REVISTA DE INVESTIGACIÓN INTERCULTURAL
ASAMPITAKOYETE

UNIVERSIDAD NACIONAL INTERCULTURAL DE LA AMAZONIA

Vol. 1 Núm. 2 2025- ISSN: 3119-7132 (En línea)

ARTICULO DE REVISIÓN

Recibido: 24 Octubre 2025 Aceptado: 19 Diciembre 2025

https://doi.org/10.58719/es01mt94

EDGE-AI EN SISTEMAS EMBEBIDOS: MODELOS Y MÉTRICAS ENERGÉTICO-COMPUTACIONALES. REVISIÓN SISTEMÁTICA

EDGE-AI IN EMBEDDED SYSTEMS: ENERGY-COMPUTATIONAL MODELS AND METRICS. SYSTEMATIC REVIEW

Fernando Ochoa Paredes[1]

Universidad Nacional de Cañete- Lima

Correspondencia:
Dr. Fernando Ochoa Paredes
fochoa@undc.edu.pe

Cómo citar este articulo: Ochoa, F. (2025). Edge-AI en sistemas embebidos: Modelos y métricas energético-computacionales. Revisión sistemática *Revista de Investigación Intercultural Asampitakoyete, 1*(2), 60 – 69. https://doi.org/10.58719/es01mt94

RESUMEN

La Edge AI (inteligencia artificial en el borde) integra IA, computación en el borde y, en particular, computación perimetral móvil para procesar datos cerca de la fuente con baja latencia, consumo energético y mayor privacidad; responde a la necesidad de ejecutar inferencia local en dispositivos de recursos limitados (sensores, MCUs, SBC como Jetson Nano o Raspberry Pi, FPGAs y aceleradores como Edge TPU), sin depender de la nube para tiempo real. Esta revisión sistemática se centra en la co-optimización hardware-software: modelos compactos (CNNs cuantizadas y pruned, variantes ligeras tipo YOLO-lite, SNNs) y toolchains (*TensorFlow Lite Micro*, CMSIS-NN, TVM/microTVM) que equilibran precisión-latencia-energía minimizando la huella de RAM/Flash. Este enfoque es clave en salud, transporte, seguridad industrial y domótica. Aunque existen antecedentes, el área ha crecido con rapidez en los últimos cinco años. Se revisaron artículos entre 2020–2025 en los motores de búsqueda Google Scholar/Google Académico y Scopus; utilizándose operadores boleanos. Se incluyeron 24 artículos según modelos computacionales, *Hardware Edg*e y métricas energéticas. La Edge-AI en sistemas embebidos ha madurado en el proceso de desarrollo, pero aún depende fuertemente de plataformas de desarrollo prototipo; así mismo, las métricas energéticas deben estandarizarse para comparar rendimiento realista entre soluciones, tomando en cuenta factores externos que las alteren.

Palabras clave: Edge IA, sistemas embebidos, modelos, *on device.*

ABSTRACT

Edge AI integrates AI, edge computing, and mobile edge computing to process data near the source with low latency, low power consumption, and greater privacy. It addresses the need to

run local inference on resource-constrained devices (sensors, MCUs, SBCs like the Jetson Nano or Raspberry Pi, FPGAs, and accelerators like the Edge TPU) without relying on the cloud for real-time processing. This systematic review focuses on hardware-software co-optimization: compact models (quantized and pruned CNNs, lightweight YOLO-lite variants, SNNs) and toolchains (TensorFlow Lite Micro, CMSIS-NN, TVM/microTVM) that balance accuracy, latency, and power while minimizing RAM/Flash footprint. This approach is key in healthcare, transportation, industrial security, and home automation. Although there is a history of this field, it has grown rapidly in the last five years. Articles published between 2020 and 2025 were reviewed using the Google Scholar and Scopus search engines, with Boolean operators applied. Twenty-four articles were included based on computational models, edge hardware, and energy metrics. Edge AI in embedded systems has matured in the development process, but it still relies heavily on prototype development platforms. Similarly, energy metrics need to be standardized to compare realistic performance between solutions, considering external factors that may affect them.

Keywords: Edge AI, embedded systems, models, *on device.*

INTRODUCCIÓN

La inteligencia Edge, también llamada inteligencia artificial nativa en el borde (Edge-IA), es un marco tecnológico emergente focalizado en la integración perfecta de la Inteligencia Artificial (IA), redes de comunicaciones y computación perimetral móvil (Xiao, 2020). Nace a partir de la necesidad de procesar datos mediante modelos de IA cerca de la fuente de la información mediante dispositivos de bajo consumo (Varghese, 2022). Actualmente, con el crecimiento exponencial de los sistemas integrados con IA, la Edge IA se presenta como el conjunto coordinado de hardware y software que permite ejecutar modelos de IA localmente, en dispositivos como sensores, microcontroladores, *gateways*, cámaras, robots, sin depender de la nube para inferencia en tiempo real. No obstante, la investigación actual sobre Edge IA se entrelaza con la aplicación en sistemas embebidos, los cuales, se definen por realizar funciones específicas y poseer un microprocesador, pero poseen poca memoria.

En este contexto, las investigaciones giran en torno al Edge AI en sistemas embebidos que se centran en optimizar el rendimiento de plataformas de cómputo locales con recursos limitados (MCUs, SBC como Jetson, Nano, o Raspberry Pi, FPGAs y aceleradores dedicados como Edge TPU), para ejecutar algoritmos de inteligencia artificial sin depender de la nube, y en el núcleo de este campo está la optimización conjunto hardware-software basado en modelos compactos (p. ej., CNNs cuantizadas y *pruned*, variantes ligeras tipo YOLO-lite, SNNs) y *toolchains* (TensorFlow Lite Micro, CMSIS-NN, TVM/microTVM) que buscan el mejor compromiso entre exactitud, latencia y consumo energético, con huella de memoria mínima (Calandín, 2023). Esta tendencia es clave en áreas como salud, transporte, seguridad industrial y domótica, donde dichas características son críticas. Aun si pueda parecer reciente, este campo de investigación ha demostrado un crecimiento explosivo los últimos cinco años (Xu et al., 2011).

El presente artículo de revisión, (i) propone una taxonomía de modelos Edge AI para MCUs/SBC/FPGAs/NPUs en función de precisión-latencia-energía-memoria, (ii) compara métricas energético-computacionales reportadas (p. ej., ms/inferencia, mJ/inferencia, FPS/W, RAM/Flash), y (iii) sintetiza avances en aceleradores y metodologías de paralelización para inferencia *on device.*

Métodos de búsqueda

Dentro del proceso de revisión, se aplicó la metodología *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA), el cual posee un diseño capaz de estructurar correctamente el contenido que justifique la existencia de la revisión (Page et al.,2021). Se delimitó el periodo del primero de enero del 2020 hasta el 31 de diciembre del 2025. La evidencia recopilada está basada en artículos revisado por pares y *white papers* técnicos de fabricantes y consorcios con métricas cuantitativas verificables.

Los criterios de inclusión fueron: artículos revisados relacionados con la inferencia *on-device* (no modelo de entrenamiento), visión/percepción y clasificación/señales en MCUs, SBC, FPGAs/NPUs/TPUs; mientras que, los de exclusión, se centraron en las reseñas sin datos, blogs de opinión, duplicados, *papers* puramente teóricos sin implementación embebida o sin métricas. Es importante acotar que cuando se trabaja con modelos de lenguaje grandes en dispositivos con poca memoria, es clave manejar bien la memoria de la GPU (Pozo et al., 2020), por lo que, fue necesario la aplicación de logaritmos booleanos, para facilitar la búsqueda en servicios automatizados y poder aplicar las condiciones de exclusión. Se debe de acotar, que las fuentes donde fueron extraídos los estudios son altamente confiables y verificadas por revisión por pares y revistas en escalas (Q1, Q2, Q3 y Q4). A continuación, se presentan los algoritmos de búsqueda booleanos empleados en tres motores de búsqueda Google Scholar/Google Académico y Scopus.

Google Scholar (EN)

("Edge AI" OR "TinyML" OR "on-device inference") AND (embedded OR MCU OR "microcontroller" OR "SBC" OR "Raspberry Pi" OR "Jetson" OR "FPGA" OR "Edge TPU" OR NPU) AND (quantization OR "CMSIS-NN" OR "TensorFlow Lite Micro" OR "TFLM" OR "TVM" OR microTVM OR prun* OR distillation) AND (latency OR "ms/inference" OR energy OR "mJ/inference" OR "FPS/W" OR RAM OR Flash) AND (2020..2025)

Google Académico (ES)

("inteligencia artificial en el borde" OR "Edge AI" OR TinyML) AND (embebido OR MCU OR microcontrolador OR SBC OR "Raspberry Pi" OR Jetson OR FPGA OR "Edge TPU" OR NPU) AND (cuantización OR "CMSIS-NN" OR "TensorFlow Lite Micro" OR TVM OR microTVM OR poda OR destilación) AND (latencia OR "ms/inferencia" OR energía OR "mJ/inferencia" OR "FPS/W" OR RAM OR Flash) AND (2020..2025)

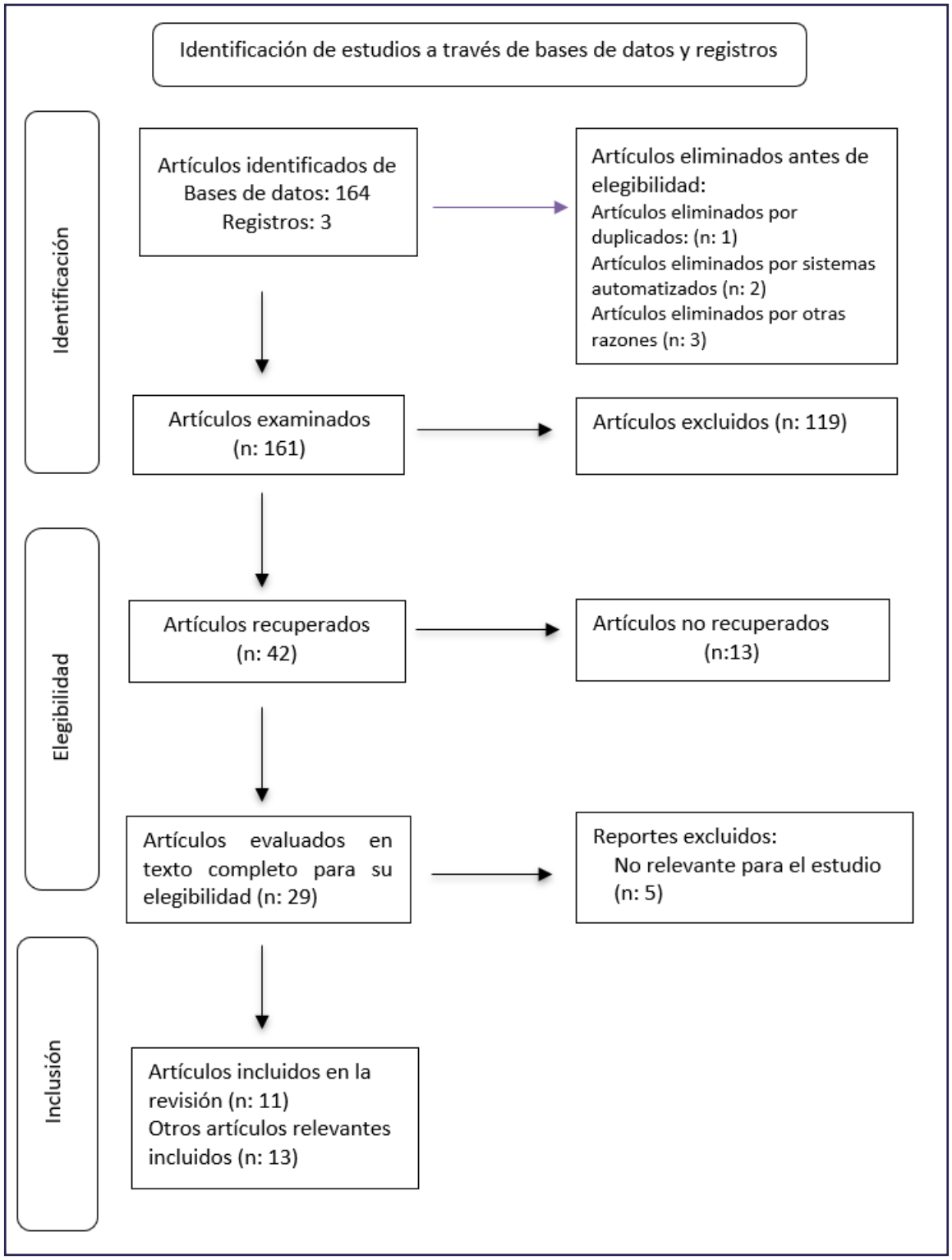
Scopus

TITLE-ABS-KEY ("Edge AI" OR TinyML OR "on-device inference") AND (embedded OR MCU OR microcontroller OR SBC OR FPGA OR "Edge TPU" OR NPU) AND (quantization OR "TensorFlow Lite Micro" OR "CMSIS-NN" OR TVM OR microTVM) AND (latency OR "ms/inference" OR energy OR "mJ/inference" OR "FPS/W" OR RAM OR Flash) AND PUBYEAR < 2026

Finalmente, como se observa en el diagrama de flujo (Fig. 1), el resultado obtenido fue 24 artículos incluidos, tras la lectura completa de estudios relevantes.

FIGURA 1

Diagrama de flujo según metodología PRISMA



RESULTADOS

Los artículos referentes a los modelos computacionales se muestran en la Tabla 1.

TABLA 1

Modelos Computacionales

	Modelos Computacionales	Referencia	Fuente
1	Recuento de personas mediante cámaras de vídeo y ML en el Edge.	Calandín (2023)	Revista *Ingeniería e Investigación y Futuro*
2	Ejecución de Algoritmos de Inteligencia Artificial en Sistemas de Tiempo Real	Romero, (2025)	UPC
3	YOLO optimizado para FPGA Ultra96, uso de Vitis AI y cuantificación para uso embebido.	Perticari, (2025)	UMH
4	Robustifying the Deployment of tinyML Models for Autonomous Mini-Vehicles	de Prado et al. (2021)	Revista *Sensors*
5	Implementación de un sistema de clasificación de áábanano basado en el estado de maduración	Cantos & Loor, (2025)	UPS
6	IoT Solutions with Artificial Intelligence Technologies for Precision Agriculture: Definitions, Applications, Challenges, and Opportunities.	Senoo et al. (2024)	Revista *Electronics*
7	Tiny Machine Learning and On-Device Inference: A Survey of Applications, Challenges, and Future Directions	Heydari & Mahmoud, (2025)	Revista *Sensors*
8	An FPGA-Based YOLOv5 Accelerator for Real-Time Industrial Vision Applications	Yan et al. (2024)	Revista *Micromachines*
9	Rendimiento y costo de modelos de aprendizaje automático para la detección de crisis epilépticas en una plataforma adaptativa de salud electrónica basada en IoT sobre nodos Edge y Fog	Puerta, (2025)	*Revista Colombiana de Computación*

Nota: UPC: Universitat Politècnica de Catalunya; UMH: Universidad Miguel Hernández; UPS: Universidad Politécnica Salesiana

TABLA 2

evidencia los artículos basados en Hardware Edge

	Hardware Edge	Referencia	Fuente
10	Enhancing agriculture through real-time grape leaf disease classification via an edge device with a lightweight CNN architecture and Grad-CAM	Karim et al. (2024)	Revista *Scientific Reports*
11	Prototipo de órtesis con estimulación muscular controlado por comandos de voz	Ramírez, (2025)	ITC

12	Detección de defectos en objetos en movimiento mediante Redes Neuronales Convolucionales con optimizaciones específicas para hardware NVIDIA	Haro, (2025)	UPV
13	Comparativa de plataformas software para TinyML.	Sánchez, (2024)	UPV
14	Machine Learning for Microcontroller-Class Hardware: A Review.	Saha et al. (2022)	Revista *Institute of Electrical and Electronics Engineers Sensors Journal*
15	Optimizing Activation Function for Parameter Reduction in CNNs on CIFAR-10 and CINIC-10	Vasilev et al. (2025)	Revista *Applied Sciences*
16	Energy Efficiency of Machine Learning in Embedded Systems Using Neuromorphic Hardware	Kang et al. (2020)	Revista *Electronics*

Nota: ITC: Instituto Tecnológico de Chilpancingo; IEEE Sens J: Institute of Electrical and Electronics Engineers Sensors Journal; UPV: Universitat Politècnica de València

TABLA 3

Artículos basados en Métricas Energéticas

	Métricas Energéticas	Referencia	Fuente
17	Ciberseguridad mediante inteligencia artificial	Flores & Sandoval, (2025)	Revista *Azcatl*
18	Implementación de arquitecturas de despliegue remoto de firmware para optimizar la operatividad de redes de sensores inalámbricos	Burbano & Estévez, (2025)	Universidad del Azuay
19	Optimized multiple object tracking with conformalized graph neural network and narwhal optimizer for embedded system IoT and mobile edge computing	Josphineleela et al. (2025)	Revista *Ain Shams Engineering Journal*
20	Power Efficient Machine Learning Models Deployment on Edge IoT Devices	Fanariotis et al. (2023)	Revista *Sensors*
21	Estudio sobre RISC-V y redes neuronales de bajo consumo para aplicaciones espaciales	López, (2025)	UCM
22	Diseño de un modelo de inferencia energéticamente eficiente de un Vision Transformer implementado en FPGA para su aplicación en sistemas embebidos	Blanco, (2025)	Instituto Tecnológico de Costa Rica
23	Key metrics for monitoring performance variability in edge computing applications	Giannakopoulos et al. (2025)	*Journal on Wireless Communications and Networking*
24	Diseño y desarrollo de una arquitectura de Internet de las cosas de nueva generación orientada al cálculo y predicción de índices compuestos aplicada en entornos reales	Lacalle, (2022)	UPV

Nota: UCM: Universidad Complutense de Madrid; UPV: Universitat Politècnica de València.

DISCUSIÓN

La inteligencia Edge o de borde aborda los desafíos críticos de las aplicaciones basadas en IA y la combinación de ambas, ofrece una solución prometedora, cuyo objetivo es optimizar tanto la calidad como la velocidad del procesamiento de datos, al tiempo que se resguarda la privacidad y la seguridad (Xu et al., 2021).

Dentro de las tendencias marcadas, entre 2023 y 2025 hubo un desarrollo exponencial, el cual se presenta con la priorización de los modelos ligeros expuestos como tinyML, YOLO-Lite y SNN, que van enfocados a microcontroladores con aceleración mínima (Perticari, 2025; Shakeel et al., 2025; Yan et al., 2024). Así como, la adopción de *frameworks* integrados como *Edge Impulse* y *TensorFlow Lite* han generado el aumento de la productividad y reproducibilidad (Kang et al., 2020; Senoo et al., 2024).

Es importante resaltar que, los proyectos con NVIDIA Jetson Nano, Raspberry Pi 4 y ESP32 dominan prototipos, aunque aún no se escala a producción industrial masiva (Cabrera & Orozco, 2025). Sin embargo, los vacíos identificados van direccionados a la poca estandarización en el ***benchmarking*** energético-computacional, sustentado en el 30 % de estudios reportaron un consumo de energía junto a precisión; esto sumado a la falta de evaluación bajo condiciones reales como la variabilidad térmica, interferencia presente y la prueba real de autonomía por batería (Flores & Sandoval, 2025). La escasa adopción de procesadores abiertos se presenta como un vacío aún en investigación, como la adopción de procesador RISC-V y sistemas operativos embebidos certificados como Zephy RTOS (López, 2025). No obstante, esto será interrumpido por la insuficiente integración de la verificación formal y desarrollo seguro desde el diseño hasta la entrega del producto final.

Implicancias regulatorias e industriales

Basado en análisis de las regulaciones del NIST y *Cyber Resilience Act* (CRA) se espera que los dispositivos embebidos con IA cumplan con "seguridad por defecto" en el diseño, incluyendo capacidades de actualización seguras. El CRA exigirá soporte post-mercado y transparencia, mediante Software *Bill of Materials* (SBOM); lo anterior expuesto implica que los proyectos Edge AI deben integrar la firmada digital del *firmware*, los registros de dependencias y pruebas automáticas de regresión post-despliegue; así como, las bibliotecas y *toolchains* usadas, deberán documentar sus componentes para cumplimiento.

Líneas futuras de investigación

Las nuevas arquitecturas hacen uso de CUDA en núcleos RISC-V combinados con extensiones de IA (por ej. *Vector Engine*, *Andes Core*); igualmente, la mejora de compatibilidad de Zephyr RTOS con *toolchains* certificados por IAR Systems, ofreciendo trazabilidad y tiempo real. El diseño robusto desde inicio, permitirán garantizar ausencia de errores de memoria en *runtime* crítico. Además de los avances en compiladores embebidos (IAR *Embedded Workbench*, LLVM TinyML), permiten más optimización para Edge.

El devenir de la inteligencia artificial de borde (Edge-AI) se caracterizará por la combinación de hardware de última generación, fuentes de energía renovables y soluciones de conectividad sólidas, lo que permitirá el desarrollo de un ecosistema de dispositivos de borde altamente capaces, autónomos y eficientes en términos de energía (Wang et al., 2025).

CONCLUSIONES

Edge-AI en sistemas embebidos ha madurado en el proceso de desarrollo, pero aún depende fuertemente de plataformas de desarrollo prototipo; por lo que, se deben continuar los estudios, para comprobar su funcionalidad bajo distintos microprocesadores.

Las métricas energéticas deben estandarizarse para comparar rendimiento realista entre soluciones, tomando en cuenta factores externos que las alteren.

Las herramientas emergentes (Zephyr, RISC-V, Rust) ofrecen bases como gran alternativa para una generación robusta de soluciones Edge industriales y regulatorias, aún si hoy se sigue investigando sus aplicaciones, la presente opción es la más esperanzadora.

Se debe promover la adopción de prácticas de diseño seguro a través de buenas prácticas, trazabilidad e interoperabilidad futura con otros sistemas.

REFERENCIAS BIBLIOGRÁFICAS

Blanco, G. (2025). *Diseño de un modelo de inferencia energéticamente eficiente de un Vision Transformer implementado en FPGA para su aplicación en sistemas embebidos*. [Tesis de Pregrado, Instituto Tecnológico de Costa Rica]. https://repositoriotec.tec.ac.cr/bitstream/handle/2238/16382/TF10069_BIB314353_Gabriel_Eduardo_Blanco-Mora.pdf?sequence=1&isAllowed=y

Burbano, J., & Estévez, M. (2025). *Implementación de arquitecturas de despliegue remoto de firmware para optimizar la operatividad de redes de sensores inalámbricos*. [Tesis de Pregrado, Universidad del Azuay]. http://dspace.uazuay.edu.ec/handle/datos/15723

Cabrera, L., & Orozco, K. Caracterización de rendimiento computacional en plataformas embebidas para aplicaciones de Edge AI en modelos de detección de personas. *Tecnología en Marcha, 38* (4), 191-201. https://doi.org/10.18845/tm.v38i4.7754

Calandín, A. (2023). *Recuento de personas mediante cámaras de vídeo y ML en el Edge* [Tesis de Maestría, Universitat Politècnica de València]. https://riunet.upv.es/handle/10251/199177

Cantos, A., & Loor, J. (2025). *Implementación de un sistema de clasificación de banano basado en el estado de maduración*. [Tesis de Pregrado, Universidad Politécnica Salesiana]. http://dspace.ups.edu.ec/handle/123456789/30353

de Prado, M., Rusci, M., Capotondi, A., Donze, R., Benini, L., & Pazos, N. (2021). Robustifying the Deployment of TinyML Models for Autonomous Mini-Vehicles. *Sensors, 21*, 1339. https://doi.org/ 10.3390/s21041339

Fanariotis, A., Orphanoudakis, T., Kotrotsios, K., Fotopoulos, V., Keramidas, G., & Karkazis, P. (2025). Power Efficient Machine Learning Models Deployment on Edge IoT Devices. *Sensors, 23*(3), 1595; https://doi.org/10.3390/s23031595

Flores, L., & Sandoval, J. (2025). Ciberseguridad con IA y métricas energéticas. *Azcatl, 4*, 32-37. doi: 10.24275/aZC2025a006

Giannakopoulos, P., van Knippenberg, B., Chandra, K., Calabretta, N., & Exarchakos, G. (2025). Key metrics for monitoring performance variability in edge computing applications. *Journal on Wireless Communications and Networking, 2025* (38). https://doi.org/10.1186/s13638-025-02469-6

Haro, A. (2025). *Detección de defectos en objetos en movimiento mediante redes neuronales convolucionales con optimizaciones específicas para hardware NVIDIA*. [Tesis de Pregrado, Universitat Politècnica de València]. https://riunet.upv.es/handle/10251/198954

Heydari, S., & Mahmoud, Q. (2025). Tiny Machine

Learning and On-Device Inference: A Survey of Applications, Challenges, and Future Directions. *Sensors, 25*(10), 3191; https://doi.org/10.3390/s25103191

Josphineleela, R., Sam, G., Ramesh, T., & Balamurugan, K. (2025). Optimized multiple object tracking with conformalized graph neural network and narwhal optimizer for embedded system IoT and mobile edge computing. *Ain Shams Engineering Journal, 16,* 103581. https://doi.org/10.1016/j.asej.2025.103581

Kang, M., Lee, Y., & Park, M. (2020). Energy Efficiency of Machine Learning in Embedded Systems Using Neuromorphic Hardware. *Electronics, 9*, 1069. doi:10.3390/electronics9071069

Karim, M., Goni, M., Nahiduzzaman, M., Ahsan, M., Haider, J., & Kowalski, M. (2024). Enhancing agriculture through real-time grape leaf disease classification via an edge device with a lightweight CNN architecture and Grad-CAM. *Science Reports, 14*(1), 16022. doi: 10.1038/s41598-024-66989-9

Lacalle, I. (2022). *Diseño y desarrollo de una arquitectura de Internet de las Cosas de nueva generación orientada al cálculo y predicción de índices compuestos aplicada en entornos reales.* [Tesis de Doctorado, Universitat Politècnica de València]. Doi: 10.4995/Thesis/10251/190634

López, D. (2025). *Estudio e implementación de redes neuronales de bajo consumo sobre RISC-V para aplicaciones espaciales*. [Tesis de Pregrado, Universidad de Complutense de Madrid]. https://hdl.handle.net/20.500.14352/124200

Page, M., McKenzie, J., Bossuyt, P., Boutron, I., Hoffmann, T., Mulrow, C., ShamseerI, L., Tetzlaff, J., Akl, E., Brennan, S., Chou, R., Glanville, J., Grimshaw, J., Hro´bjartsson, A., Lalu, M., Li, T., Loder, E., Mayo, E., McDonald, S., Moher, D. (2021) The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine 18*(3), e1003583. https://doi.org/10.1371/journal.pmed.1003583

Perticari, J. (2025*). Implementación de los modelos de redes neuronales convolucionales Yolo en un dispositivo AVNET ULTRA 96 V2 mediante VITIS-AI*. [Universidad Miguel Hernández de Elche]. http://dspace.umh.es/bitstream/11000/36047/1/TFG-Perticari%20Ventura%2C%20Joaqu%C3%ADn.pdf

Pozo, D., Monge, C., & Recalde, P. (2025). Optimización de modelos de lenguaje grande: Una aproximación con Quantization, ONNX y Pruning para aplicaciones en tiempo real. *Revista Ingeniería e Innovación del Futuro, 4*(2), 130–146. https://doi.org/10.62465/riif.v4n2.2025.177

Puerta, G. (2025). Rendimiento y costo de modelos de aprendizaje automático para la detección de crisis epilépticas en una plataforma adaptativa de salud electrónica basada en IoT sobre nodos Edge y Fog. *Revista Colombiana de Computación*, *26*(1). https://doi.org/10.29375/25392115.5480

Ramírez, M. (2025). *Prototipo de órtesis con estimulación muscular controlado por comandos de voz*. [Tesis de Maestría, Instituto Tecnológico de Chilpancingo]. https://rinacional.tecnm.mx/bitstream/TecNM/10923/1/TESIS%20MARIA%20ISABEL.pdf

Romero, L. (2025). *Ejecución de Algoritmos de Inteligencia Artificial en Sistemas de Tiempo Real.*[Tesis de Maestría, Universitat Politècnica de Catalunya]. https://upcommons.upc.edu/server/api/core/bitstreams/c8652cd2-6aad-4375-a907-0cf3cf63eb37/content

Saha, S., Sandha, S., & Srivastava, M. (2022). Machine Learning for Microcontroller-Class Hardware: A Review. *Institute of Electrical and Electronics Engineers Sensors Journal, 22*(22):21362-21390. doi: 10.1109/jsen.2022.3210773

Sánchez, J. (2024). *Comparativas de plataformas software para TinyML* [Tesis de Maestría, Universitat Politècnica de València]. https://riunet.upv.es/server/api/core/bitstreams/cbac7ab1-a167-48bb-8467-a7b9c8287701/content

Senoo, E., Anggraini, L., Kumi, J., Karolina, L, Akansah, E., Sulyman, H., Mendonça, I., & Aritsugi, M. (2024). IoT Solutions with Artificial Intelligence Technologies for Precision Agriculture: Definitions, Applications, Challenges, and Opportunities. *Electronics, 13,* 1894. https://doi.org/10.3390/ electronics13101894

Shakeel, M., Sharatchandra, B., Javed, A., Harkin, J., & Finlay, D. (2025). Toward TinyDPFL systems for real-time cardiac healthcare: Trends, challenges, and system-level perspectives on AI algorithms, hardware, and edge intelligence. *Journal of Systems Architecture, 168,* 103587. https://doi.org/10.1016/j.sysarc.2025.103587

Varghese, B., Wang, N., Bermbach, D., Hong, C., De Lara, E., Shi, W., & Stewart, C. (2021). A Survey on Edge Performance Benchmarking. *ACM Computing Surveys, 54*(3), 1-33. https://doi.org/10.1145/3444692

Vasilev, V., Shterev, V., & Nenova, M. (2025). Optimizing Activation Function for Parameter Reduction in CNNs on CIFAR-10 and CINIC-10. *Applied Sciences, 15*, 4292. https:// doi.org/10.3390/app15084292

Villota, J. (2025). *Diseño e implementación de solución de Deep Learning sobre un sistema embebido, para la selección automática de mango "Tommy Atkins"*. [Tesis de Maestría, Escuela Colombiana de Ingeniería]. https://repositorio.escuelaing.edu.co/bitstreams/14018709-5511-4932-8f6e-49d0e427f8f0/download

Wang, T., Guo, J., Zhang, B., Yang, G., & Li, D. (2025). Deploying AI on Edge: Advancement and Challenges in Edge Intelligence. *Mathematics, 13*, 1878. https://doi.org/10.3390/ math13111878

Xu, D., Li, T., Li, Y., Su, X., Tarkoma, S., Jiang, T., Crowcroft, J., & Hui, P. (2021). Edge Intelligence: Empowering Intelligence to the Edge of Network. *Proceedings of the IEEE, 109*, (11), 1778-1837. DOI: 10.1109/JPROC.2021.3119950

Yan, Z., Zhang, B., & Wang, D. (2024). An FPGA-Based YOLOv5 Accelerator for Real-Time Industrial Vision Applications. *Micromachines*, *15*, 1164. https://doi.org/10.3390/ mi15091164