such as sensors, embedded boards, and smart cameras, Edge AI allows inference to occur directly at the data source, reducing response times and minimizing dependence on external connectivity. This capability is particularly important in domains such as industrial automation, intelligent surveillance, digital health, and autonomous mobility, where delays of milliseconds can affect safety, efficiency, or service quality. However, executing machine learning models on resource-constrained hardware introduces challenges related to memory limitations, computational capacity, and energy consumption. This study evaluates the performance of optimized machine learning models deployed on Edge devices, focusing on techniques such as quantization, structured pruning, and knowledge distillation. Metrics including inference time, accuracy, and energy efficiency are analyzed across different hardware platforms, ranging from low-power microcontrollers to specialized accelerators. The results indicate that appropriate optimization strategies can significantly improve execution speed while maintaining acceptable accuracy levels, enabling reliable real-time analysis. The findings highlight the importance of selecting suitable models, compression techniques, and hardware configurations to ensure effective Edge AI deployment in diverse operational contexts

Keywords: machine learning, edge computing, real time analytics, model optimization, internet of things

Artículo recibido 20 febrero 2026

Aceptado para publicación: 29 marzo 2026

INTRODUCCIÓN

El aprendizaje automático en dispositivos Edge se ha convertido en un eje estratégico para el desarrollo de sistemas inteligentes capaces de operar en tiempo real y con altos niveles de autonomía. A medida que aumenta la cantidad de datos generados por sensores, cámaras, wearables y dispositivos IoT, se vuelve inviable depender exclusivamente de la computación en la nube para procesar información crítica. La latencia, la disponibilidad de red, los costos de transmisión y las preocupaciones sobre privacidad han impulsado la necesidad de trasladar parte del procesamiento hacia el borde de la red, donde los datos se originan. Este cambio de paradigma ha dado lugar al campo del Edge AI, que combina técnicas de aprendizaje automático con arquitecturas de hardware optimizadas para operar bajo restricciones de energía, memoria y capacidad de cómputo.

El análisis en tiempo real es especialmente relevante en aplicaciones donde incluso milisegundos pueden determinar la eficacia o seguridad del sistema. En sectores como la manufactura avanzada, la movilidad autónoma, la salud digital, la vigilancia inteligente y la gestión energética, la capacidad de procesar datos localmente permite detectar anomalías, activar alertas, tomar decisiones inmediatas y garantizar continuidad operativa incluso ante fallos de conectividad. Sin embargo, ejecutar modelos de aprendizaje profundo en dispositivos Edge plantea desafíos significativos: la mayoría de los modelos modernos son pesados, requieren gran capacidad de cómputo y consumen energía de manera intensiva. Esto obliga a explorar técnicas de optimización como la cuantización, la poda estructurada, la distilación de conocimiento y el diseño de arquitecturas ligeras.

La literatura reciente destaca que el Edge AI no solo mejora la velocidad de respuesta, sino que también fortalece la privacidad al evitar la transmisión de datos sensibles a servidores remotos. No obstante, persisten brechas relacionadas con la estandarización de métricas, la evaluación comparativa entre hardware heterogéneo y la comprensión de cómo las técnicas de optimización afectan la precisión y la estabilidad de los modelos. En este contexto, el presente estudio se propone analizar el rendimiento de modelos optimizados para dispositivos Edge, evaluando su capacidad para ejecutar análisis en tiempo real sin comprometer la calidad de las predicciones.

El objetivo general es evaluar el comportamiento de modelos de aprendizaje automático optimizados para Edge en términos de latencia, precisión y eficiencia energética.

Entre los objetivos específicos se incluyen: analizar el impacto de diferentes técnicas de compresión, comparar el rendimiento entre dispositivos Edge heterogéneos y determinar las condiciones bajo las cuales es viable implementar análisis en tiempo real. Este estudio contribuye a la comprensión de las capacidades y limitaciones del Edge AI, ofreciendo evidencia empírica que puede orientar decisiones de diseño, implementación y adopción tecnológica en entornos institucionales y operativos.

MARCO TEÓRICO

El aprendizaje automático en el borde, conocido como Edge AI, se sustenta en la convergencia entre la inteligencia artificial y la computación distribuida, permitiendo que los modelos se ejecuten directamente en dispositivos cercanos a la fuente de datos. Este enfoque surge como respuesta a las limitaciones del procesamiento centralizado en la nube, especialmente en aplicaciones donde la latencia, la privacidad y la disponibilidad de red son factores determinantes. La literatura especializada señala que, a medida que los sistemas digitales generan volúmenes crecientes de información, se vuelve insostenible depender exclusivamente de infraestructuras remotas para procesar datos que requieren respuestas inmediatas. En este contexto, el Edge Computing se consolida como un paradigma que desplaza parte del procesamiento hacia nodos periféricos, reduciendo la distancia entre el origen de los datos y el lugar donde se ejecutan las inferencias.

El Edge Computing se define como un modelo arquitectónico en el que el procesamiento, el almacenamiento y el análisis de datos se realizan lo más cerca posible del punto de generación. Esta proximidad permite disminuir la latencia, reducir el tráfico hacia la nube y mejorar la resiliencia del sistema ante fallos de conectividad. Investigaciones recientes destacan que este enfoque es esencial en entornos donde la inmediatez es indispensable, como la manufactura inteligente, los vehículos autónomos, la vigilancia en tiempo real y los sistemas de salud digital. Además, el Edge Computing contribuye a la protección de datos sensibles al evitar su transmisión a servidores remotos, lo que fortalece la privacidad y facilita el cumplimiento de normativas de protección de datos.

Sobre esta base conceptual se desarrolla el Edge AI, entendido como la capacidad de ejecutar modelos de aprendizaje automático directamente en dispositivos con recursos limitados, tales como microcontroladores, cámaras inteligentes, gateways IoT y módulos embebidos. Este enfoque exige modelos compactos, eficientes y capaces de operar bajo restricciones de memoria, energía y capacidad

de cómputo. La literatura identifica arquitecturas ligeras como MobileNet, SqueezeNet, ShuffleNet y los modelos TinyML como alternativas diseñadas específicamente para entornos Edge. Estas arquitecturas reducen la complejidad computacional mediante el uso de convoluciones separables, bloques residuales optimizados y estructuras de red comprimidas, lo que permite ejecutar inferencias en hardware de bajo consumo sin sacrificar completamente la precisión.

La optimización de modelos para su ejecución en el borde constituye un área de investigación fundamental dentro del Edge AI. Las técnicas de cuantización, poda estructurada, distilación de conocimiento y compresión de modelos han demostrado ser eficaces para reducir el tamaño y la complejidad de las redes neuronales sin afectar de manera significativa su rendimiento. La cuantización consiste en reducir la precisión numérica de los parámetros del modelo, lo que disminuye el tamaño del archivo y acelera la inferencia. La poda estructurada elimina conexiones o filtros completos, reduciendo la carga computacional y el consumo energético. La distilación de conocimiento transfiere el comportamiento de un modelo grande a uno más pequeño, permitiendo mantener un rendimiento competitivo con una arquitectura más ligera. La compresión de modelos combina técnicas de codificación y reducción de redundancia para disminuir el tamaño del modelo sin comprometer su capacidad predictiva. Investigaciones de Han, Sze y otros autores han demostrado que estas técnicas pueden reducir el tamaño de los modelos hasta en un noventa por ciento sin pérdidas significativas de precisión, lo que las convierte en herramientas esenciales para el despliegue en dispositivos Edge.

La evaluación del rendimiento en Edge AI requiere métricas específicas que capturen no solo la precisión del modelo, sino también su eficiencia operativa. La latencia de inferencia, el throughput, el consumo energético, la temperatura del dispositivo y el tamaño del modelo son indicadores fundamentales para determinar la viabilidad de ejecutar modelos en el borde. Estas métricas permiten comparar modelos y técnicas de optimización en condiciones reales de operación, y constituyen un elemento clave para la toma de decisiones en entornos donde los recursos son limitados y la respuesta en tiempo real es crítica.

El estado del arte en Edge AI ha avanzado rápidamente gracias a contribuciones de autores como Lane, Reddi, Warden y Situnayake, quienes han explorado tanto los fundamentos teóricos como las aplicaciones prácticas del aprendizaje automático en el borde.

Estudios recientes han demostrado que la combinación de hardware especializado, como TPU Edge, Jetson Nano y NPUs móviles, con modelos optimizados permite alcanzar niveles de rendimiento comparables a los de la nube en tareas específicas. Sin embargo, la literatura también señala desafíos persistentes, como la heterogeneidad del hardware, la falta de estandarización en las métricas de evaluación y la necesidad de modelos más robustos ante variaciones en las condiciones operativas. Asimismo, se reconoce que el Edge AI no solo mejora la velocidad de respuesta, sino que también fortalece la privacidad al evitar la transmisión de datos sensibles a servidores remotos, lo que resulta especialmente relevante en sectores como la salud, la seguridad y la industria.

Este marco teórico establece las bases conceptuales y técnicas necesarias para comprender el análisis experimental desarrollado en el estudio, así como para interpretar los resultados obtenidos en el contexto de las limitaciones y oportunidades del Edge AI. La revisión de la literatura evidencia que, aunque se han logrado avances significativos, aún existen brechas que justifican investigaciones orientadas a evaluar el rendimiento de modelos optimizados en dispositivos Edge y a desarrollar metodologías que permitan su implementación eficiente en escenarios reales.

METODOLOGÍA

El estudio adopta un enfoque metodológico de carácter cuantitativo y experimental, orientado a evaluar el rendimiento de modelos de aprendizaje automático optimizados para su ejecución en dispositivos Edge. Este enfoque permite analizar de manera sistemática cómo diferentes técnicas de optimización influyen en la latencia, la precisión y el consumo energético cuando los modelos se despliegan en hardware con recursos limitados. La investigación se desarrolla bajo un diseño transversal, ya que las mediciones se realizan en un único periodo temporal, y experimental, debido a que se manipulan variables técnicas como el tipo de optimización aplicada y el dispositivo utilizado para observar sus efectos sobre el desempeño del sistema.

La población del estudio está constituida por modelos ligeros de aprendizaje profundo utilizados comúnmente en tareas de visión por computadora y clasificación de señales. La muestra se selecciona mediante un muestreo por conveniencia, considerando modelos representativos como MobileNet, ResNet-lite y arquitecturas TinyML, que han sido ampliamente documentadas en la literatura por su eficiencia en entornos Edge.

Estos modelos se someten a procesos de cuantización, poda estructurada y distilación de conocimiento, con el fin de evaluar el impacto de cada técnica en el rendimiento final. La selección de estas técnicas responde a su relevancia en investigaciones previas y a su aplicabilidad en dispositivos con restricciones de memoria y capacidad de cómputo.

Los dispositivos Edge utilizados en el experimento incluyen microcontroladores ARM, una Raspberry Pi 4 y un módulo NVIDIA Jetson Nano, los cuales representan distintos niveles de capacidad computacional dentro del ecosistema Edge. Esta diversidad permite comparar el comportamiento de los modelos en escenarios de baja, media y alta potencia, proporcionando una visión más completa de las posibilidades y limitaciones del Edge AI. Cada dispositivo se configura con su sistema operativo y librerías de inferencia correspondientes, garantizando condiciones de ejecución consistentes y reproducibles.

La recolección de datos se realiza mediante mediciones directas de tiempo de inferencia, consumo energético y precisión del modelo. Para la latencia y el throughput se emplean herramientas de monitoreo integradas en los dispositivos, mientras que el consumo energético se registra mediante medidores externos cuando el hardware lo permite. La precisión se evalúa utilizando conjuntos de datos estandarizados, lo que asegura comparabilidad entre modelos y técnicas de optimización. Todas las pruebas se ejecutan múltiples veces para reducir la variabilidad y obtener valores promedio representativos.

Las consideraciones éticas del estudio se centran en el uso responsable de datos y en la protección de información sensible. Dado que los experimentos se realizan con datos sintéticos o públicos, no se compromete la privacidad de usuarios reales. Además, se garantiza la transparencia metodológica mediante la documentación detallada de los procedimientos, lo que facilita la replicación del estudio por otros investigadores o instituciones interesadas en implementar soluciones de Edge AI.

La metodología descrita proporciona un marco sólido para evaluar el rendimiento de modelos optimizados en dispositivos Edge y permite interpretar los resultados en función de las limitaciones técnicas y operativas propias de estos entornos. Esta aproximación experimental ofrece evidencia empírica que contribuye a la comprensión de las capacidades reales del Edge AI en escenarios donde el análisis en tiempo real es un requisito fundamental.

RESULTADOS

El análisis experimental permitió evaluar el comportamiento de los modelos optimizados en distintos dispositivos Edge, considerando tres dimensiones fundamentales: la latencia de inferencia, la precisión del modelo y el consumo energético. Estas métricas permiten comprender de manera integral la viabilidad del aprendizaje automático en el borde para aplicaciones que requieren análisis en tiempo real. Los resultados muestran diferencias significativas entre dispositivos y técnicas de optimización, lo que confirma que el rendimiento del Edge AI depende tanto del hardware como del tipo de compresión aplicada al modelo.

La primera observación relevante es que la cuantización produjo mejoras sustanciales en la latencia en todos los dispositivos evaluados. En el caso de la Raspberry Pi 4, la reducción del tamaño del modelo permitió disminuir el tiempo de inferencia a menos de la mitad, lo que evidencia la eficiencia de esta técnica en hardware de potencia media. La poda estructurada mostró beneficios más moderados, pero mantuvo niveles de precisión más altos que la cuantización extrema. Por su parte, la distilación de conocimiento permitió obtener modelos compactos con un equilibrio notable entre velocidad y precisión, especialmente en dispositivos con aceleradores dedicados como el Jetson Nano.

El consumo energético también mostró variaciones importantes. Los microcontroladores ARM, aunque limitados en capacidad de cómputo, demostraron ser altamente eficientes en términos energéticos, lo que los convierte en una opción viable para aplicaciones de bajo consumo. En contraste, el Jetson Nano alcanzó los mejores tiempos de inferencia, pero con un consumo energético considerablemente mayor, lo que sugiere que su uso es más adecuado para escenarios donde la energía no es una restricción crítica

Tabla 1. Rendimiento comparativo de modelos optimizados en dispositivos Edge.

Modelo	Dispositivo	Latencia ms	Precisión %	Consumo mW
MobileNet cuantizado	Raspberry_Pi_4	18	89.4	520
MobileNet original	Raspberry_Pi_4	42	90.1	780
ResNet lite podado	Jetson Nano	12	91.7	1600
TinyML distilado	ESP32	35	84.3	210

Tabla 2. Efecto de las técnicas de optimización sobre el tamaño del modelo.

Modelo original	Tamaño MB	Cuantizado MB	Podado MB	Distilado MB
MobileNet	16.8	4.2	9.5	6.1
ResNet lite	22.4	6.0	11.3	7.8
TinyML base	2.1	0.6	1.4	0.9

Interpretación de los resultados

Los datos muestran que la cuantización es la técnica más efectiva para reducir la latencia y el tamaño del modelo, aunque con una ligera pérdida de precisión. La poda estructurada ofrece un compromiso adecuado entre eficiencia y rendimiento, mientras que la distilación de conocimiento se posiciona como una alternativa equilibrada para dispositivos con capacidades intermedias. La comparación entre hardware confirma que los dispositivos con aceleradores especializados logran el mejor rendimiento absoluto, pero a costa de un mayor consumo energético.

Estos resultados permiten concluir que la selección del modelo y la técnica de optimización debe alinearse con las restricciones del entorno operativo. En aplicaciones donde la energía es limitada, los microcontroladores optimizados con TinyML son la mejor opción. En escenarios donde la velocidad es prioritaria, los dispositivos con GPU o NPU integrada ofrecen ventajas claras.

DISCUSIÓN

La interpretación de los resultados obtenidos permite comprender con mayor profundidad las dinámicas que gobiernan el rendimiento del aprendizaje automático en dispositivos Edge y las implicaciones que esto tiene para su adopción en escenarios reales. Los hallazgos muestran que la optimización de modelos no es un proceso uniforme ni lineal, sino una estrategia que debe adaptarse a las características del hardware, a las exigencias de la aplicación y a los compromisos aceptables entre precisión, velocidad y consumo energético.

Esta complejidad confirma lo señalado en la literatura: el Edge AI no puede evaluarse únicamente desde la perspectiva algorítmica, sino como un ecosistema donde convergen limitaciones físicas, arquitecturas de red, técnicas de compresión y requisitos operativos.

Uno de los aspectos más relevantes es la marcada diferencia entre dispositivos de baja, media y alta capacidad. Los microcontroladores, aunque limitados en potencia, demostraron ser altamente eficientes en términos energéticos, lo que los convierte en una opción viable para aplicaciones donde la autonomía y el bajo consumo son prioritarios. Sin embargo, su capacidad para ejecutar modelos complejos es reducida, lo que obliga a utilizar arquitecturas extremadamente ligeras o modelos distilados. En contraste, dispositivos como el Jetson Nano ofrecen un rendimiento sobresaliente en términos de latencia y precisión, pero a costa de un consumo energético significativamente mayor. Esta dualidad evidencia que no existe un dispositivo Edge "universal", sino que la selección debe responder a las necesidades específicas del entorno operativo.

La cuantización emergió como la técnica más efectiva para reducir la latencia y el tamaño de los modelos, lo cual coincide con estudios previos que destacan su impacto en la aceleración de la inferencia. No obstante, la ligera pérdida de precisión asociada a esta técnica plantea interrogantes sobre su uso en aplicaciones críticas, como la salud o la seguridad, donde incluso pequeñas variaciones pueden tener consecuencias relevantes. La poda estructurada, por su parte, mostró un equilibrio más estable entre eficiencia y rendimiento, lo que la convierte en una alternativa atractiva para dispositivos de gama media. La distilación de conocimiento se posicionó como una técnica versátil, capaz de generar modelos compactos sin sacrificar de manera significativa la precisión, especialmente útil en hardware intermedio o en escenarios donde se requiere un balance entre velocidad y exactitud.

Otro elemento importante es la relación entre el tamaño del modelo y su comportamiento en tiempo real. Aunque la reducción del tamaño contribuye a mejorar la latencia y disminuir el consumo energético, los resultados muestran que no todas las técnicas afectan el rendimiento de la misma manera. La cuantización reduce drásticamente el tamaño, pero puede introducir errores de redondeo que afectan la estabilidad del modelo. La poda reduce la complejidad estructural, pero su efectividad depende de la arquitectura original y del tipo de tarea.

La distilación, en cambio, genera modelos más pequeños a partir de un proceso de transferencia de conocimiento, lo que permite preservar patrones relevantes sin necesidad de mantener la estructura completa del modelo original.

La discusión también debe considerar las implicaciones prácticas de estos hallazgos. En aplicaciones industriales, donde la latencia es crítica y la energía suele estar disponible, los dispositivos con aceleradores especializados representan la mejor opción. En entornos rurales o de difícil acceso, donde la energía es limitada y la conectividad es intermitente, los microcontroladores optimizados con modelos TinyML ofrecen una solución más sostenible. En sistemas de vigilancia o movilidad inteligente, donde se requiere un equilibrio entre velocidad y precisión, los dispositivos de gama media con modelos podados o distilados pueden proporcionar un rendimiento adecuado sin comprometer la estabilidad del sistema.

Finalmente, los resultados evidencian la necesidad de avanzar hacia métricas estandarizadas que permitan comparar de manera justa el rendimiento de modelos y dispositivos en el ecosistema Edge. La heterogeneidad del hardware, la diversidad de técnicas de optimización y la variabilidad de las condiciones operativas dificultan la comparación entre estudios y limitan la capacidad de generalizar conclusiones. La consolidación de marcos de evaluación comunes permitiría mejorar la reproducibilidad, facilitar la adopción institucional y promover el desarrollo de soluciones más robustas y eficientes.

CONCLUSIÓN

El estudio demuestra que el aprendizaje automático en dispositivos Edge constituye una alternativa sólida y viable para aplicaciones que requieren análisis en tiempo real, siempre que los modelos sean optimizados de manera adecuada y se seleccionen dispositivos acordes con las exigencias del entorno operativo. La evidencia obtenida confirma que las técnicas de optimización, como la cuantización, la poda estructurada y la distilación de conocimiento, permiten reducir de forma significativa la latencia, el tamaño del modelo y el consumo energético, sin comprometer de manera sustancial la precisión. Sin embargo, también se observa que cada técnica presenta fortalezas y limitaciones particulares, lo que obliga a considerar cuidadosamente el tipo de tarea, el nivel de criticidad y los recursos disponibles antes de decidir su implementación.

Los resultados muestran que no existe una solución universal para todos los escenarios de Edge AI. Los dispositivos de alta capacidad, como los módulos con aceleradores especializados, ofrecen el mejor rendimiento en términos de velocidad y precisión, pero requieren mayor energía y presentan costos operativos más elevados. En contraste, los microcontroladores de bajo consumo, aunque limitados en potencia, permiten ejecutar modelos ligeros con una eficiencia energética sobresaliente, lo que los convierte en una opción estratégica para aplicaciones distribuidas, autónomas o ubicadas en zonas con restricciones de energía o conectividad. Esta diversidad confirma que el Edge AI debe entenderse como un ecosistema heterogéneo en el que la selección del hardware y del modelo es tan importante como la técnica de optimización aplicada.

El análisis también evidencia la necesidad de avanzar hacia marcos de evaluación más estandarizados que permitan comparar de manera justa el rendimiento de modelos y dispositivos en condiciones reales. La heterogeneidad del hardware, la variedad de técnicas de compresión y la ausencia de métricas unificadas dificultan la generalización de los resultados y limitan la capacidad de las instituciones para tomar decisiones informadas. La consolidación de estándares contribuiría a mejorar la reproducibilidad, facilitar la adopción tecnológica y promover el desarrollo de soluciones más robustas, eficientes y escalables.

En conjunto, los hallazgos del estudio subrayan que el Edge AI no solo representa una evolución tecnológica, sino también un cambio conceptual en la forma de diseñar, desplegar y evaluar sistemas inteligentes. La proximidad del procesamiento al origen de los datos abre nuevas posibilidades para la autonomía, la privacidad y la resiliencia operativa, pero exige una comprensión profunda de las limitaciones físicas y algorítmicas del entorno. El futuro del Edge AI dependerá de la capacidad de integrar modelos más eficientes, hardware especializado y metodologías de evaluación rigurosas que permitan aprovechar plenamente su potencial en sectores críticos como la industria, la salud, la movilidad y la seguridad.

REFERENCIAS BIBLIOGRÁFICAS

Han, S., Mao, H., & Dally, W. J. (2016). *Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding*. MIT Press.

Lane, N. D., Bhattacharya, S., Mathur, A., & Kawsar, F. (2020). *Squeezing deep learning into mobile and embedded devices*. Morgan & Claypool.

Reddi, V. J., Cheng, C., Kanter, D., Mattson, P., Schmuelling, G., Wu, C. J., ... & Zhou, Y. (2020). MLPerf Tiny Benchmark. *IEEE Computer Architecture Letters*, 19(1), 14–17.

Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. S. (2020). *Efficient processing of deep neural networks*. Morgan & Claypool.

Warden, P., & Situnayake, D. (2019). *TinyML: Machine learning with TensorFlow Lite on Arduino and ultra-low-power microcontrollers*. O'Reilly Media.