CAPÍTULO 4

Inteligencia artificial en el borde para salud digital sobre redes de telecomunicaciones con conectividad limitada

Edge artificial intelligence for digital health over telecommunications networks with constrained connectivity

1 Leandro Antonio Pazmiño Ortiz
Máster en Ciberseguridad / Master of Telecommunications Engineering
Escuela Politécnica Nacional / Escuela de Formación de Tecnólogos, Ecuador
leandro.pazmino@epn.edu.ec
https://orcid.org/0000-0002-9241-8409

DOI https://doi.org/10.64644/9789907956313_4 RECIBIDO 15/09/2026 | ACEPTADO 14/10/2026

RESUMEN

La salud digital basada en dispositivos conectados puede ampliar la monitorización y el apoyo clínico más allá del hospital, pero los enfoques dependientes de la nube pierden resiliencia cuando la conectividad es intermitente, costosa o de baja capacidad. Este capítulo presenta una revisión sistematizada de la literatura sobre inteligencia artificial en el borde (Edge AI) aplicada a la salud digital, con un periodo de cobertura bibliográfica comprendido entre el 1 de enero de 2020 y el 9 de septiembre de 2026 y una búsqueda ejecutada y auditada el 9 de septiembre de 2026 en PubMed y OpenAlex, reportada conforme a PRISMA 2020 y PRISMA-S. Se identificaron 1 031 registros (296 en PubMed y 735 en OpenAlex). Un identificador de PubMed no pudo recuperarse en la exportación, por lo que el conjunto auditable previo a la deduplicación quedó en 1 030 registros; tras eliminar 140 duplicados por DOI y 11 tipos de publicación manifiestamente no elegibles, se cribaron 879 registros por título y resumen. Se buscaron 248 informes primarios a texto completo, de los cuales se recuperaron 129; la evaluación a texto completo produjo un corpus primario de 86 estudios con ejecución demostrada de IA en el borde y un corpus secundario de 29 revisiones. La evaluación metodológica se realizó por diseño mediante una lista estructurada de validez técnica (n = 71), QUADAS-2 (n = 10) y PROBAST (n = 5); además, las cinco revisiones sistemáticas del corpus secundario fueron consideradas para AMSTAR 2, con una revisión evaluable a texto completo y cuatro no evaluables por falta de recuperación del texto completo. Predominan los prototipos experimentales (n = 40) y las validaciones técnicas (n = 31) sobre microcontroladores (n = 28), servidores de borde (n = 27) y wearables (n = 17), con aplicaciones concentradas en cardiología (n = 20), movilidad y caídas (n = 14) y neuroseñales (n = 8). La producción se acelera desde 2024 y se concentra en India, Italia y Estados Unidos. La evaluación de calidad revela documentación limitada sobre reproducibilidad, escasez de comparadores explícitos y validación mayoritariamente retrospectiva o de laboratorio. La cuantificación, la poda y las arquitecturas compactas permiten inferencia local incluso en microcontroladores, pero la evidencia longitudinal en poblaciones rurales o desatendidas sigue siendo escasa. Se propone un modelo de referencia

funcional basado en degradación controlada, interoperabilidad, seguridad por diseño y evaluación del ciclo de vida.

Palabras clave: inteligencia artificial en el borde, salud digital, internet de las cosas médicas, aprendizaje automático embebido, conectividad limitada, equidad en salud

ABSTRACT

Digital health based on connected devices can extend monitoring and clinical decision support beyond hospitals, but cloud-dependent approaches lose resilience when connectivity is intermittent, costly, or capacity-constrained. This chapter presents a systematized, fully documented and auditable literature review on edge artificial intelligence (Edge AI) applied to digital health, covering literature published from 1 January 2020 to 9 September 2026, with the search executed and audited on 9 September 2026 in PubMed and OpenAlex and reported following PRISMA 2020 and PRISMA-S. A total of 1,031 records were identified (296 in PubMed and 735 in OpenAlex). One PubMed identifier could not be retrieved from the export, leaving 1,030 auditable records before deduplication; after removing 140 DOI duplicates and 11 manifestly ineligible publication types, 879 records were screened by title and abstract. Full texts were sought for 248 primary reports, of which 129 were retrieved; full-text assessment yielded a primary corpus of 86 studies with demonstrated edge AI execution and a secondary corpus of 29 reviews. Methodological appraisal was design-specific, using a structured technical-validity checklist (n = 71), QUADAS-2 (n = 10), and PROBAST (n = 5); the five systematic reviews in the secondary corpus were also considered for AMSTAR 2, with one review assessable in full text and four non-assessable because the full text could not be retrieved. Experimental prototypes (n = 40) and technical validations (n = 31) predominate, targeting microcontrollers (n = 28), edge servers (n = 27), and wearables (n = 17), with applications concentrated in cardiology (n = 20), mobility and falls (n = 14), and neural signals (n = 8). Output accelerates from 2024 and concentrates in India, Italy, and the United States. Quality appraisal reveals sparse reproducibility reporting, limited explicit baselines, and mostly retrospective or laboratory validation. Quantization, pruning, and compact architectures enable on-device inference even on microcontrollers, but longitudinal evidence in rural or underserved populations remains scarce. A functional reference model is proposed, based on controlled degradation, interoperability, security by design, and lifecycle evaluation.

Keywords: edge artificial intelligence, digital health, internet of medical things, embedded machine learning, constrained connectivity, health equity

1. Introducción

La salud digital contemporánea se apoya cada vez más en biosensores, dispositivos portátiles, teléfonos inteligentes, sistemas de monitorización remota y plataformas del Internet de las Cosas Médicas (IoMT). Estos sistemas producen señales fisiológicas continuas —electrocardiografía (ECG), fotopletismografía (PPG), movimiento, temperatura, saturación de oxígeno u otros indicadores— cuyo valor depende de transformar datos en información accionable con una latencia compatible con la necesidad clínica. Durante la primera etapa de esta transformación, el patrón dominante consistió en captar localmente y transmitir datos hacia centros de datos para almacenamiento y analítica. Ese modelo conserva ventajas para entrenamiento, análisis longitudinal y coordinación a gran escala, pero su dependencia de un enlace estable se vuelve una

limitación estructural cuando la red es intermitente, cara o insuficiente (Pereira et al., 2024; Rocha et al., 2024).

Esta limitación es especialmente relevante cuando la digitalización pretende ampliar el acceso. La cobertura o el uso de Internet a escala nacional no equivalen a disponibilidad continua, baja latencia, asequibilidad ni calidad de servicio en el punto de atención. En Ecuador, la ENEMDU de julio de 2025 reportó que el 80,1 % de las personas de cinco años o más había utilizado Internet en los doce meses anteriores; la proporción fue 85,3 % en áreas urbanas y 68,5 % en áreas rurales (Instituto Nacional de Estadística y Censos, 2025).

Estas cifras describen uso, no capacidad para sostener telemedicina síncrona o transmisión permanente de bioseñales. Un estudio localizado en hospitales generales de Imbabura encontró diferencias marcadas entre establecimientos urbanos y rurales: la dimensión de infraestructura y conectividad alcanzó 60 % en los primeros y 15 % en los segundos, mientras que el índice de viabilidad para telemedicina fue 62,5 % frente a 18,75 %, respectivamente (Armas Cárdenas et al., 2026). El estudio no puede extrapolarse a todo el país, pero ilustra cómo la disponibilidad nominal de Internet puede coexistir con restricciones operativas severas.

La inteligencia artificial en el borde responde a este problema al desplazar una parte de la inferencia desde la nube hacia el punto donde se generan los datos. En sentido amplio, Edge AI abarca un continuo de distribución del cómputo que se extiende desde dispositivos finales y microcontroladores hasta teléfonos inteligentes, pasarelas, nodos fog, infraestructura de computación de borde de acceso múltiple (Multi-access Edge Computing, MEC) y servicios en la nube.

TinyML ocupa el extremo con mayores restricciones de memoria, capacidad de cómputo y energía, donde modelos compactos pueden ejecutarse directamente sobre microcontroladores (Keivanimehr & Akbari, 2025; Pazmiño-Ortiz et al., 2026; Rocha et al., 2024). Esta distribución no elimina a la nube. En arquitecturas maduras, el entrenamiento intensivo, la supervisión de flotas, el análisis poblacional y algunas tareas de actualización permanecen en niveles superiores, mientras que el borde asume funciones sensibles a latencia, disponibilidad, privacidad o ancho de banda.

La promesa técnica es clara: si una señal se interpreta localmente, el sistema puede transmitir una alerta, una característica o un resumen en lugar del flujo crudo. Esta reducción del volumen de datos es especialmente valiosa en redes de baja tasa y gran alcance. Sin embargo, el argumento debe formularse con precisión.

LoRaWAN, por ejemplo, no impone un límite universal de 51 bytes; su presupuesto de carga útil depende de la región y de la data rate definidos por los parámetros regionales (LoRa Alliance, 2025). Aun así, las restricciones de tasa, tiempo en aire y carga útil hacen que la transmisión continua de señales biomédicas de alta resolución sea poco apropiada en numerosos modos de largo alcance. El valor de Edge AI reside, por tanto, en convertir un problema de streaming continuo en un problema de telemetría selectiva.

El segundo desafío es computacional. La literatura de TinyML ha demostrado que la cuantificación, la poda, la destilación de conocimiento y la selección de arquitecturas eficientes

pueden reducir de forma sustancial el tamaño y el costo de inferencia (Keivanimehr & Akbari, 2025; Yahyati et al., 2025). Pero la evidencia tampoco permite fijar un consumo energético universal. Existen prototipos con cifras muy bajas a nivel de modelo o núcleo de inferencia y otros sistemas completos que consumen órdenes de magnitud más debido al sensor, pantalla, radio, acondicionamiento analógico y plataforma de cómputo.

Un clasificador de arritmias de 2026 ejecutado en Arduino UNO, por ejemplo, consiguió 97,6 % de exactitud de benchmark, menos de 24 KB de memoria y aproximadamente 200 ms por latido (Zambrano-de la Torre et al., 2026), mientras que otros trabajos reportan perfiles de energía muy diferentes (Hizem et al., 2025). Comparar Edge AI exige especificar qué se mide —modelo, microcontrolador (microcontroller unit, MCU) activo o sistema completo— y bajo qué ciclo de trabajo.

La descentralización también modifica la superficie de riesgo. La permanencia local de los datos puede reducir la exposición de bioseñales crudas en tránsito, pero los dispositivos periféricos están físicamente más expuestos, deben recibir actualizaciones seguras y pueden ser objeto de extracción de claves, manipulación de firmware, ataques adversarios o inferencia sobre parámetros del modelo (Li et al., 2024; Messinis et al., 2024; Naghib et al., 2025; Shammar et al., 2025).

Del mismo modo, el aprendizaje federado no equivale automáticamente a privacidad: las actualizaciones de modelo pueden filtrar información y la comunicación iterativa puede ser prohibitiva en redes limitadas (Li et al., 2025; Mohammadi et al., 2026). La adopción de criptografía ligera estandarizada, como Ascon en NIST SP 800-232, mejora las opciones para dispositivos restringidos, pero su ventaja concreta depende de la implementación y no debe reducirse a porcentajes energéticos universales (Turan et al., 2025).

Además, la interoperabilidad es una condición de integración clínica. Los dispositivos pueden intercambiar información mediante estándares de la familia IEEE 11073 para Service-oriented Device Connectivity (SDC) en entornos de punto de atención, mientras que HL7 FHIR proporciona recursos y formatos interoperables para el intercambio de información sanitaria (HL7 International, 2023; IEEE & ISO, 2024). Una pasarela de borde puede actuar como traductor semántico, pero esa función requiere perfiles, terminologías, mapeo de identificadores, seguridad y validación; no basta con convertir un objeto local a JSON. La norma IEEE/ISO 11073-10700-2024, de hecho, advierte que implementar el protocolo SDC no demuestra por sí mismo la seguridad clínica, la efectividad ni la ciberseguridad de las funciones conjuntas (IEEE & ISO, 2024).

El vacío más importante es conceptual y empírico: resiliencia técnica no es sinónimo de inclusión sanitaria. Poder inferir sin conexión elimina una barrera, pero no resuelve por sí solo el costo total de propiedad, la capacitación, el mantenimiento, la accesibilidad para personas con discapacidad, la alfabetización digital, el sesgo demográfico, la gobernanza de datos ni la sostenibilidad. Las guías de OMS y UIT sobre telemedicina accesible sitúan la accesibilidad y la inclusión como propiedades de diseño del servicio completo, no únicamente de su conectividad (World Health Organization & International Telecommunication Union, 2024). De forma análoga, el marco regulatorio de la OMS para IA en salud exige considerar el uso previsto, la validación externa, la calidad de los datos, la supervisión y el ciclo de vida (World Health Organization, 2023).

A partir del vacío identificado, el capítulo adopta un enfoque interdisciplinario centrado en tres ejes: trazabilidad metodológica, precisión de las afirmaciones técnicas y separación entre factibilidad de ingeniería y resultados de equidad. El objetivo general es analizar cómo Edge AI puede contribuir a sistemas de salud digital resilientes en entornos con conectividad limitada, identificando arquitecturas, mecanismos de optimización, aplicaciones, beneficios, riesgos y vacíos de implementación. La revisión se estructura mediante seis preguntas de investigación (RQ1–RQ6) que conectan directamente la evidencia técnica con la seguridad, la interoperabilidad y la inclusión.

2. Metodología de la revisión

El capítulo se formuló como una revisión sistematizada con síntesis narrativa comparativa. PRISMA 2020 se utilizó como marco de transparencia para estructurar la identificación, selección y síntesis de la evidencia, mientras que PRISMA-S orientó la documentación de las estrategias de búsqueda (Page et al., 2021; Rethlefsen et al., 2021).

A diferencia de versiones preliminares de este capítulo, en las que no se conservaba un registro auditable de la ejecución de la búsqueda, esta versión documenta íntegramente el proceso ejecutado el 9 de septiembre de 2026: ecuaciones por base de datos, exportaciones, deduplicación, decisiones individuales de cribado y conteos por fase. Los archivos de auditoría (consultas, registros exportados, bitácora de decisiones y libro de trabajo de evaluación a texto completo) se conservan junto con los materiales del capítulo y permiten verificar cada cifra reportada.

2.1. Preguntas de investigación

Las preguntas se definieron para cubrir el modelo dispositivo-borde-niebla-nube (Device-Edge-Fog-Cloud), las restricciones de recursos y los resultados clínicos y sociales, sin convertir el corpus en una revisión genérica de IoT o de IA médica. La Tabla 1 presenta las seis preguntas de investigación y el tipo de evidencia requerido para responderlas.

Tabla 1

Preguntas de investigación y evidencia analítica asociada

Código	Pregunta de investigación	Evidencia buscada
RQ1	¿Qué arquitecturas de Edge AI y tecnologías habilitadoras se utilizan en salud digital?	Ubicación de inferencia, hardware, frameworks, gateways, fog, MEC
RQ2	¿Qué enfoques permiten operar modelos de IA bajo restricciones de cómputo, memoria y energía?	Cuantificación, poda, destilación, NAS, inferencia activada por eventos
RQ3	¿Qué aplicaciones sanitarias presentan mayor madurez tecnológica y de validación?	Cardiovascular, PPG, glucosa, caídas, neuroseñales, imagen, monitorización
RQ4	¿Qué beneficios y compromisos se reportan en latencia, ancho de banda y disponibilidad?	Comparaciones extremo a extremo, costo de comunicación, operación sin conexión
RQ5	¿Qué retos de ciberseguridad, privacidad, confianza e interoperabilidad se abordan?	Criptografía ligera, actualizaciones, canales laterales, FHIR, IEEE 11073
RQ6	¿Qué evidencia empírica existe sobre la contribución a la inclusión y la equidad?	Despliegues de campo, accesibilidad, costo, resultados de salud

Nota. RQ = research question; NAS = neural architecture search; MEC = multi-access edge computing. Elaboración propia.

2.2. Alcance temporal, unidad de análisis y fuentes

El periodo de interés abarca del 1 de enero de 2020 al 9 de septiembre de 2026, fecha de ejecución documentada de la búsqueda. La unidad de análisis del corpus científico primario es una publicación revisada por pares —artículo original o trabajo de congreso completo— que demuestra, con métodos y resultados suficientes, la ejecución total o parcial de una función de IA en un destino de borde (microcontrolador, wearable, teléfono inteligente, gateway, nodo fog, servidor de borde o infraestructura MEC), con una aplicación sanitaria explícita. Se aceptaron publicaciones en inglés o español.

Las fuentes de búsqueda ejecutadas fueron PubMed/MEDLINE y OpenAlex, consultadas el 9 de septiembre de 2026. Se intentó además una consulta a IEEE Xplore; el acceso no devolvió resultados auditables (registros exportables con metadatos verificables), por lo que IEEE Xplore no se incluye entre las fuentes de esta versión. Esta limitación se declara explícitamente y se retoma en la sección 4.7. Las revisiones, surveys y mapeos sistemáticos recuperados durante el cribado se mantuvieron como evidencia secundaria para cartografiar el campo, mientras que normas, documentos regulatorios y estadísticas oficiales se utilizaron únicamente como fuentes contextuales. Esta separación evita mezclar rendimiento experimental con recomendaciones normativas o indicadores de conectividad.

Para preservar la validez de la síntesis se distinguen tres estratos: corpus científico primario, síntesis científica secundaria y fuentes contextuales. Las afirmaciones sobre desempeño, consumo, latencia, memoria, validación o aplicaciones se sustentan prioritariamente en estudios primarios;

las revisiones se emplean para triangulación y contextualización; y las fuentes oficiales fundamentan regulación, estándares, conectividad, accesibilidad y sostenibilidad. De forma complementaria, se utilizaron estudios técnicos adyacentes procedentes de dominios no sanitarios cuando aportaban evidencia directamente transferible a componentes de Edge AI relevantes para el Internet de las Cosas Médicas (IoMT), por ejemplo, investigaciones sobre criptografía ligera o seguridad de microcontroladores. Estas publicaciones se utilizaron exclusivamente para caracterizar mecanismos técnicos y no se contabilizaron como parte del corpus primario (véase la Tabla 2).

Tabla 2

Clasificación de fuentes y reglas de uso en la síntesis

Conjunto	Contenido	N	Uso en la síntesis
Corpus primario	Estudios con ejecución demostrada de IA en el borde y aplicación sanitaria	86	Base de las afirmaciones sobre desempeño, recursos y validación
Evidencia secundaria	Revisiones, surveys y mapeos sobre Edge AI en salud	29	Triangulación, cartografía del campo, localización de estudios
Fuentes contextuales	Normas, guías regulatorias, estadísticas oficiales, estudios técnicos adyacentes	—	Regulación, estándares, conectividad; caracterización de mecanismos técnicos

Nota. Las fuentes contextuales no se contabilizan como estudios primarios. Elaboración propia con base en Page et al. (2021) y Rethlefsen et al. (2021).

2.3. Estrategia de búsqueda y selección: ejecución documentada

La búsqueda se ejecutó el 9 de septiembre de 2026 en dos bases de datos. En PubMed/MEDLINE se utilizó la ecuación: (“edge artificial intelligence” OR “edge AI” OR “edge intelligence” OR “on-device AI” OR “embedded machine learning” OR TinyML OR “distributed intelligence”) AND (healthcare OR medical OR biomedical OR “digital health” OR wearable OR “Internet of Medical Things” OR IoMT OR telemedicine OR “remote patient monitoring”), con filtros de fecha de publicación entre el 01/01/2020 y el 09/09/2026 e idioma inglés o español. La interfaz reportó 296 identificadores; la exportación en XML permitió parsear 295 registros (el identificador PMID 38564569 no pudo recuperarse de la exportación).

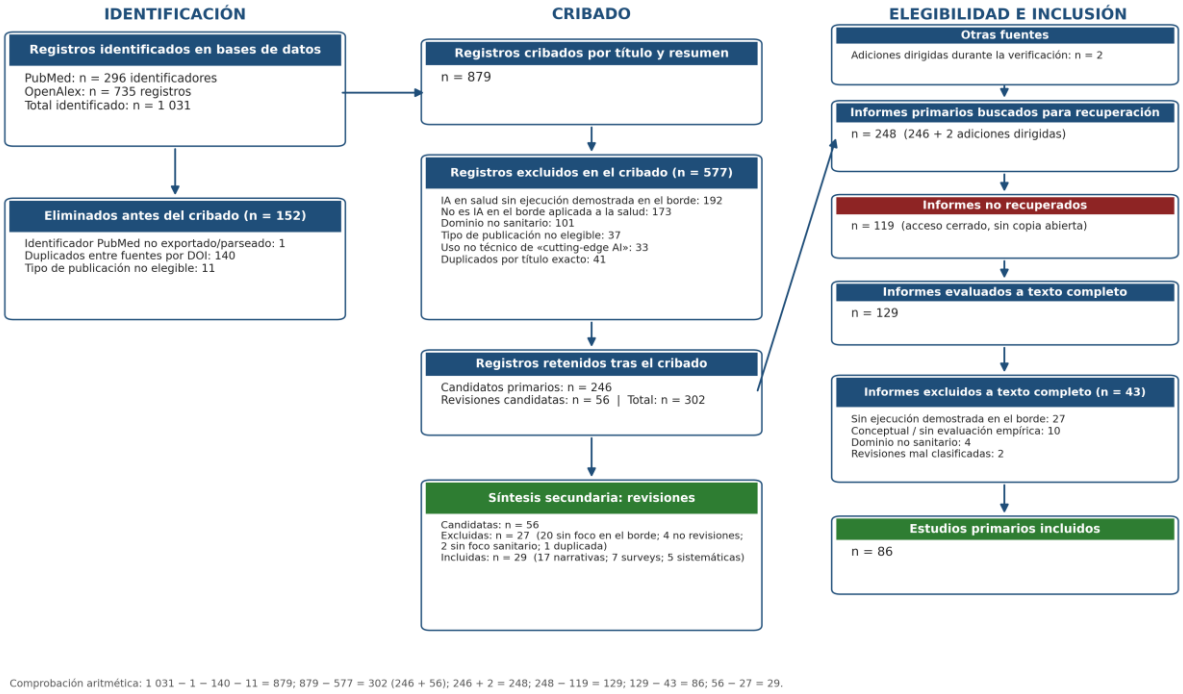
En OpenAlex se ejecutó una búsqueda equivalente por título y resumen con los mismos bloques conceptuales y filtros de fecha, idioma y tipo de documento: el estimador de la interfaz indicó 734 registros, mientras que la paginación por cursor devolvió 735 registros exportados. Esta discrepancia de un registro corresponde al comportamiento del estimador durante la exportación y se declara sin atribuirle una causa no verificada.

Se identificaron 1031 registros en las dos fuentes (296 identificadores en PubMed y 735 registros exportados de OpenAlex). Un identificador de PubMed (PMID 38564569) no pudo recuperarse en la exportación XML, de modo que el corpus auditable previo a la deduplicación

quedó constituido por 1030 registros parseados/exportados (295 de PubMed y 735 de OpenAlex). La deduplicación entre fuentes por DOI identificó 140 duplicados y dejó 890 registros únicos; posteriormente se eliminaron 11 registros por tipo de publicación manifiestamente no elegible, por lo que 879 registros pasaron al cribado por título y resumen. La Figura 1 representa de forma explícita esta aritmética y separa las dos adiciones dirigidas identificadas durante la verificación, mientras que la Tabla 3 resume los conteos por fase.

Figura 1

Diagrama de flujo PRISMA 2020 de la identificación, el cribado y la inclusión de estudios



Nota. PubMed aportó 296 identificadores, de los cuales 295 pudieron recuperarse y parsearse; el PMID 38564569 quedó documentado como identificador no exportado. OpenAlex aportó 735 registros exportados por cursor, aunque su estimador de interfaz mostró 734. Las dos adiciones dirigidas (n = 2) se incorporaron antes de la recuperación a texto completo. Elaboración propia conforme a PRISMA 2020 (Page et al., 2021).

El cribado por título y resumen se aplicó a los 879 registros con decisiones individuales registradas. Los motivos de exclusión fueron: IA en salud sin ejecución demostrada en el borde (n = 192), literatura no correspondiente a IA en el borde aplicada a la salud (n = 173), dominio no sanitario (n = 101), tipo de publicación no elegible (n = 37), uso no técnico de “cutting-edge AI” (n = 33) y duplicados por título exacto (n = 41).

En total, 577 registros fueron excluidos en esta fase. Los 56 candidatos a revisión secundaria fueron adjudicados uno por uno: se conservaron 29 (17 narrativas, 7 surveys y 5 sistemáticas) y se descartaron 27 (20 sin foco en el borde, 4 que no eran revisiones, 2 sin foco sanitario y 1 duplicada). Durante esta adjudicación se detectó un duplicado adicional dentro del conjunto de revisiones, por lo que el total de duplicados por título exacto asciende a 42.

De los 246 candidatos primarios, más dos adiciones dirigidas de estudios elegibles identificados durante la verificación (total 248), se recuperaron 129 informes a texto completo mediante acceso editorial, repositorios abiertos y manuscritos de autor. Los 119 informes no recuperados corresponden a publicaciones de acceso cerrado sin copia abierta localizable; permanecen sin evaluar y se listan en la bitácora de recuperación.

Tabla 3

Flujo PRISMA 2020: conteos por fase de la ejecución documentada

Fase	Descripción	n
Identificación	Identificadores/registros identificados en PubMed	296
Identificación	Registros exportados en OpenAlex	735
Identificación	Total identificado en las fuentes	1 031
Precribado	Identificador PubMed no recuperado en la exportación XML	1
Precribado	Registros auditables parseados o exportados antes de deduplicación	1 030
Precribado	Duplicados eliminados entre fuentes por DOI	140
Precribado	Registros únicos tras deduplicación por DOI	890
Precribado	Registros eliminados automáticamente por tipo de publicación no elegible	11
Cribado	Registros cribados por título y resumen	879
Cribado	Registros excluidos en el cribado	577
Cribado	Candidatos primarios a texto completo	246
Cribado	Revisiones candidatas para síntesis secundaria	56
Otras fuentes	Adiciones dirigidas identificadas durante la verificación	2
Elegibilidad	Informes primarios buscados para recuperación (246 + 2)	248
Elegibilidad	Informes primarios no recuperados	119
Elegibilidad	Informes recuperados y evaluados a texto completo	129
Elegibilidad	Informes excluidos tras evaluación a texto completo	43
Inclusión	Estudios primarios incluidos	86
Síntesis secundaria	Revisiones excluidas tras adjudicación	27
Síntesis secundaria	Revisiones incluidas	29

Nota. La aritmética auditada es: 1 031 − 1 − 140 − 11 = 879 registros cribados; 879 − 577 = 302 registros retenidos (246 candidatos primarios + 56 revisiones); 246 + 2 adiciones dirigidas = 248 informes primarios buscados; 248 − 119 = 129 informes evaluados; 129 − 43 = 86 estudios primarios incluidos; y 56 − 27 = 29 revisiones incluidas como evidencia secundaria. Elaboración propia.

La evaluación a texto completo excluyó 43 informes: 27 por no demostrar ejecución de IA en el borde —18 con cómputo de escritorio o en la nube sin componente de borde, y 9 reclasificados tras una readjudicación estricta por basarse en simulación (OMNeT++, Cooja o datos sintéticos) o en un procesador sustituto sin ejecución en el destino de borde declarado—, 10 por carácter conceptual sin evaluación empírica, 4 por dominio no sanitario y 2 revisiones mal clasificadas. Los 9 estudios

reclasificados se conservan como fuentes contextuales cuando aportan evidencia técnica transferible. El corpus primario final comprende 86 estudios.

2.4. Criterios de inclusión y exclusión

La Tabla 4 presenta los criterios operativos aplicados en las dos etapas de elegibilidad: cribado por título y resumen, y evaluación a texto completo. El criterio determinante del corpus primario fue la demostración explícita de ejecución de una función de IA en un destino de borde; los estudios que solo simulaban el despliegue en estaciones de trabajo sin un objetivo de borde identificable fueron excluidos del corpus primario y, cuando aportaban evidencia técnica transferible, considerados como fuentes contextuales.

Tabla 4

Criterios operativos de inclusión y exclusión

Dimensión	Incluir	Excluir
Dominio	Aplicación clínica, sanitaria, biomédica, telemedicina o monitorización de la salud	IoT industrial, ciudades inteligentes, agricultura, transporte u otros dominios no sanitarios
IA en el borde	Inferencia o aprendizaje ejecutado total o parcialmente en dispositivo, wearable, teléfono, gateway, fog, servidor de borde o MEC	Dispositivo usado solo como colector pasivo; IA ejecutada únicamente en la nube o simulada en estación de trabajo sin objetivo de borde
Evidencia empírica	Evaluación con métodos, datos y métricas reportadas de forma verificable	Trabajos puramente conceptuales, editoriales o sin evaluación empírica
Tipo de publicación	Artículos originales y trabajos de congreso completos revisados por pares	Resúmenes cortos, editoriales, preprints, tesis, capítulos sin revisión identificable
Idioma y fecha	Inglés o español; publicación entre 2020-01-01 y 2026-09-09	Otros idiomas; publicaciones fuera del periodo
Versiones múltiples	Versión de revista ampliada cuando duplica sustancialmente al congreso	Duplicados bibliográficos o versiones preliminares superadas

Nota. Los códigos operativos de exclusión (E1–E8) y su correspondencia con estas dimensiones se documentan en el Apéndice B. Elaboración propia.

2.5. Extracción de datos y evaluación de la calidad metodológica

La ficha de extracción registró: objetivo clínico y uso previsto; población y procedencia de los datos; modalidad de señal; arquitectura y ubicación de la inferencia; hardware y entorno de ejecución; tamaño del modelo, memoria y precisión numérica; latencia de inferencia; consumo o energía por inferencia y frontera de medición; protocolo de comunicación y volumen transmitido; comportamiento ante desconexión; mecanismos de ciberseguridad; interoperabilidad; diseño de validación; comparador; métricas clínicas; y resultados de equidad o accesibilidad. Para prevenir comparaciones espurias, las métricas de recursos se comparan solo cuando comparten denominador

y frontera de medición (por ejemplo, microcontrolador activo frente a microcontrolador activo, o energía del sistema completo frente a sistema completo).

La madurez se clasificó mediante dos ejes independientes. El primero es el Technology Readiness Level (TRL), escala de nueve niveles desde principios básicos (TRL 1) hasta sistemas probados en operación (TRL 9) (National Aeronautics and Space Administration, 2023), con una adaptación operacional para Edge AI sanitario: TRL 1–2, concepto o aplicación formulada; TRL 3, prueba de concepto analítica o experimental; TRL 4, prototipo integrado validado en laboratorio; TRL 5, validación en un entorno relevante que aproxima el flujo clínico; TRL 6, demostración de un prototipo funcional en entorno relevante; TRL 7, demostración del sistema en un entorno operativo real; TRL 8, sistema completo cualificado para su uso previsto; y TRL 9, sistema utilizado de forma sostenida en operación con evidencia posdespliegue. La asignación se realiza al nivel más alto explícitamente sustentado por el estudio; el TRL no se interpreta como autorización regulatoria ni como evidencia de beneficio clínico.

El segundo eje es el nivel de madurez de validación (NMV), escala analítica definida para esta revisión: NMV1, validación retrospectiva interna o benchmark en un conjunto de datos público; NMV2, validación independiente, temporal o externa sin diseño prospectivo de impacto; NMV3, estudio prospectivo o piloto con usuarios de la población objetivo; NMV4, evaluación clínica prospectiva comparativa o multicéntrica; y NMV5, evidencia de implementación longitudinal o posdespliegue con resultados clínicos, de acceso, costos, seguridad o equidad. Esta jerarquía es una herramienta analítica del capítulo y no sustituye instrumentos formales de riesgo de sesgo ni sistemas como GRADE. Cuando la información publicada no permite una clasificación inequívoca, se registra como no reportado.

La evaluación de la calidad metodológica y, cuando correspondía al diseño, del riesgo de sesgo se aplicó a los 86 estudios primarios mediante instrumentos diferenciados. Los 10 estudios de precisión diagnóstica se evaluaron con QUADAS-2 (Whiting et al., 2011) en los dominios de selección de pacientes, prueba índice, estándar de referencia y flujo/tiempos. Los 5 modelos de predicción se evaluaron con PROBAST (Wolff et al., 2019) en participantes, predictores, desenlace y análisis.

Los 71 estudios restantes —prototipos experimentales y validaciones técnicas— se examinaron con una lista estructurada de validez técnica definida para esta revisión, centrada en la correspondencia entre plataforma y destino de borde, la ejecución real frente a simulación, la separación entre entrenamiento y prueba, la presencia de comparadores, la documentación de la frontera de medición de recursos y la suficiencia del protocolo de evaluación. Esta lista se utiliza como evaluación de calidad y validez técnica, no como sustituto de una herramienta clínica validada de riesgo de sesgo. No se calcularon puntajes agregados; cada dominio o ítem recibió un juicio sustentado y los resultados se detallan por estudio en los archivos de auditoría.

Las 29 revisiones del corpus secundario se clasificaron por tipo (17 narrativas, 7 surveys y 5 sistemáticas) y se utilizaron para triangulación. Las 5 revisiones sistemáticas se sometieron a AMSTAR 2 (Shea et al., 2017): una revisión pudo evaluarse con texto completo y obtuvo confianza global críticamente baja por defectos críticos identificados; las otras cuatro no pudieron

evaluarse formalmente porque el texto completo no fue recuperado, por lo que se registran como "no evaluables" y no se les asigna una categoría de confianza global.

Las revisiones narrativas y los surveys no se calificaron con AMSTAR 2. Tampoco se generó una calificación GRADE global, porque las RQ1–RQ6 abarcan tareas, señales, plataformas y desenlaces heterogéneos sin un efecto común susceptible de síntesis cuantitativa única. La síntesis se realiza por categorías analíticas y los resultados individuales se interpretan condicionados a su plataforma, población y protocolo.

2.6. Trazabilidad PRISMA y comprobaciones de consistencia

Para hacer auditable la selección, se diferencian explícitamente identificadores, registros, informes y estudios. La estrategia y los conteos de identificación se documentan en el Apéndice A; los motivos operativos de exclusión y sus códigos se presentan en el Apéndice B; y el Apéndice C enumera los 86 estudios primarios incluidos. La comprobación aritmética del flujo es completa: 1 031 identificadores/registros identificados menos 1 identificador PubMed no exportado, 140 duplicados por DOI y 11 tipos no elegibles producen 879 registros cribados; tras excluir 577 quedan 302 registros retenidos, correspondientes a 246 candidatos primarios y 56 revisiones.

A los 246 candidatos primarios se añadieron 2 estudios identificados durante la verificación, por lo que se buscaron 248 informes; 119 no se recuperaron, 129 se evaluaron a texto completo, 43 se excluyeron y 86 se incluyeron. En la rama secundaria, 56 revisiones candidatas menos 27 excluidas produjeron 29 revisiones incluidas. Estos valores son los que se reproducen en la Figura 1 y la Tabla 3.

3. Desarrollo de la revisión

El corpus primario comprende 86 estudios con ejecución demostrada de IA en el borde. Por metodología de evaluación, predominan los prototipos experimentales (n = 40) y las validaciones técnicas (n = 31), frente a 10 estudios de precisión diagnóstica y 5 modelos de predicción.

Para la caracterización temporal se auditó el año bibliográfico explícitamente consignado en las 86 referencias del Apéndice C: 3 estudios corresponden a 2020, 1 a 2021, 10 a 2022, 6 a 2023, 14 a 2024, 26 a 2025 y 25 a 2026; una referencia del corpus no consigna año en el registro bibliográfico consolidado y se mantiene como "no determinado" para evitar imputación. Con esta regla reproducible, 51 de 86 estudios (59,3 %) corresponden a 2025–2026, aun cuando 2026 es un año parcial con corte al 9 de septiembre.

La caracterización geográfica disponible muestra mayor representación de India (n = 14), Italia (n = 11), Estados Unidos (n = 10) y China (n = 8), seguidos por Corea del Sur (n = 4), España (n = 3) y México (n = 3). Los conteos de país explícitamente consolidados en la versión recibida suman 76 estudios; para los 10 restantes no se conserva en la tabla de extracción una asignación inequívoca

de país, por lo que se reportan como “no determinable en el registro consolidado” en lugar de imputar una afiliación.

Los destinos de borde más frecuentes son los microcontroladores (n = 28), los servidores de borde (n = 27) y los wearables (n = 17), seguidos por configuraciones con múltiples destinos (n = 7), teléfonos inteligentes (n = 6) y nodos fog (n = 2); estas categorías técnicas pueden solaparse cuando un estudio despliega más de un destino. Por aplicación sanitaria, cardiología —arritmia y ECG/PPG— es el dominio más estudiado (n = 20), seguido por movilidad, rehabilitación y detección de caídas (n = 14), neurología y neuroseñales (n = 8), monitorización general del paciente (n = 6) e imagen médica diagnóstica (n = 5). La Tabla 5 consolida la caracterización solicitada.

La evaluación metodológica y del riesgo de sesgo (Tabla 6) muestra dos patrones. En los 71 prototipos y validaciones técnicas, la ejecución en hardware real y la documentación de recursos están mejor descritas que la reproducibilidad y la caracterización de los datos: 51 estudios reportan medición en el dispositivo y 19 quedaron inciertos; el registro restante no presenta una codificación inequívoca en el resumen consolidado.

Asimismo, 55 reportan métricas de recursos; 17 describen con claridad la muestra o el conjunto de datos y 53 quedaron inciertos, con un registro no determinable; y 28 incluyen una línea base explícita y 42 quedaron inciertos, nuevamente con un registro no determinable. La reproducibilidad, evaluada mediante la disponibilidad sistemática de código y datos, quedó incierta en los 71. En QUADAS-2 (n = 10), la selección de pacientes y el estándar de referencia fueron inciertos en los 10 casos; solo un estudio alcanzó riesgo bajo en prueba índice y flujo/tiempos.

En PROBAST (n = 5), los cuatro dominios quedaron inciertos en los cinco estudios por información insuficiente. En AMSTAR 2, una de las cinco revisiones sistemáticas fue evaluable y obtuvo confianza críticamente baja; las otras cuatro se clasificaron como no evaluables por falta de texto completo. Estos resultados no invalidan el corpus, pero reducen la fuerza con la que las métricas técnicas pueden extrapolarse a eficacia clínica.

Tabla 5

Caracterización del corpus primario (n = 86)

Dimensión	Categoría	n
Año de publicación	2020 / 2021 / 2022 / 2023 / 2024 / 2025 / 2026 / no determinado	3 / 1 / 10 / 6 / 14 / 26 / 25 / 1
Metodología / diseño de evaluación	Prototipo experimental	40
	Validación técnica	31
	Precisión diagnóstica	10
	Modelo de predicción	5

Dimensión	Categoría	n
Destino de borde	Microcontrolador (MCU)	28
	Servidor de borde	27
	Wearable	17
	Múltiple	7
	Teléfono inteligente	6
	Fog	2
Aplicación sanitaria	Cardiología: arritmia y ECG/PPG	20
	Movilidad, rehabilitación y detección de caídas	14
	Otras aplicaciones clínicas y de salud pública	13
	Neurología y neuroseñales (EEG, crisis, BCI)	8
	Monitorización general del paciente / IoMT	6
	Imagen médica diagnóstica	5
	Asistencia a discapacidad visual	3
	Diabetes: estimación de glucosa	3
	Sueño, presión arterial, apoyo a la decisión, salud materna, seguridad del personal (2 c/u)	10
	Dispensación de medicamentos, oftalmología, UCI, bioseñales (1 c/u)	4
País (afiliación)	India / Italia / Estados Unidos / China	14 / 11 / 10 / 8
	Corea del Sur	4
	España / México	3 / 3
	Otros países (1–2 estudios c/u)	23
	No determinable en el registro consolidado	10

Nota. Para país, los conteos explícitamente disponibles suman 76 y los 10 casos no asignables se reportan como “no determinable en el registro consolidado”. Las categorías de destino de borde pueden solaparse. La dimensión “Metodología / diseño de evaluación” resume los principales enfoques del corpus: prototipo experimental, validación técnica, precisión diagnóstica y modelo de predicción. Elaboración propia a partir de la extracción a texto completo y de la auditoría bibliográfica del Apéndice C. El año se tomó del dato explícito de cada referencia; un registro sin año no fue imputado.

Tabla 6

Evaluación metodológica y del riesgo de sesgo: resumen por herramienta

Herramienta	n	Hallazgos principales
Lista estructurada de validez técnica	71	Medición en el dispositivo: sí 51, incierto 19, no determinable en el resumen consolidado 1. Métricas de recursos: sí 55; los 16 restantes no se desagregaron de forma inequívoca en el resumen consolidado. Muestra/datos descritos con claridad: sí 17, incierto 53, no determinable 1. Comparación con línea base: sí 28, incierto 42, no determinable 1. Reproducibilidad (código/datos): incierta en 71. Limitaciones discutidas: sí 18. Instrumento de calidad/validez técnica, no herramienta clínica validada de riesgo de sesgo.
QUADAS-2 (precisión diagnóstica)	10	Selección de pacientes: incierto 10. Prueba índice: bajo 1, incierto 9. Estándar de referencia: incierto 10. Flujo y tiempos: bajo 1, incierto 9. Dominios mayoritariamente no evaluables con certeza por información insuficiente.
PROBAST (modelos de predicción)	5	Participantes, predictores, desenlace y análisis: inciertos en los 5 estudios por información insuficiente del protocolo de validación.
AMSTAR 2 (revisiones sistemáticas)	5	Una revisión evaluable con texto completo: confianza críticamente baja por cuatro defectos críticos. Cuatro revisiones: no evaluables por falta de texto completo; no se asignó categoría de confianza global.

Nota. Los juicios de QUADAS-2 y PROBAST corresponden a riesgo de sesgo por dominio; AMSTAR 2 corresponde a confianza en revisiones sistemáticas. La lista de validez técnica es un instrumento estructurado de calidad metodológica para estudios de ingeniería y no se interpreta como una herramienta clínica validada de riesgo de sesgo. "Incierto" refleja información insuficiente, no necesariamente un defecto confirmado. Elaboración propia.

3.1. Categoría analítica I: arquitecturas y distribución del cómputo de IA

Las revisiones recientes coinciden en que Edge AI sanitario no describe una única arquitectura, sino un continuo de ubicaciones de cómputo (Keivanimehr & Akbari, 2025; Pazmiño-Ortiz et al., 2026; Pereira et al., 2024; Rocha et al., 2024; Yahyati et al., 2025). El extremo de menor capacidad comprende microcontroladores y wearables con inferencia local; teléfonos y gateways pueden agregar datos, ejecutar modelos de mayor complejidad y traducir protocolos; servidores edge o fog concentran cargas de varios dispositivos con baja latencia; y la nube mantiene ventajas para entrenamiento intensivo, gestión de flotas, análisis poblacional y almacenamiento longitudinal. La decisión de colocación debe considerar simultáneamente latencia clínica, disponibilidad del enlace, confidencialidad, energía, memoria, costo de comunicación y requisitos de integración.

La contribución diferencial de Edge AI frente a un enfoque cloud-centric es arquitectónica: desacopla funciones sensibles al tiempo y a la disponibilidad de una conexión WAN continua. Cuando la inferencia ocurre cerca de la fuente, el sistema puede responder sin round-trip a la nube, reducir la exposición de datos crudos en tránsito, transformar streaming continuo en telemetría selectiva y mantener un subconjunto funcional durante interrupciones. Estas ventajas no convierten al borde en sustituto universal de la nube: el procesamiento local introduce límites de memoria y energía, mayor exposición física, complejidad de actualización y necesidad de observabilidad distribuida. Por ello, para conectividad limitada el patrón más robusto suele ser híbrido, con funciones críticas localizadas en el borde y capacidades no críticas o de gran escala en niveles superiores.

En el plano de interoperabilidad, IEEE 11073 SDC y HL7 FHIR ocupan roles distintos y complementarios. IEEE/ISO 11073-10700-2024 define requisitos base para participantes en un sistema de conectividad orientada a servicios en dispositivos médicos de punto de atención (IEEE & ISO, 2024). FHIR R5, por su parte, especifica recursos y representaciones —incluidos JSON y XML— para intercambio de información sanitaria (HL7 International, 2023). Un gateway puede mapear una observación o inferencia local a recursos FHIR, pero la semántica, identificación, perfilado y seguridad deben diseñarse explícitamente. No es correcto asumir que IEEE 11073 "se convierte" automáticamente en FHIR.

La Tabla 7 resume los principales patrones arquitectónicos y sus compromisos de conectividad, latencia y riesgo.

Tabla 7

Patrones arquitectónicos de Edge AI en salud digital

Patrón	Lugar de inferencia	Dependencia de conectividad	Contribución diferencial	Riesgo/limitación
Cloud-centric	Centro de datos remoto	Alta y continua	Entrenamiento, escalabilidad y análisis global; administración centralizada.	Dependencia de red; latencia variable; mayor transporte de datos crudos.
Gateway/Fog	Teléfono, hub clínico o servidor local	LAN local; WAN puede ser intermitente	Agregación local, modelos mayores, traducción de protocolos y continuidad parcial.	Punto local de fallo; energía, mantenimiento y gestión de versiones.
On-device / TinyML	Sensor, wearable o MCU	Puede operar sin WAN	Respuesta inmediata, mínima telemetría, autonomía y localización de datos.	Memoria/energía limitadas; exposición física; actualización y drift.
Híbrida device-edge-cloud	Inferencia distribuida según criticidad	Adaptativa	Combina resiliencia local con supervisión, entrenamiento y análisis global.	Mayor complejidad de orquestación, validación, sincronización y MLOps.

Nota. Elaboración propia a partir de Keivanimehr y Akbari (2025), Pazmiño-Ortiz et al. (2026), Pereira et al. (2024) y Rocha et al. (2024).

3.2. Categoría analítica II: eficiencia bajo restricciones de recursos

La cuantificación es una de las técnicas de optimización más consolidadas en la literatura revisada. La cuantificación de FP32 a INT8 reduce teóricamente a una cuarta parte el almacenamiento requerido por los pesos y puede aprovechar instrucciones enteras eficientes. Sin embargo, la ganancia real depende de activaciones, buffers intermedios, operadores soportados y librerías. En un autoencoder de ECG de 2026, la cuantificación posentrenamiento redujo el modelo de 5,44 KB a 1,36 KB, con una disminución pequeña del F1 (88,67 % a 88,36 %) (Suryanarayanan &

Gajula, 2026; simulación en MATLAB, evidencia contextual). Este resultado demuestra eficiencia de memoria en una tarea concreta; no demuestra por sí solo consumo energético ni validez clínica.

Cuando la cuantificación posterior al entrenamiento (post-training quantization, PTQ) degrada el modelo, el entrenamiento consciente de cuantificación (quantization-aware training, QAT) puede recuperar desempeño al simular durante el entrenamiento los errores de redondeo. No obstante, la superioridad de QAT frente a PTQ no es universal y debe evaluarse por tarea. La poda elimina pesos, parámetros o canales y puede inducir dispersidad (sparsity), mientras que la destilación transfiere conocimiento de un modelo maestro a uno compacto. La búsqueda de arquitecturas y los modelos diseñados desde el inicio para microcontroladores evitan que la compresión sea una corrección tardía (Keivanimehr & Akbari, 2025; Yahyati et al., 2025).

La memoria de activaciones suele ser tan importante como el tamaño de los pesos. Por ello, una práctica de documentación adecuada debe incluir Flash/ROM, SRAM pico, arena de tensores, latencia, ciclos y frecuencia de reloj. El benchmark de Arduino UNO publicado en 2026 es ilustrativo: un 1D-CNN cuantificado para tres clases de arritmia utilizó aproximadamente 21,8 KB de Flash y 1,7 KB de SRAM, con cerca de 200 ms por latido y 97,6 % de exactitud sobre MIT-BIH (Zambrano-de la Torre et al., 2026). Los autores lo interpretan correctamente como prueba de factibilidad; las señales de voluntarios sanos se usaron para verificar el funcionamiento, no para estimar el rendimiento diagnóstico.

La energía requiere aún más cautela. Un artículo de 2025 reportó 0,024 mW para un sistema de detección de anomalías ECG optimizado (Hizem et al., 2025), mientras que prototipos que incorporan sensado, pantalla, radio o plataformas diferentes presentan presupuestos mucho mayores. Estas discrepancias no son necesariamente contradicciones: pueden reflejar límites de medición distintos. El criterio de comparación debe ser energía por inferencia o potencia media del sistema bajo un ciclo de trabajo definido, con instrumentación y condiciones descritas.

Finalmente, la eficiencia no se limita al modelo. La inferencia jerárquica y event-driven puede ejecutar un detector barato permanentemente y activar un modelo mayor solo ante eventos sospechosos. Del mismo modo, el sistema puede cambiar de política según batería, calidad de señal o conectividad. En salud digital, esta capacidad adaptativa es particularmente relevante porque la prioridad no es maximizar throughput sino conservar función segura durante largos periodos.

La Tabla 8 compara las técnicas de optimización más utilizadas y los riesgos que deben controlarse durante el despliegue.

Tabla 8

Técnicas de optimización para dispositivos con recursos restringidos

Técnica	Qué reduce	Beneficio esperado	Riesgo para controlar
Cuantificación PTQ/QAT	Memoria y costo aritmético	Modelos INT8 con menor huella de memoria y posibilidad de cómputo entero eficiente	Pérdida de sensibilidad; operadores no soportados.
Poda/dispersidad (sparsity)	Parámetros y MACs	Menor cómputo si el runtime explota sparsity	Irregularidad; posible degradación de robustez.
Destilación	Complejidad del modelo	Estudiante compacto con comportamiento del maestro	Transferencia de sesgos y pérdida de calibración.
Arquitecturas eficientes / NAS	Complejidad desde el diseño	Mejor ajuste al hardware objetivo	Costo de búsqueda y dependencia de plataforma.
Inferencia event-driven	Duty cycle y radio/cómputo activo	Mayor autonomía	Riesgo de que el detector ligero omita eventos.
Offloading adaptativo	Carga local	Balance entre dispositivo, gateway y nube	Dependencia de estimación de red y políticas seguras.

Nota. PTQ = post-training quantization; QAT = quantization-aware training; NAS = neural architecture search; MAC = multiply–accumulate. Elaboración propia a partir de Hizem et al. (2025), Keivanimehr y Akbari (2025), Suryanarayanan y Gajula (2026; evidencia de simulación), Yahyati et al. (2025) y Zambrano-de la Torre et al. (2026).

3.3. Categoría analítica III: aplicaciones sanitarias y madurez de validación

La monitorización cardiovascular es el dominio con evidencia más visible. Su combinación de señales temporales, necesidad de respuesta rápida y dispositivos portátiles ha impulsado clasificación de arritmias, detección de anomalías y estimación de eventos en hardware restringido (Hizem et al., 2025; Keivanimehr & Akbari, 2025; Suryanarayanan & Gajula, 2026; Zambrano-de la Torre et al., 2026). También aparecen aplicaciones sobre PPG, estimación no invasiva de glucosa, detección de estrés, caídas y análisis de señales multimodales (Ibrar et al., 2026; Zeynali et al., 2025). La diversidad demuestra flexibilidad técnica, pero no implica igual madurez clínica.

Un patrón recurrente es el uso de conjuntos de datos públicos para entrenamiento y evaluación, seguido por el despliegue del modelo en hardware. Ese diseño demuestra factibilidad de despliegue, pero responde de forma incompleta a la cuestión de si el sistema funciona de manera segura y generalizable en el contexto previsto. El estudio de arritmia en Arduino (Zambrano-de la Torre et al., 2026) es ejemplar porque distingue la exactitud obtenida en MIT-BIH de la prueba de funcionamiento con voluntarios. Esa separación debería ser una norma de documentación.

La estimación de glucosa mediante PPG y TinyML ilustra un área de alto interés clínico, pero con desafíos de validez, ya que factores fisiológicos, demográficos y de adquisición pueden alterar la señal (Zeynali et al., 2025). Del mismo modo, los sistemas de detección de caídas o estrés evaluados en conjuntos de datos controlados pueden degradarse al cambiar el dispositivo, la población o las actividades. La madurez de Edge AI sanitario debe valorarse por validación externa,

diversidad de usuarios, condiciones de movimiento, calidad de señal y consecuencias de falsos negativos o positivos, no únicamente por exactitud.

La Tabla 9 reúne estudios representativos para ilustrar la diversidad de aplicaciones y la diferencia entre madurez tecnológica y nivel de evidencia. Los TRL son estimaciones analíticas basadas en lo que cada publicación demuestra explícitamente; no equivalen a aprobación clínica o regulatoria.

Tabla 9

Estudios representativos: madurez tecnológica y de validación

Estudio representativo	Aplicación / plataforma	Resultado técnico reportado	TRL / NMV y cautela interpretativa
Zambrano-de la Torre et al., 2026	Arritmia; 1D-CNN INT8 en Arduino UNO	97,6 % en MIT-BIH; <24 KB; ~200 ms/latido	TRL 4; NMV1. Prototipo integrado en hardware; ausencia de partición interpaciente (riesgo de fuga de datos).
Hizem et al., 2025	Anomalía ECG; edge/TinyML	92,3 % y 0,024 mW reportados	TRL 4; NMV1. Validación técnica; la cifra energética solo es comparable dentro de su frontera de medición.
Zeynali et al., 2025	Glucosa no invasiva con PPG y TinyML	STM32H743IIT6; 6 388 sujetos/35 358 registros + 67 externos; RMSE 19,7 mg/dL; 100 % Clarke A+B; 6,4 s/inferencia	TRL 4–5; NMV2. Validación externa incluida; la generalización depende de la variabilidad fisiológica.
Ibrar et al., 2026	Detección de caídas multimodal; TinyML federado (TinyCFL)	ESP32-CAM, Nano 33 y Arduino UNO; exactitud 98,95 %, AUC 99,92 %	TRL 4; NMV1. Colaboración federada demostrada en hardware; falta validación prospectiva.
Abdullah et al., 2026	Apnea del sueño; transformer evolutivo en Raspberry Pi 4	<2 W; 72 h de operación offline; sincronización HL7/FHIR	TRL 5; NMV2. Operación prolongada sin conexión; métricas casi perfectas con n = 17 en prueba (cautela).
Wakili & Bakkali, 2026	Arquitectura móvil offline; smartphones de entrada	Random Forest 98,25 %, <100 ms (repositorio público); usabilidad exploratoria	Texto completo no recuperado: sin evaluación formal. Cifras procedentes del repositorio de código, no del artículo.

Nota. Los resultados corresponden a las condiciones experimentales de cada estudio y no son directamente comparables. Suryanarayanan y Gajula (2026) se reclasificó como evidencia contextual por basarse en simulación en MATLAB sin ejecución en el destino de borde. TRL = technology readiness level; NMV = nivel de madurez de validación definido para esta revisión. Elaboración propia.

3.4. Categoría analítica IV: conectividad, latencia, ancho de banda y disponibilidad

Uno de los beneficios de Edge AI mejor sustentados conceptual y experimentalmente en la literatura revisada es la reducción del volumen de datos que debe transportarse hacia

infraestructuras remotas. En monitorización continua, el costo de comunicación suele superar el de una inferencia local ligera, especialmente cuando la radio debe permanecer activa o cuando el enlace se tarifa por volumen o energía. El procesamiento local permite transmitir estados, características, scores o fragmentos seleccionados, manteniendo en el dispositivo el flujo de alta frecuencia.

LoRaWAN constituye un caso ilustrativo. La especificación regional vigente RP002-1.0.5 define parámetros que varían por región y data rate (LoRa Alliance, 2025); por ello, 51 bytes es un valor posible en ciertos modos de baja tasa, no un límite universal. La conclusión de ingeniería permanece: los modos de gran alcance operan con presupuestos de datos pequeños y restricciones de tiempo en aire, por lo que el streaming de bioseñales no es el uso natural de LPWAN. Edge AI puede transformar la aplicación hacia mensajes esporádicos de alarma, calidad, tendencia o resumen.

La arquitectura offline-first incorpora persistencia local y permite preservar funciones previamente definidas durante periodos sin conectividad WAN. Ante pérdida de WAN, el sistema debe conservar la decisión, el contexto mínimo, la marca temporal, la calidad de señal y la proveniencia necesaria para una reconciliación posterior. El patrón store-and-forward es más apropiado que una dependencia de sincronización continua. Cuando regresa el enlace, el gateway puede sincronizar recursos interoperables y aplicar políticas de prioridad, evitando que una cola histórica desplace una alerta urgente.

La latencia también debe medirse end-to-end. Una inferencia de 10 ms no garantiza una alerta de 10 ms si el sistema espera una ventana de señal de varios segundos, realiza filtrado, acceso a almacenamiento o actualización de interfaz. La evaluación correcta descompone adquisición, preprocesamiento, inferencia, decisión, comunicación y presentación. Para un sistema de apoyo clínico, la métrica relevante es el tiempo hasta una acción segura y no únicamente el tiempo del kernel neuronal.

La Tabla 10 relaciona los beneficios potenciales de red con los mecanismos Edge AI que los producen y las métricas necesarias para verificarlos.

Tabla 10

Beneficios de red, mecanismos Edge AI y métricas de evaluación

Beneficio potencial	Mecanismo Edge AI	Compromiso introducido	Métrica recomendada
Menor ancho de banda	Inferencia y extracción local de características	Pérdida de información si solo se transmite la decisión	Bytes/evento, radio-on time, tasa de eventos y política de envío.
Menor latencia	Decisión cerca del sensor	Modelo local más pequeño o menos preciso	Latencia end-to-end y percentiles, no solo inferencia.
Mayor disponibilidad	Operación offline y store-and-forward	Estado local que debe reconciliarse	Tiempo útil sin WAN, pérdida de eventos, consistencia tras recuperación.
Menor dependencia de nube	Funciones mínimas autónomas	Complejidad de ciclo de vida en el dispositivo	Capacidad funcional por estado de conectividad.
Potencial ahorro energético	Menos transmisiones	Cómputo local consume energía	Energía por inferencia (J/inferencia), energía por evento transmitido (J/evento) y energía total del sistema bajo un ciclo de trabajo definido

Nota. WAN = wide area network. Elaboración propia a partir de LoRa Alliance (2025), Rocha et al. (2024) y Wakili y Bakkali (2026; texto completo no recuperado: caracterización basada en metadatos y repositorio público).

3.5. Categoría analítica V: ciberseguridad, privacidad, confianza e interoperabilidad

El procesamiento en el borde permite mantener localmente determinados datos y reducir la transmisión de bioseñales crudas; sin embargo, esta residencia local de los datos no debe equipararse automáticamente con una mayor privacidad. Las revisiones de seguridad IoMT identifican amenazas en autenticación, comunicaciones, malware, disponibilidad, acceso físico e integridad del dispositivo (Messinis et al., 2024; Naghib et al., 2025; Shammar et al., 2025). Cuando la IA se ejecuta en el extremo, a estas amenazas se añaden robo o sustitución del modelo, manipulación de entradas, deriva no detectada, ataques a la cadena de actualización y filtraciones derivadas de telemetría o entrenamiento distribuido.

Los canales laterales ejemplifican el problema físico. TinyPower mostró que las trazas de potencia o electromagnéticas pueden explotarse mediante redes neuronales podadas para recuperar claves del Advanced Encryption Standard (AES) en microcontroladores y matrices de puertas programables en campo (field-programmable gate arrays, FPGA) (Li et al., 2024). El estudio no es sanitario, pero resulta directamente relevante como modelo de amenaza para dispositivos médicos físicamente accesibles. La defensa requiere una combinación de implementaciones criptográficas resistentes, aislamiento, diseño de hardware, control de acceso y evaluación específica de side channels.

NIST estandarizó en 2025 la familia Ascon para criptografía ligera en dispositivos restringidos, incluyendo Ascon-AEAD128 y funciones hash/XOF (Turan et al., 2025). Este respaldo es más sólido que afirmar ahorros energéticos fijos frente a AES. La ventaja de Ascon es ofrecer una

alternativa estandarizada de criptografía ligera para dispositivos restringidos; cualquier comparación con AES debe realizarse sobre el MCU, compilador y modo de operación concretos. La guía FDA de febrero de 2026 refuerza que la ciberseguridad de dispositivos médicos debe gestionarse dentro del sistema de calidad y del ciclo de vida, no como una capa añadida al final (U.S. Food and Drug Administration, 2026).

El aprendizaje federado puede reducir la centralización de datos, pero conserva costos de comunicación, heterogeneidad de datos no independientes ni idénticamente distribuidos (non-IID), ataques a actualizaciones y riesgo de fuga. Una revisión de 2025 concluyó que gran parte de la literatura de FL en salud no estaba preparada para uso clínico por problemas de privacidad, generalización, sesgo y costo de comunicación (Li et al., 2025). Los diseños federados para TinyML, como TinyCFL (Ibrar et al., 2026), son relevantes porque aproximan la colaboración al hardware real, aunque su adopción en redes intermitentes exige compresión, selección de clientes, sincronización tolerante a fallos y evaluación de energía.

La privacidad diferencial agrega garantías formales, pero introduce un trade-off cuantificable. Una revisión de 2026 encontró que presupuestos de privacidad más estrictos pueden degradar el rendimiento y, en algunos casos, ampliar brechas entre subgrupos (Mohammadi et al., 2026). Para Edge AI orientado a la inclusión, esto significa que privacidad, exactitud y equidad deben auditarse conjuntamente. El uso de DP o FL no garantiza por sí mismo una mayor equidad.

En seguridad clínica, privacidad e interoperabilidad, las normas y guías convergen en un enfoque de ciclo de vida. La OMS destaca el uso previsto, la validación externa, la calidad y representatividad de los datos, la protección de la privacidad, la supervisión y la monitorización (World Health Organization, 2023). Los principios de Good Machine Learning Practice añaden la necesidad de datos representativos, independencia adecuada de los conjuntos de prueba, evaluación del equipo humano-IA, información clara al usuario y seguimiento del modelo desplegado (U.S. Food and Drug Administration, 2025). IEEE 11073 SDC aporta mecanismos de interoperabilidad de dispositivos, mientras FHIR aporta intercambio semántico de información (HL7 International, 2023; IEEE & ISO, 2024). Ninguno reemplaza el análisis de riesgos clínicos ni la validación end-to-end del comportamiento conjunto.

La Tabla 11 organiza las principales superficies de ataque y las mitigaciones prioritarias para un sistema Edge AI sanitario.

Tabla 11

Superficies de ataque, mitigaciones y fuentes de evidencia

Superficie	Amenaza relevante	Mitigaciones prioritarias	Evidencia/estándar
Dispositivo físico	Extracción de claves/modelo, manipulación, side channels	Secure boot, almacenamiento de claves, aislamiento, hardening, pruebas físicas	TinyPower (Li et al., 2024); FDA (U.S. Food and Drug Administration, 2026).
Firmware/modelo	OTA maliciosa, reversión a una versión anterior (rollback), modelo obsoleto, drift	Firmas, versionado, rollback protection, monitorización y plan de actualización	FDA (U.S. Food and Drug Administration, 2026); OMS (World Health Organization, 2023).
Comunicación	Intercepción, replay, DoS	AEAD, autenticación mutua, rate limiting, redundancia	NIST Ascon (Turan et al., 2025); revisiones IoMT (Messinis et al., 2024; Naghib et al., 2025; Shammar et al., 2025).
Aprendizaje distribuido	Inversion/poisoning, clientes heterogéneos, alto tráfico	Secure aggregation, DP, robust aggregation, selección de clientes, compresión	FL healthcare (Li et al., 2025); DP (Mohammadi et al., 2026).
Interoperabilidad	Mapeo semántico incorrecto, pérdida de contexto	Perfiles, terminologías, proveniencia, validación de recursos y pruebas de integración	FHIR (HL7 International, 2023); IEEE 11073 SDC (IEEE & ISO, 2024).
Uso clínico	Automation bias, falsas alarmas, falta de generalización	Human-in-the-loop, uso previsto, validación externa, monitorización posdespliegue	OMS (World Health Organization, 2023).

Nota. Las fuentes específicas se indican en la última columna; las medidas deben interpretarse dentro de un proceso de gestión de riesgos y ciclo de vida. AEAD = authenticated encryption with associated data; DoS = denial of service; DP = differential privacy; FL = federated learning; OTA = over-the-air. Elaboración propia.

3.6. Categoría analítica VI: inclusión digital, equidad y sostenibilidad

La evidencia revisada permite diferenciar con claridad la resiliencia técnica de la equidad sanitaria. Edge AI puede preservar determinadas funciones de inferencia y alerta cuando la conectividad WAN es intermitente o inexistente; sin embargo, esta capacidad no demuestra por sí sola mejoras de equidad. La transición desde la resiliencia técnica hacia resultados de inclusión depende de la infraestructura, la asequibilidad, los recursos humanos, la accesibilidad, el diseño del servicio, la integración clínica y los resultados sanitarios. Por tanto, la contribución de Edge AI a la equidad debe considerarse una hipótesis de impacto que requiere validación empírica.

El contexto ecuatoriano ilustra el problema. El uso de Internet fue menor en zonas rurales que urbanas en 2025 (Instituto Nacional de Estadística y Censos, 2025), y el estudio de hospitales generales de Imbabura documentó una viabilidad de telemedicina muy baja en establecimientos rurales (Armas Cárdenas et al., 2026). Edge AI podría reducir el requisito de conexión continua en algunos flujos de monitorización, pero no resuelve problemas de disponibilidad de equipos, capacitación, mantenimiento, suministro eléctrico, referencia clínica o integración con el sistema

nacional. Por ello, el caso ecuatoriano justifica investigar soluciones offline-first, pero no permite afirmar que Edge AI ya haya mejorado indicadores de salud.

La arquitectura RuralEdgeHealth de 2026 constituye una evidencia más directa de factibilidad en entornos restringidos: ejecuta clasificación en teléfonos Android de entrada, integra almacenamiento cifrado y reporta una evaluación exploratoria de usabilidad (Wakili & Bakkali, 2026). Los propios autores la presentan como fase de factibilidad y no como evidencia definitiva de impacto poblacional. Esa cautela debería extenderse a toda la literatura de Edge AI "inclusiva".

La accesibilidad también excede la conectividad. El conjunto de herramientas OMS-UIT de 2024 enfatiza barreras para personas con discapacidad y otras poblaciones marginadas (World Health Organization & International Telecommunication Union, 2024). En un sistema Edge AI, una alarma local que no puede ser percibida, comprendida o accionada por el usuario no es inclusiva, aunque funcione offline. Deben incorporarse multimodalidad, lenguaje claro, localización cultural, compatibilidad con tecnologías de apoyo, posibilidad de asistencia y rutas de escalamiento humano.

El sesgo de datos es otra dimensión. Los modelos compactos suelen entrenarse en conjuntos de datos de referencia porque reducen los costos de desarrollo; sin embargo, las prestaciones pueden variar entre poblaciones, sensores y condiciones de adquisición. La compresión no crea el sesgo, pero puede interactuar con él. La privacidad diferencial añade otro compromiso porque los subgrupos pequeños pueden sufrir un mayor deterioro (Mohammadi et al., 2026). La evaluación de equidad requiere métricas desagregadas y validación local.

Finalmente, la sostenibilidad no puede omitirse. El Global E-waste Monitor 2024 estimó que en 2022 se generaron 62 millones de toneladas de residuos electrónicos y que solo el 22,3 % se documentó como formalmente recolectado y reciclado (International Telecommunication Union & United Nations Institute for Training and Research, 2024). Esta cifra es global y no mide específicamente dispositivos médicos Edge AI; sirve como señal de contexto. El diseño para zonas de recursos limitados debería considerar vida útil, baterías reemplazables o recargables, reparabilidad, disponibilidad de repuestos y planes de fin de vida. Una solución barata de adquirir, pero cara de mantener puede profundizar la inequidad.

La Tabla 12 diferencia los aportes potenciales de Edge AI de los determinantes que requieren intervención organizacional, económica o social.

Tabla 12

Dimensiones de inclusión y métricas para evaluar equidad

Dimensión de inclusión	Qué puede aportar Edge AI	Qué Edge AI no resuelve por sí solo	Indicador de evaluación
Conectividad	Inferencia offline; telemetría esporádica	Cobertura física para escalamiento o sincronización	Horas/días operativos sin WAN; eventos sincronizados.
Asequibilidad	Menos tráfico y posible hardware de bajo costo	TCO, reposición, licencias, soporte y energía	Costo por paciente/año; costo por evento clínicamente útil.
Accesibilidad	Alertas locales y apps offline	Discapacidad, lenguaje, alfabetización y usabilidad	Tasa de éxito, errores de uso y cumplimiento de criterios de accesibilidad aplicables (p. ej., WCAG para interfaces digitales)
Equidad algorítmica	Personalización/potencial de adaptación local	Representatividad de dataset y sesgo estructural	Sensibilidad/especificidad/calibración por subgrupo.
Capacidad del sistema	Menor dependencia de centros de datos	Personal, protocolos clínicos, referencia y mantenimiento	Tiempo a intervención; carga de trabajo; disponibilidad del servicio.
Sostenibilidad	Menos tráfico/cómputo remoto en algunos casos	Baterías, obsolescencia, e-waste	Vida útil, reparabilidad, energía total y fin de vida.

Nota. TCO = costo total de propiedad; WCAG = Web Content Accessibility Guidelines; WAN = wide area network. Elaboración propia a partir de Armas Cárdenas et al. (2026), Instituto Nacional de Estadística y Censos (2025), International Telecommunication Union y United Nations Institute for Training and Research (2024), y World Health Organization e International Telecommunication Union (2024).

3.7. Síntesis transversal: vacíos de investigación y requisitos mínimos de reporte

La síntesis muestra que la frontera de investigación ha pasado de demostrar la factibilidad de ejecutar IA en el borde a demostrar que un sistema completo puede mantenerse seguro, interoperable y clínicamente útil durante su ciclo de vida. Los vacíos prioritarios son los siguientes:

- Validación prospectiva y externa: realizar más estudios con TRL 5–7 y NMV3–NMV5, con pacientes, múltiples centros y condiciones de uso reales.
- Evaluación comparativa reproducible de recursos (benchmarking): distinguir la energía atribuible a la inferencia, el consumo del microcontrolador durante la ejecución, el consumo de la radio y el consumo energético del sistema completo; reportar energía por inferencia, potencia media y ciclo de trabajo bajo condiciones explícitamente definidas.
- Robustez frente al ruido y la deriva: evaluar movimiento, sensores distintos, cambios demográficos y degradación del dispositivo.
- Ciclo de vida seguro: actualizaciones firmadas, inventario de versiones, monitorización de deriva, reversión segura (rollback) y respuesta a vulnerabilidades.
- Interoperabilidad verificable: perfiles FHIR y especificaciones IEEE 11073, proveniencia, terminologías y pruebas end-to-end con EHR.

- Evidencia de inclusión: medir acceso, uso, costo, tiempo a atención, abandono, resultados y brechas por subgrupo.
- Evaluación económica y ambiental: costo total de propiedad, mantenibilidad, reparabilidad, baterías y fin de vida.

4. Discusión

La evidencia revisada permite reformular el papel de Edge AI en salud digital con mayor precisión. Su aporte principal no es reemplazar la nube, sino distribuir funciones para que las tareas clínicamente críticas no dependan de la disponibilidad instantánea de un centro remoto. Esta diferencia es particularmente relevante en redes intermitentes o costosas, donde la resiliencia depende de definir estados operativos, prioridades de telemetría, almacenamiento local, sincronización, ciberseguridad y supervisión humana (human-in-the-loop). En consecuencia, la unidad real de evaluación no debe ser el modelo aislado, sino el sistema sociotécnico completo: sensor, dispositivo, enlace, software, personal, flujo clínico, mantenimiento y gobernanza.

4.1. Compromisos estructurales (trade-offs)

El primer compromiso contrapone la complejidad del modelo y el presupuesto de recursos. Reducir parámetros y precisión numérica puede hacer viable el despliegue, pero la exactitud promedio no es suficiente: deben preservarse sensibilidad, calibración y desempeño en subgrupos clínicamente relevantes. La compresión debe incorporarse a la validación, no tratarse como una etapa de ingeniería posterior.

El segundo compromiso contrapone la localización de los datos y la superficie de ataque. Mantener los datos en el dispositivo reduce la exposición en tránsito y el almacenamiento central, pero incrementa la importancia del firmware, el hardware, las actualizaciones y el acceso físico. La seguridad “del borde” debe presupuestarse en memoria, cómputo y energía desde el diseño.

El tercer compromiso contrapone autonomía y gobernanza. Un dispositivo capaz de inferir sin red debe saber qué acciones están autorizadas durante la desconexión. En aplicaciones de apoyo al diagnóstico, una alerta local puede ser apropiada; una acción terapéutica autónoma exige un nivel de validación y control mucho mayor. La degradación funcional controlada no implica mantener toda la funcionalidad, sino preservar el subconjunto de funciones previamente definido como seguro, útil y autorizado para el estado operativo correspondiente.

El cuarto compromiso contrapone privacidad colaborativa y costo de comunicación. FL reduce la necesidad de centralizar datos, pero puede demandar múltiples rondas de transferencia y no elimina los ataques de inferencia. En conectividad limitada, un esquema jerárquico, asincrónico o con clientes seleccionados puede ser preferible, pero debe evaluarse frente a una línea base local y no asumirse superior por definición.

4.2. Propuesta: modelo de referencia funcional para Edge AI resiliente y orientado a la inclusión

A partir de la evidencia revisada y de las limitaciones identificadas, se propone un modelo de referencia funcional para Edge AI resiliente y orientado a la inclusión en salud digital bajo condiciones variables de conectividad. El modelo formaliza la adaptación de las funciones del sistema ante cambios en la disponibilidad y calidad del enlace mediante degradación funcional controlada, persistencia local, priorización de comunicaciones y recuperación segura. Asimismo, incorpora como requisitos transversales la interoperabilidad, la ciberseguridad, la privacidad, la accesibilidad, la representatividad de los datos, la sostenibilidad y la evaluación de equidad. El modelo constituye una síntesis conceptual derivada de las RQ1–RQ6 y no representa una arquitectura específica de implementación, una norma técnica ni un sistema clínicamente validado.

La Figura 2 presenta un modelo de referencia funcional basado en modos operativos para Edge AI en salud digital bajo conectividad variable. El modelo se organiza en dos dimensiones complementarias. La primera representa la cadena funcional del sistema, desde la adquisición de datos y la inferencia local hasta la persistencia, la comunicación y los servicios proporcionados por pasarelas o infraestructura en la nube. La segunda representa cuatro modos operativos determinados por las condiciones de conectividad: conectado, degradado, desconectado y recuperación.

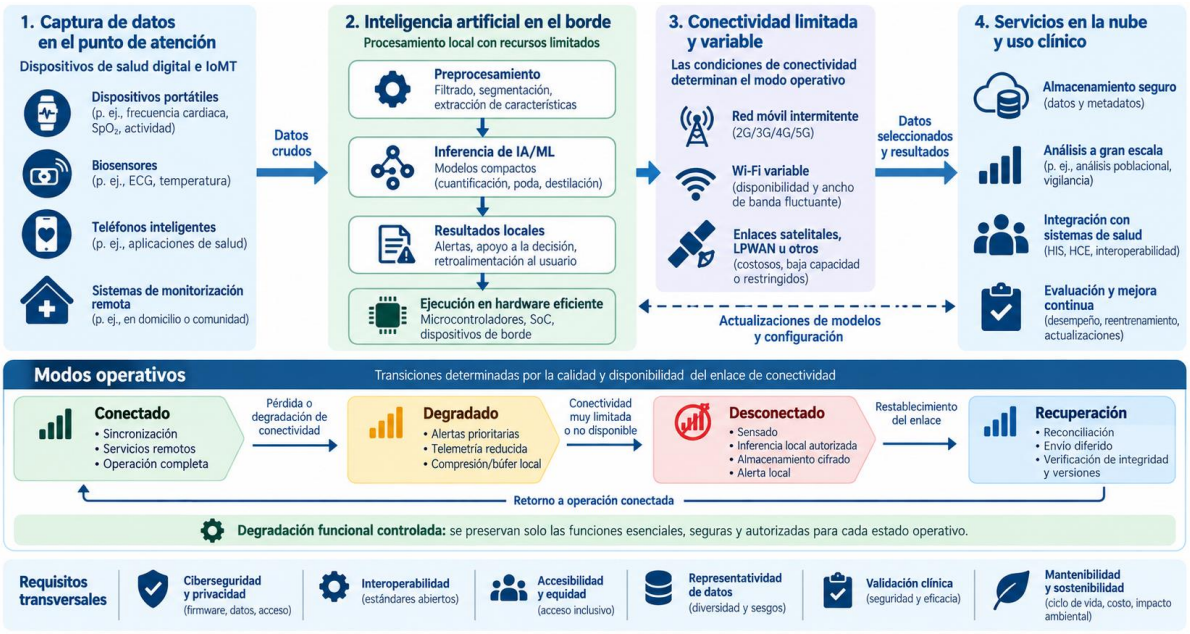
Ante la degradación o pérdida del enlace, las funciones del sistema se adaptan mediante mecanismos de degradación funcional controlada, preservando localmente aquellas capacidades previamente definidas como esenciales, seguras y autorizadas para el uso previsto. Una envolvente transversal incorpora requisitos de ciberseguridad, privacidad, interoperabilidad, accesibilidad, representatividad de los datos, validación clínica, mantenibilidad y sostenibilidad.

El modelo contempla cuatro modos operativos: conectado, degradado, desconectado y recuperación. En modo conectado, el sistema puede sincronizar datos, políticas y modelos de acuerdo con su uso previsto. En modo degradado, prioriza alertas y reduce la telemetría no esencial. En modo desconectado, conserva adquisición, inferencia local autorizada, interacción segura y almacenamiento cifrado, mientras que las funciones que requieren contexto remoto quedan suspendidas o se sustituyen por procedimientos humanos previamente definidos. Durante la recuperación, el sistema sincroniza información según prioridades, conserva la proveniencia, verifica la integridad y las versiones, y resuelve conflictos antes de restablecer el funcionamiento completo.

La Tabla 13 resume los cuatro modos operativos del modelo de referencia funcional propuesto y los controles que deben preservarse ante cambios en las condiciones de conectividad.

Figura 2

Modelo de referencia funcional basado en modos operativos para Edge AI en salud digital bajo conectividad variable



Nota. La figura representa un modelo de referencia funcional conceptual y no una arquitectura específica de implementación, un dispositivo clínico concreto ni un sistema clínicamente validado. Elaboración propia a partir de la síntesis narrativa de la evidencia revisada.

Tabla 13

Modos operativos y controles del modelo de referencia funcional propuesto

Estado	Funciones que continúan	Funciones que se restringen	Controles críticos
Conectado	Inferencia local + sincronización + servicios remotos	No se restringen funciones por conectividad; permanecen las restricciones clínicas, de seguridad y de política definidas para el uso previsto	Autenticación, monitorización, versionado y QoS
Degradado	Inferencia local, alertas prioritarias, resumen	Telemetría no esencial, entrenamiento federado intensivo	Priorización, política de reintentos y control de congestión, compresión y búfer local
Desconectado	Sensado, inferencia local autorizada para operación desconectada según el uso previsto, alerta local y persistencia	Consulta a servicios en la nube, actualización remota y agregación global	Estado seguro, supervisión humana, cifrado local, reloj confiable y preservación de proveniencia
Recuperación	Reconciliación, envío diferido, verificación de versión	Acciones automáticas que dependan de datos no reconciliados	Orden de prioridad, idempotencia, integridad y auditoría

Nota. El modelo constituye una propuesta funcional derivada de la síntesis narrativa y no representa por sí mismo una arquitectura específica de implementación ni un sistema clínicamente validado. QoS = quality of service. Elaboración propia.

4.3. Implicaciones para Ecuador y otros entornos de conectividad desigual

En Ecuador, la brecha urbano-rural observada en el uso de Internet y en la viabilidad institucional de la telemedicina justifica explorar arquitecturas que no necesiten conectividad permanente (Armas Cárdenas et al., 2026; Instituto Nacional de Estadística y Censos, 2025). El argumento es extensible, con cautela, a países y regiones donde coexisten cobertura nominal, baja capacidad, interrupciones eléctricas, costos de datos elevados, limitada disponibilidad de especialistas y heterogeneidad de equipamiento. En estos contextos, la ventaja de Edge AI no reside solo en disminuir la latencia, sino en reducir la dependencia de servicios remotos para funciones mínimas seguras. Sin embargo, una arquitectura técnicamente autónoma puede fracasar si no existen repuestos, personal capacitado, procesos de calibración, soporte de firmware, rutas de escalamiento clínico o presupuesto para reposición.

Para centros de salud o brigadas móviles, un teléfono de entrada puede ser un gateway más realista que un servidor fog dedicado; para wearables de larga duración, un MCU puede filtrar y detectar eventos; y para hospitales con LAN estable, un servidor local puede ejecutar modelos mayores y traducción FHIR. La selección tecnológica debe partir del flujo clínico: evento que se debe detectar, actor responsable, tolerancia a la desconexión, consecuencias de falsos negativos o positivos, datos mínimos que se deben conservar y tiempo máximo hasta la intervención.

Solo después se definen el hardware, el enlace y la ubicación de la inferencia. En países con brechas tecnológicas, la escalabilidad depende además de compras públicas sostenibles, disponibilidad regional de componentes, documentación en el idioma local, formación técnica y capacidad de operar durante fallos parciales; replicar una solución urbana sin rediseñarla para el contexto suele trasladar, en lugar de resolver, la dependencia de infraestructura.

4.4. Sostenibilidad económica, mantenimiento y escalabilidad

La sostenibilidad económica debe evaluarse mediante el costo total de propiedad (TCO), no únicamente por el precio del dispositivo. Una solución Edge AI puede reducir el gasto de transmisión o de uso de la nube, pero trasladar costos a la adquisición de gateways, el almacenamiento local, el reemplazo de baterías, la calibración, la gestión de claves, la actualización de firmware, la monitorización de la deriva del modelo, el soporte de campo y la reposición por obsolescencia.

En despliegues rurales, los costos logísticos y el tiempo fuera de servicio pueden superar el valor del hardware. Por ello, los estudios deberían reportar el costo por paciente/año, el costo por evento clínicamente útil, la vida esperada, la frecuencia de mantenimiento y los recursos humanos requeridos, además de comparar con alternativas realistas como cloud-only, store-and-forward sin IA o procesos clínicos convencionales.

La escalabilidad introduce un segundo problema: administrar cientos o miles de nodos convierte una prueba de concepto en una operación MLOps/DevSecOps distribuida. La operación requiere inventario de activos, identidad de dispositivo, telemetría mínima, actualización segura y recuperable, políticas de compatibilidad de modelos, observabilidad del rendimiento, detección

de deriva y mecanismos para retirar versiones inseguras. La arquitectura debe tolerar que una parte de la flota permanezca temporalmente desactualizada por falta de conectividad sin producir estados clínicos incoherentes. El diseño modular, el uso de estándares interoperables y la separación entre funciones clínicas críticas y servicios auxiliares favorecen el escalamiento y la sustitución tecnológica, mientras que las dependencias propietarias rígidas pueden aumentar el lock-in y los costos del ciclo de vida.

4.5. Regulación, ética y gobernanza de datos clínicos en el borde

La ejecución local no elimina las obligaciones regulatorias asociadas al uso de IA en salud. La OMS enfatiza un uso previsto claramente definido, validación externa, calidad y representatividad de los datos, supervisión humana, ciberseguridad, documentación y monitorización durante el ciclo de vida (World Health Organization, 2023). De forma complementaria, los principios internacionales de Good Machine Learning Practice destacan equipos multidisciplinarios, separación apropiada entre conjuntos de datos, evaluación del desempeño del equipo humano-IA, información clara al usuario y monitorización de modelos desplegados (U.S. Food and Drug Administration, 2025). En Edge AI estos requisitos deben materializarse en controles que funcionen aun cuando la conectividad sea limitada: versionado local, trazabilidad de inferencias, registro de configuración, detección de fallos, actualización firmada y estados seguros ante pérdida de sincronización.

La privacidad requiere gobernar todo el ciclo del dato, no solo evitar que la señal cruda viaje a la nube. Deben definirse finalidad, base jurídica, minimización, retención local, cifrado en reposo y tránsito, gestión de claves, acceso basado en roles, borrado, respaldo, sincronización y tratamiento de telemetría. Las actualizaciones de modelos, los registros (logs) y los gradientes federados también pueden revelar información sensible.

En Ecuador, la Ley Orgánica de Protección de Datos Personales reconoce los datos de salud como categorías de protección reforzada y exige principios de licitud, finalidad, proporcionalidad, seguridad y responsabilidad; su Reglamento General desarrolla las obligaciones aplicables al tratamiento (Asamblea Nacional del Ecuador, 2021; Presidencia de la República del Ecuador, 2023). Por tanto, una implementación sanitaria debe incorporar la evaluación de riesgos de privacidad y seguridad desde el diseño, especialmente cuando procesa datos biométricos, monitorización continua o decisiones automatizadas.

Desde la ética, la autonomía técnica del dispositivo aumenta la necesidad de límites explícitos a la autonomía decisional. Debe distinguirse entre una inferencia informativa, una alerta que requiere confirmación humana y una acción terapéutica automatizada. El nivel de supervisión debe ser proporcional al riesgo del uso previsto. Además, equidad, accesibilidad y privacidad pueden entrar en tensión: una política de minimización demasiado agresiva puede eliminar datos necesarios para auditar sesgos; una técnica de privacidad diferencial puede degradar más el desempeño de subgrupos pequeños; y una interfaz offline puede excluir a usuarios con discapacidad si no incorpora diseño accesible. En consecuencia, la gobernanza ética debe tratar privacidad, seguridad, desempeño y equidad como objetivos simultáneos y medibles, no como atributos independientes.

4.6. Agenda de investigación orientada a impacto

Para cerrar la brecha entre demostración técnica y salud inclusiva, las futuras investigaciones deberían diseñarse considerando variables de resultado en cuatro dimensiones complementarias:

- Técnica: latencia end-to-end, energía total, memoria pico, disponibilidad sin red, bytes transmitidos y tasa de recuperación.
- Clínica: sensibilidad, especificidad, calibración, tiempo hasta intervención, carga de falsas alarmas y seguridad por uso previsto.
- Equidad y experiencia: desempeño por subgrupo, accesibilidad, tasa de uso, abandono, comprensión de alertas y tiempo de atención en población objetivo.
- Económica y ambiental: costo total de propiedad, costo por caso útil, mantenimiento, vida de batería, reparabilidad y fin de vida.

Un ensayo o piloto que reporte únicamente exactitud y latencia no puede responder RQ6. Del mismo modo, un estudio de equidad sin instrumentación de calidad de servicio y disponibilidad no puede atribuir mejoras a la arquitectura Edge AI. La investigación más valiosa será aquella que conecte las cuatro capas y compare Edge AI con alternativas realistas: cloud-only, reglas simples, store-and-forward sin IA o atención habitual.

4.7. Limitaciones de la revisión

Esta revisión presenta limitaciones que condicionan la fuerza de sus conclusiones. Primero, 119 de los 248 informes buscados a texto completo no pudieron recuperarse por tratarse de publicaciones de acceso cerrado sin copia abierta localizable. El corpus evaluado (129 informes) podría no representar al conjunto de la literatura elegible, y los estudios no recuperados permanecen sin evaluar en la bitácora de recuperación.

Segundo, la evaluación de la calidad estuvo limitada por la documentación insuficiente de los estudios primarios: la mayoría de los juicios “incierto” en QUADAS-2 (10 estudios), PROBAST (5 estudios) y la lista de verificación técnica (71 estudios) reflejan información insuficiente en la extracción, no necesariamente defectos confirmados. En particular, la reproducibilidad fue incierta en los 71 estudios de validez técnica porque la disponibilidad de código y datos no se capturó sistemáticamente; esta es una limitación de la extracción que debe corregirse en actualizaciones futuras.

Tercero, se aplicó AMSTAR 2 (Shea et al., 2017) a las 5 revisiones sistemáticas del corpus secundario. Solo una pudo evaluarse con texto completo —la revisión sobre wearables de EEG intraauricular— y obtuvo confianza críticamente baja por cuatro defectos críticos: ausencia de protocolo previo, de lista de estudios excluidos, de evaluación del riesgo de sesgo de los estudios incluidos y de consideración del sesgo en la interpretación. Las cuatro revisiones restantes no pudieron recuperarse a texto completo; por ello, sus ítems se consideran no evaluables y no se les asigna una categoría global de confianza. Esta limitación reduce la capacidad de usar el corpus

secundario como evidencia de alta certeza, razón por la que las revisiones se emplean únicamente para triangulación y contextualización.

Cuarto, la búsqueda se ejecutó en dos bases de datos (PubMed y OpenAlex). El intento de consulta a IEEE Xplore no devolvió resultados auditables y la fuente se excluyó con declaración explícita; la literatura gris y regional puede estar subrepresentada, y solo se aceptaron publicaciones en inglés o español. La discrepancia de un registro entre el estimador de OpenAlex (734) y la exportación por cursor (735) se declara sin atribuirle causa no verificada.

Quinto, nueve estudios inicialmente candidatos fueron reclasificados como evidencia contextual tras una readjudicación estricta del criterio de ejecución demostrada en el borde (simulación o procesador sustituto). La decisión está documentada por estudio, pero el límite entre prototipo de laboratorio y simulación conserva un componente de juicio que debe interpretarse con cautela.

Sexto, la heterogeneidad de tareas, señales, plataformas, diseños de validación y fronteras de medición impide un metanálisis global y obliga a una síntesis narrativa; las cifras de estudios individuales son resultados condicionados a su protocolo. Las escalas TRL y NMV son herramientas analíticas de esta revisión y no equivalen a aprobación regulatoria ni a evidencia de eficacia clínica. Finalmente, los datos de 2026 son parciales (corte el 9 de septiembre) y la rápida evolución de hardware y runtimes hace que las cifras de memoria o energía envejezcan antes que los principios arquitectónicos.

5. Conclusiones

5.1. Aportes originales y conclusiones integradas

El primer aporte del capítulo es haber reconstruido y documentado de forma íntegra y trazable el proceso de revisión: se identificaron 1 031 registros/identificadores en PubMed y OpenAlex; un identificador de PubMed no pudo exportarse, por lo que quedaron 1 030 registros auditables antes de la deduplicación; 879 registros fueron cribados, 129 informes primarios se evaluaron a texto completo y el corpus final comprende 86 estudios primarios con ejecución demostrada de IA en el borde, además de 29 revisiones como evidencia secundaria. La revisión establece también una separación explícita entre resiliencia técnica, madurez tecnológica y evidencia clínica.

Edge AI puede conservar adquisición, filtrado, inferencia y alerta cuando la WAN se degrada, y técnicas como cuantificación, poda y destilación hacen viable el despliegue en hardware restringido —28 estudios reportan microcontroladores—. Sin embargo, la evaluación metodológica muestra que la factibilidad de despliegue no equivale a efectividad clínica: predominan prototipos experimentales (n = 40) y validaciones técnicas (n = 31), la documentación sobre reproducibilidad es insuficiente y son escasos los estudios prospectivos, multicéntricos o posdespliegue.

El segundo aporte consiste en conceptualizar Edge AI como un problema de distribución funcional del cómputo y no como una oposición binaria entre borde y nube. En redes con capacidad

limitada, la inferencia local permite transformar flujos continuos de bioseñales en telemetría selectiva, reducir la dependencia de intercambios de ida y vuelta con servicios remotos y mantener funciones previamente definidas durante periodos de conectividad degradada o ausente. La nube conserva un papel relevante en el entrenamiento, la coordinación, la gestión del ciclo de vida y el análisis longitudinal. El modelo de referencia funcional propuesto (Figura 2) sintetiza esta relación mediante cuatro modos operativos —conectado, degradado, desconectado y recuperación— y establece mecanismos de degradación funcional controlada, persistencia con integridad, priorización de comunicaciones y reconciliación interoperable.

El tercer aporte consiste en incorporar al mismo modelo de referencia funcional requisitos transversales de ciberseguridad, interoperabilidad, privacidad, sostenibilidad, accesibilidad y evaluación de equidad. Mantener determinados datos localmente puede reducir su exposición durante la transmisión, pero incrementa la relevancia de la seguridad del hardware y del firmware, la gestión de claves, las actualizaciones y el acceso físico. IEEE 11073 SDC y HL7 FHIR pueden contribuir a articular el flujo dispositivo–pasarela–historia clínica electrónica, pero no sustituyen la validación semántica, técnica ni clínica. En este marco, Edge AI debe entenderse como un habilitador potencial de resiliencia técnica y de inclusión, no como una tecnología intrínsecamente equitativa; todavía se requiere evidencia longitudinal que demuestre mejoras sostenidas de acceso, resultados, costo-efectividad y reducción de disparidades en poblaciones rurales y desatendidas.

5.2. Recomendaciones para actores clave

Para investigadores, se recomienda diseñar estudios que reporten simultáneamente rendimiento clínico y técnico, proporcionen información suficiente para determinar la madurez tecnológica (TRL) y la madurez de validación (NMV) mediante criterios explícitos, incluyan validación externa y prospectiva, comparadores realistas y métricas de energía de extremo a extremo, conectividad, seguridad, accesibilidad y subgrupos. También deberían publicarse artefactos de reproducibilidad —configuración de hardware, versiones, protocolos de medición y perfiles de interoperabilidad— sin comprometer información sensible.

Para responsables de políticas públicas, la prioridad es evitar que las estrategias de salud digital dependan de una conectividad ideal. Los programas deberían exigir interoperabilidad, protección de datos, ciberseguridad por ciclo de vida, evaluación de TCO, accesibilidad y planes de mantenimiento; financiar pilotos en regiones con brechas reales; y definir mecanismos de evaluación que midan acceso, tiempo hasta la atención, continuidad del servicio y equidad, además de la exactitud. Para organizaciones sanitarias, se recomienda seleccionar Edge AI a partir del flujo clínico y del riesgo, establecer gobernanza de modelos y datos, un inventario de dispositivos, actualización segura, procedimientos de operación offline, supervisión humana y criterios de escalamiento o retiro del sistema.

5.3. Trabajos futuros

La prioridad de validación clínica es avanzar desde NMV1–NMV2 hacia estudios prospectivos y multicéntricos que evalúen desempeño en la población objetivo, robustez ante cambios de sensor

y deriva, calibración, carga de falsas alarmas, tiempo hasta la intervención y eventos de seguridad. Los estudios deberían comparar Edge AI con alternativas pertinentes —cloud-only, reglas simples, store-and-forward sin IA y atención habitual— para estimar el valor incremental de la inteligencia local.

En interoperabilidad, los trabajos futuros deben validar flujos de extremo a extremo desde el sensor hasta la historia clínica electrónica (electronic health record, EHR), incluyendo pruebas de integración, mapeos semánticos, proveniencia, identificación, perfiles FHIR, especificaciones IEEE 11073, sincronización después de periodos sin conexión y consistencia de versiones. Una línea especialmente relevante es comprobar cómo los sistemas preservan significado clínico y trazabilidad cuando operan durante horas o días sin conectividad y luego reconcilian eventos locales con registros institucionales.

Finalmente, la evaluación del impacto social debe superar indicadores de factibilidad. Se requieren estudios longitudinales que midan adopción, accesibilidad, costos para pacientes y organizaciones, continuidad de atención, desempeño por subgrupo, alfabetización digital, carga para el personal, mantenimiento, residuos electrónicos y efectos distributivos. Solo la convergencia de validación clínica, interoperabilidad verificable y evidencia social permitirá determinar cuándo Edge AI mejora realmente la salud digital en entornos con conectividad limitada.

Referencias bibliográficas

Los 86 estudios incluidos en la revisión se listan en el Apéndice C.

Armas Cárdenas, J. C., Lucio Vásquez, E. M., Puetate Huera, G. H., & Grijalva Maigua, C. N. (2026). La brecha digital, impacto en la implementación de telemedicina en los Hospitales Generales de Imbabura. Metanoia: Revista de Ciencia, Tecnología e Innovación, 12(1), 413–428. https://doi.org/10.61154/metanoia.v12i1.4226

Asamblea Nacional del Ecuador. (2021). Ley Orgánica de Protección de Datos Personales. Registro Oficial, Quinto Suplemento No. 459. https://www.registroficial.gob.ec/quinto-suplemento-al-registro-oficial-no-459/

Hizem, M., Bousbia, L., Ben Dhiab, Y., Aoueileyine, M. O.-E., & Bouallegue, R. (2025). Reliable ECG anomaly detection on edge devices for Internet of Medical Things applications. Sensors, 25(8), 2496. https://doi.org/10.3390/s25082496

HL7 International. (2023). FHIR Release 5 (v5.0.0). https://hl7.org/fhir/R5/

Ibrar, K., Fusco, P., Rimoli, G. P., Palmieri, F., & Ficco, M. (2026). Cross-modal federated TinyML for MCU-based Internet of Medical Things. IEEE Journal of Biomedical and Health Informatics. Advance online publication. https://doi.org/10.1109/JBHI.2026.3679191

Institute of Electrical and Electronics Engineers & International Organization for Standardization. (2024). ISO/IEEE International Standard: Health informatics—Device interoperability—Part 10700: Point-of-care medical device communication—Standard for base requirements for participants in a Service-oriented Device Connectivity (SDC) system (IEEE/ISO 11073-10700-2024). https://standards.ieee.org/ieee/11073-10700/11716/

Instituto Nacional de Estadística y Censos. (2025). Tecnologías de la Información y Comunicación—TIC 2025: Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU).

International Telecommunication Union & United Nations Institute for Training and Research. (2024). The Global E-waste Monitor 2024. https://www.itu.int/pub/D-GEN-E_WASTE.01-2024

Keivanimehr, A. R., & Akbari, M. (2025). TinyML and edge intelligence applications in cardiovascular disease: A survey. Computers in Biology and Medicine, 186, 109653. https://doi.org/10.1016/j.compbiomed.2025.109653

Li, H., Ninan, M., Wang, B., & Emmert, J. M. (2024). TinyPower: Side-channel attacks with tiny neural networks. In 2024 IEEE International Symposium on Hardware Oriented Security and Trust (HOST) (pp. 320–331). IEEE. https://doi.org/10.1109/HOST55342.2024.10545382

Li, M., Xu, P., Hu, J., Tang, Z., & Yang, G. (2025). From challenges and pitfalls to recommendations and opportunities: Implementing federated learning in healthcare. Medical Image Analysis, 101, 103497. https://doi.org/10.1016/j.media.2025.103497

LoRa Alliance. (2025, October 8). RP002-1.0.5 LoRaWAN regional parameters. https://resources.lora-alliance.org/document/rp002-1-0-5-lorawan-regional-parameters

Messinis, S., Temenos, N., Protonotarios, N. E., Rallis, I., Kalogeras, D., & Doulamis, N. (2024). Enhancing Internet of Medical Things security with artificial intelligence: A comprehensive review. Computers in Biology and Medicine, 170, 108036. https://doi.org/10.1016/j.compbiomed.2024.108036

Mohammadi, M., Vejdanihemmat, M., Lotfinia, M., Rusu, M., Truhn, D., Maier, A., & Arasteh, S. T. (2026). Differential privacy for medical deep learning: Methods, tradeoffs, and deployment implications. npj Digital Medicine, 9, 93. https://doi.org/10.1038/s41746-025-02280-z

Naghib, A., Gharehchopogh, F. S., & Zamanifar, A. (2025). A comprehensive and systematic literature review on intrusion detection systems in the Internet of Medical Things: Current status, challenges, and opportunities. Artificial Intelligence Review, 58, 114. https://doi.org/10.1007/s10462-024-11101-w

National Aeronautics and Space Administration. (2023, September 27). Technology readiness levels. https://www.nasa.gov/directorates/somd/space-communications-navigation-program/technology-readiness-levels/

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Pazmiño-Ortiz, L., Cuenca-Sánchez, A., & Loarte-Cajamarca, B. (2026). Distributed intelligence in the Artificial Intelligence of Things: A review of artificial intelligence workload placement across the device-edge-fog-cloud continuum. Future Internet, 18(6), 296. https://doi.org/10.3390/fi18060296

Pereira, C. V. F., de Oliveira, E. M., & de Souza, A. D. (2024). Machine learning applied to edge computing and wearable devices for healthcare: Systematic mapping of the literature. Sensors, 24(19), 6322. https://doi.org/10.3390/s24196322

Presidencia de la República del Ecuador. (2023). Reglamento General de la Ley Orgánica de Protección de Datos Personales (Decreto Ejecutivo No. 904). Registro Oficial, Tercer Suplemento No. 435. https://www.registroficial.gob.ec/tercer-suplemento-al-registro-oficial-no-435/

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., & Koffel, J. B. (2021). PRISMA-S: An extension to the PRISMA Statement for reporting literature searches in systematic reviews. Systematic Reviews, 10, 39. https://doi.org/10.1186/s13643-020-01542-z

Rocha, A., Monteiro, M., Mattos, C., Dias, M., Soares, J., Magalhães, R., & de Macêdo, J. A. F. (2024). Edge AI for Internet of Medical Things: A literature review. Computers & Electrical Engineering, 116, 109202. https://doi.org/10.1016/j.compeleceng.2024.109202

Shammar, E., Cui, X., Zahary, A., Alsamhi, S. H., & Al-qaness, M. A. A. (2025). Threat to trust: A systematic review on Internet of Medical Things security. Journal of Parallel and Distributed Computing, 206, 105172. https://doi.org/10.1016/j.jpdc.2025.105172

Shea, B. J., Reeves, B. C., Wells, G., Thouless, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ, 358, j4008. https://doi.org/10.1136/bmj.j4008

Suryanarayanan, K., & Gajula, R. (2026). Memory-efficient CNN autoencoder for real-time ECG anomaly detection on TinyML-enabled edge devices. Future Internet, 18(6), 286. https://doi.org/10.3390/fi18060286

Turan, M. S., McKay, K., Kang, J., Kelsey, J., & Chang, D. (2025). Ascon-based lightweight cryptography standards for constrained devices: Authenticated encryption, hash, and extendable output functions (NIST Special Publication 800-232). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-232

U.S. Food and Drug Administration. (2025). Good machine learning practice for medical device development: Guiding principles. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles

U.S. Food and Drug Administration. (2026, February). Cybersecurity in medical devices: Quality management system considerations and content of premarket submissions: Guidance for industry and Food and Drug Administration staff. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/cybersecurity-medical-devices-quality-management-system-considerations-and-content-premarket

Wakili, A., & Bakkali, S. (2026). RuralEdgeHealth: A sustainable offline mobile edge AI reference architecture for intelligent IoT sensor classification in resource-constrained environments. Array, 30, 100912. https://doi.org/10.1016/j.array.2026.100912

Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M. G., Sterne, J. A. C., Bossuyt, P. M. M., & QUADAS-2 Group. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine, 155(8), 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., Mallett, S., & PROBAST Group. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Annals of Internal Medicine, 170(1), 51–58. https://doi.org/10.7326/M18-1376

World Health Organization & International Telecommunication Union. (2024). Implementation toolkit for accessible telehealth services. https://www.who.int/publications/i/item/9789240094161

World Health Organization. (2023). Regulatory considerations on artificial intelligence for health. https://www.who.int/publications/i/item/9789240078871

Yahyati, C., Lamaakal, I., Maleh, Y., El Makkaoui, K., Ouahbi, I., Almousa, M., & Abd El-Latif, A. A. (2025). A systematic review of state-of-the-art TinyML applications in healthcare, education, and transportation. IEEE Access, 13, 204513–204562. https://doi.org/10.1109/ACCESS.2025.3633575

Zambrano-de la Torre, M., Guzman-Alfaro, S., Acuña-Correa, A., Soto-Murillo, M. A., Guzmán-Fernández, M., Robles-Ortiz, R., Villagrana-Bañuelos, K. E., Arceo-Olague, J. G., Espino-Salinas, C. H., Sánchez-Reyna, A. G., & Cuevas-Rodriguez, E. O. (2026). Real-time cardiac arrhythmia classification using TinyML on ultra-low-cost microcontrollers: A feasibility study for resource-constrained environments. Bioengineering, 13(5), 532. https://doi.org/10.3390/bioengineering13050532

Zeynali, M., Alipour, K., Tarvirdizadeh, B., & Ghamari, M. (2025). Non-invasive blood glucose monitoring using PPG signals with various deep learning models and implementation using TinyML. Scientific Reports, 15, 581. https://doi.org/10.1038/s41598-024-84265-8

Apéndice A

Estrategia de búsqueda ejecutada (9 de septiembre de 2026)

La búsqueda se ejecutó el 9 de septiembre de 2026. La Tabla A1 resume la estrategia por base de datos; los archivos de consulta, exportación y bitácora se conservan en los materiales de auditoría del capítulo.

Tabla A1

Estrategia de búsqueda ejecutada por base de datos

Base de datos	Estrategia	Fecha	Registros
PubMed / MEDLINE (interfaz web + E-utilities)	Ecuación núcleo (bloque tecnología AND bloque salud) con filtros: fecha de publicación 2020-01-01 a 2026-09-09; idioma inglés o español	2026-09-09	296 identificadores; 295 registros parseados (PMID 38564569 no recuperado)
OpenAlex	Búsqueda equivalente por título y resumen con los mismos bloques conceptuales; filtros de fecha, idioma y tipo de documento	2026-09-09	735 registros exportados (estimador de interfaz: 734)
IEEE Xplore	Consulta programática	2026-09-09	Sin resultados auditables; fuente excluida con declaración explícita

Nota. Elaboración propia con base en PRISMA-S (Rethlefsen et al., 2021).

Ecuación núcleo en PubMed (texto decodificado): (“edge artificial intelligence” OR “edge AI” OR “edge intelligence” OR “on-device AI” OR “embedded machine learning” OR TinyML OR “distributed intelligence”) AND (healthcare OR medical OR biomedical OR “digital health” OR wearable OR “Internet of Medical Things” OR IoMT OR telemedicine OR “remote patient monitoring”). Bloques conceptuales: tecnología (Edge AI, on-device AI, embedded ML, TinyML, distributed intelligence); salud (healthcare, medical, biomedical, digital health, wearable, IoMT, telemedicina, remote patient monitoring). La deduplicación entre fuentes se realizó por DOI exacto (n = 140 duplicados). Antes del cribado se eliminaron automatizadamente 11 registros por tipo de publicación manifiestamente no elegible.

Apéndice B

Códigos de exclusión, ficha de extracción y evaluación

Códigos operativos de exclusión aplicados, con conteos por fase:

Tabla B1

Códigos de exclusión y conteos

Código	Motivo	Cribado (n = 577)	Texto completo (n = 43)
E1	Dominio no sanitario	101	4
E2	No es IA en el borde aplicada a la salud	173	—
E3	IA en salud sin ejecución demostrada en el borde (incluye simulación/sustitutos)	192	27 *
E4	Uso no técnico de "cutting-edge AI"	33	—
E5	Tipo de publicación no elegible	37 (+ 11 automatizados pre-cribado)	—
E6	Duplicado (DOI entre fuentes / título exacto)	41 (+ 140 DOI pre-cribado)	—
E7	Conceptual sin evaluación empírica	—	10
E8	Revisión mal clasificada	—	2

Nota. * Incluye 9 estudios reclasificados tras readjudicación estricta (simulación en OMNeT++/Cooja/datos sintéticos o procesador sustituto). Un duplicado adicional se detectó dentro del conjunto de revisiones secundarias (total 42 duplicados por título). Elaboración propia.

La ficha de extracción registró: objetivo clínico y uso previsto; población y procedencia de datos; modalidad de señal; arquitectura y ubicación de la inferencia; hardware y entorno de ejecución; tamaño del modelo, memoria pico y precisión numérica; latencia de inferencia; consumo o energía por inferencia y frontera de medición; protocolo de comunicación y volumen transmitido; comportamiento ante desconexión; mecanismos de ciberseguridad; interoperabilidad; diseño de validación; comparador; métricas clínicas; costo reportado; accesibilidad; y resultados de equidad.

Evaluación de calidad de los estudios primarios: QUADAS-2 para 10 estudios de precisión diagnóstica (selección de pacientes, prueba índice, estándar de referencia, flujo y tiempos); PROBAST para 5 modelos de predicción (participantes, predictores, desenlace y análisis); y lista estructurada de validez técnica para 71 prototipos/validaciones (descripción de muestra/datos, comparación con línea base, medición en el dispositivo, documentación de recursos, reproducibilidad y discusión de limitaciones). No se calcularon puntajes agregados; los juicios se conservaron por dominio o ítem con su justificación. La lista técnica se interpreta como evaluación de calidad/validez de ingeniería, no como herramienta clínica validada de riesgo de sesgo.

Las 5 revisiones sistemáticas del corpus secundario se consideraron con AMSTAR 2 (16 ítems; Shea et al., 2017). Una revisión fue evaluable con texto completo y obtuvo confianza críticamente baja; las cuatro restantes se registraron como no evaluables por falta de recuperación del texto completo y no recibieron una categoría global de confianza. Los juicios disponibles por ítem se conservan en los archivos de auditoría (amstar2_reviews.csv/json).

Escalas de madurez: TRL 1–9 adaptado a Edge AI sanitario (1–2 concepto; 3 prueba de concepto; 4 prototipo en laboratorio; 5 validación en entorno relevante; 6 demostración de prototipo funcional; 7 demostración en entorno operativo; 8 sistema cualificado; 9 operación sostenida) y NMV 1–5 (1 retrospectivo/benchmark; 2 validación independiente o externa; 3 prospectivo/piloto; 4 clínico prospectivo comparativo o multicéntrico; 5 implementación longitudinal). Ambas son herramientas analíticas de esta revisión.

Apéndice C

Estudios primarios incluidos (n = 86)

Referencias bibliográficas de los estudios del corpus primario, generadas a partir de los metadatos disponibles en OpenAlex y los DOI registrados. La lista se ordena alfabéticamente. Los campos bibliográficos se conservan conforme al registro de auditoría; los casos con metadatos incompletos requieren verificación bibliográfica antes de una normalización estricta a APA 7.ª edición.

Abdullah, Fatima, Z., Oropeza Rodríguez, J. L., & Sosa, O. E. (2026). EvoApneaFormer: an IoT and prognostic evolutionary deep learning-based framework for real-time multi-event sleep apnea disorder detection and remote monitoring. *Frontiers in Bioengineering and Biotechnology*, 14. https://doi.org/10.3389/fbioe.2026.1755071

Alcaraz-Ortiz, D., Zapata-Pérez, J. F., & Martinez-Alajarin, J. (2026). Cloud-to-edge deployment of optimized nnU-Net for ischemic stroke lesion segmentation on resource-constrained embedded devices. *Sensors*, 26(11), 3322. https://doi.org/10.3390/s26113322

Ali, A., Montanaro, T., Sergi, I., Carrisi, S., Galli, D., Distante, C., & Patrono, L. (2025). An innovative IoT and edge intelligence framework for monitoring elderly people using anomaly detection on data from non-wearable sensors. *Sensors*, 25(6), 1735. https://doi.org/10.3390/s25061735

Alshehri, B., Aljhani, N., Albalawi, A., Abdulghani, W., Alfawzan, T., & Alkhodair, A. J. (2026). Edge-AI enabled wearables for construction safety: Real-time physiological monitoring and localised data processing. *Future Internet*, 18(6), 293. https://doi.org/10.3390/fi18060293

An, X., Shi, S., Wang, Q., Yu, Y., & Liu, Q. (2024). Research on a lightweight arrhythmia classification model based on knowledge distillation for wearable single-lead ECG monitoring systems. *Sensors*, 24(24), 7896. https://doi.org/10.3390/s24247896

Andò, B., Baglio, S., Manenti, M., Finocchiaro, V., Marletta, V., Rajan, S., Nehary, E. A., Dibilio, V., Zappia, M., & Mostile, G. (2025). Investigating performance of an embedded machine learning solution for classifying postural behaviors. *Sensors*, 25(14), 4262. https://doi.org/10.3390/s25144262

Arthi, R., & Krishnaveni, S. (2024). Optimized tiny machine learning and explainable AI for trustable and energy-efficient fog-enabled healthcare decision support system. *International Journal of Computational Intelligence Systems*, 17(1). https://doi.org/10.1007/s44196-024-00631-4

Balaban, B. M., Sacala, I., & Petrescu-Nita, A. C. (2025). A federated cyber-physical platform for real-time coordination of emergency medical service via 112 integrations. *Journal of Control Engineering and Applied Informatics*, 27(3). https://doi.org/10.61416/ceai.v27i3.9635

Bappy, A. S., Seppänen, T., & Hoque, M. Z. (2026). SENSEYE: a resource-aware visionary framework for assisting individuals with visual disabilities. *Scientific Reports*, 16(1). https://doi.org/10.1038/s41598-026-51257-9

Brennan, D., & Galvin, P. (2024). Evaluation of a machine learning algorithm to classify ultrasonic transducer misalignment and deployment using TinyML. *Sensors*, 24(2), 560. https://doi.org/10.3390/s24020560

Cinotti, E., Gragnaniello, M., Parlato, S., Centracchio, J., Andreozzi, E., Bifulco, P., Riccio, M., & Esposito, D. (2025). An edge AI approach for low-power, real-time atrial fibrillation detection on wearable devices based on heartbeat intervals. *Sensors*, 25(23), 7244. https://doi.org/10.3390/s25237244

D, S., S, V., Ramesh Yankanchi, S., & Manasa, S. (2025). Edge computing and blockchain-based data security in IoMT. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(06), 124–140. https://doi.org/10.3991/ijoe.v21i06.54403

Dong, F., Han, C., Xu, P., Chhatwal, J., Jiang, X., Wang, T., Hsu, A., Baek, M., Wu, D., Li, R., Jiang, Y., Tian, B., Fang, J. Y., & Niu, S. (2026). A biomimetic, ultralow-power edge-AI-empowered and self-sustaining gait analysis system. *Science Advances*, 12(34). https://doi.org/10.1126/sciadv.aeh9625

Dr. G. Sophia Reena (2023). Posture guardian with smart muscle strain detection and correction using TinyML. *Tuijin Jishu/Journal of Propulsion Technology*, 44(4), 5187–5194. https://doi.org/10.52783/tjjpt.v44.i4.1873

El Mrabet, A., Tber, A., Benaly, M., Hlou, L., & El Gouri, R. (2023). Unlocking doors: A TinyML-based approach for real-time face mask detection in door lock systems. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 11(2). https://doi.org/10.52549/ijeei.v11i2.4756

Frey, S., Spacone, G., Cossettini, A., Guermandi, M., Schilk, P., Benini, L., & Kartsch, V. (2026). BioGAP-Ultra: A modular edge-AI platform for wearable multimodal biosignal acquisition and processing. *IEEE Transactions on Biomedical Circuits and Systems*, 20(3), 399–415. https://doi.org/10.1109/tbcas.2026.3652501

Ghosh, S., Bhuvanakantham, R., Sindhujaa, P., Harishita, P. B., Mohan, A., Gulyás, B., Máthé, D., & Padmanabhan, P. (2026). A cloud-aware scalable architecture for distributed edge-enabled BCI biosensor system. *Biosensors*, 16(3), 157. https://doi.org/10.3390/bios16030157

Ghosh, S., Sindhujaa, P., Senthil Kumar, P., Mohan, A., Mahalakshmi, P., Gulyás, B., Máthé, D., & Padmanabhan, P. (2026). Hybrid edge–cloud asymmetric analytics for portable multimodal BCI biosensors. *Biosensors*, 16(7), 394. https://doi.org/10.3390/bios16070394

Giordano, M., Dheman, K., & Magno, M. (2025). SepAl: Sepsis alerts on low-power wearables with digital biomarkers and on-device tiny machine learning. *IEEE Sensors Journal*, 25(5), 7858–7866. https://doi.org/10.1109/jsen.2024.3424655

Gomes, R. G., Vwalika, B., Lee, C., Willis, A., Sieniek, M., Price, J. T., Chen, C., Kasaro, M. P., Taylor, J. A., Stringer, E. M., McKinney, S. M., Sindano, N., Dahl, G. E., Goodnight, W., Gilmer, J., Chi, B. H., Lau, C., Spitz, T., Saensuksopa, T., . . . Shetty, S. (2022). A mobile-optimized artificial intelligence system for gestational age and fetal malpresentation assessment. *Communications Medicine*, 2(1). https://doi.org/10.1038/s43856-022-00194-5

Gong, E. J., & Bang, C. S. (2025). Edge artificial intelligence device in real-time endoscopy for the classification of colonic neoplasms. *Diagnostics*, 15(12), 1478. https://doi.org/10.3390/diagnostics15121478

Gong, E. J., Bang, C. S., & Lee, J. J. (2024). Edge artificial intelligence device in real-time endoscopy for classification of gastric neoplasms: Development and validation study. *Biomimetics*, 9(12), 783. https://doi.org/10.3390/biomimetics9120783

Gragnaniello, M., Borghese, A., Marrazzo, V. R., Maresca, L., Breglio, G., Irace, A., & Riccio, M. (2024). Real-time myocardial infarction detection approaches with a microcontroller-based edge-AI device. *Sensors*, 24(3), 828. https://doi.org/10.3390/s24030828

Gragnaniello, M., Marrazzo, V. R., Borghese, A., Maresca, L., Breglio, G., & Riccio, M. (2024). Edge-AI enabled wearable device for non-invasive type 1 diabetes detection using ECG signals. *Bioengineering*, 12(1), 4. https://doi.org/10.3390/bioengineering12010004

Gudiño-Ochoa, A., García-Rodríguez, J. A., Cuevas-Chávez, J. I., Ochoa-Ornelas, R., Navarrete-Guzmán, A., Vidrios-Serrano, C., & Sánchez-Arias, D. A. (2024). Enhanced diabetes detection and blood glucose prediction using TinyML-integrated e-nose and breath analysis: A novel approach combining synthetic and real-world data. *Bioengineering*, 11(11), 1065. https://doi.org/10.3390/bioengineering11111065

Gudiño-Ochoa, A., García-Rodríguez, J. A., Ochoa-Ornelas, R., Cuevas-Chávez, J. I., & Sánchez-Arias, D. A. (2024). Noninvasive diabetes detection through human breath using TinyML-powered e-nose. *Sensors*, 24(4), 1294. https://doi.org/10.3390/s24041294

He, N., Yang, L., Hu, X., & He, Y. (2026). Locally-deployed vs. cloud-based AI in healthcare: evaluating DeepSeek-R1:8b, DeepSeek-R1, and ChatGPT o3-mini-high for complex medical diagnostics. *Frontiers in Digital Health*, 8. https://doi.org/10.3389/fdgth.2026.1785443

Heo, S., Choi, T., & Choi, W. (2026). Clinical validation of an on-device AI-driven real-time human pose estimation and exercise prescription program: Prospective single-arm quasi-experimental study. *Healthcare*, 14(4), 482. https://doi.org/10.3390/healthcare14040482

Hizem, M., Aoueileyine, M. O. E., Belhaouari, S. B., EL Omri, A., & Bouallegue, R. (2025). Sustainable e-health: Energy-efficient tiny AI for epileptic seizure detection via EEG. *Biomedical Engineering and Computational Biology*, 16. https://doi.org/10.1177/11795972241283101

Hizem, M., Bousbia, L., Ben Dhiab, Y., Aoueileyine, M. O. E., & Bouallegue, R. (2025). Reliable ECG anomaly detection on edge devices for Internet of Medical Things applications. *Sensors*, 25(8), 2496. https://doi.org/10.3390/s25082496

Horesh, Y., Oz Rokach, R., Kolben, Y., & Nachman, D. (2025). Real-time monitoring of personal protective equipment adherence using on-device artificial intelligence models. *Sensors*, 25(7), 2003. https://doi.org/10.3390/s25072003

Huang, K. Y., Hsu, Y. L., Chung, C. L., Chen, H. C., Horng, M. H., Lin, C. H., Liu, C. S., & Xu, J. L. (2025). Enhancing healthcare AI stability with edge computing and machine learning for extubation prediction. *Scientific Reports*, 15(1). https://doi.org/10.1038/s41598-025-02317-z

Huang, Z., Herbozo Contreras, L. F., Leung, W. H., Yu, L., Truong, N. D., Nikpour, A., & Kavehei, O. (2024). Efficient edge-AI models for robust ECG abnormality detection on resource-constrained hardware. *Journal of Cardiovascular Translational Research*, 17(4), 879–892. https://doi.org/10.1007/s12265-024-10504-y

Huang, Z., Yu, L., Herbozo Contreras, L. F., & Kavehei, O. (2025). Efficient and secure µ-Training and µ-Fine-Tuning for edge-based TinyML with future-guided self-distillation. Machine Learning: Health, 1(1), 015002. https://doi.org/10.1088/3049-477x/add26a

Ibrar, K., Fusco, P., Rimoli, G. P., Palmieri, F., & Ficco, M. (2026). Cross-modal federated TinyML for MCU-based Internet of Medical Things. *IEEE Journal of Biomedical and Health Informatics*. https://doi.org/10.1109/jbhi.2026.3679191

Johnvictor, A. C., Poonkodi, M., Prem Sankar, N., & VS, T. (2024). TinyML-based lightweight AI healthcare mobile chatbot deployment. *Journal of Multidisciplinary Healthcare*, 17, 5091–5104. https://doi.org/10.2147/jmdh.s483247

Jun, M. (2025). Platform framework for blockchain-enhanced healthcare AIoT systems. *Frontiers in Communications and Networks*, 6. https://doi.org/10.3389/frcmn.2025.1538965

Karpagam, P., Karthikeyan, M., Kalpana, G., & Suresh, A. (2026). Edge-AI enabled secure IoT framework for real-time patient monitoring and anomaly detection in smart healthcare systems. *Scientific Reports*, 16(1). https://doi.org/10.1038/s41598-026-59602-8

Kartali, A., Koshtrevska, M., Jokić, S., Gligorić, N., & Machidon, O. M. (2026). Privacy-preserving adaptive arrhythmia monitoring with wearable ECG, smartphone PPG, and federated on-device learning. *Frontiers in Digital Health*, 8. https://doi.org/10.3389/fdgth.2026.1900115

Kasper, K. A., Thien, R., Stuart, T., Kim, J., Bhatia, A., & Gutruf, P. (2025). Wearable AI for on-device frailty assessment. *Nature Communications*, 17(1). https://doi.org/10.1038/s41467-025-67728-y

Khalid, M. I., Hussain, A., Hussain, N., & Alkhalifah, T. (2026). Lightweight and interpretable edge intelligence AI with intrusion detection for trustworthy cardiac arrhythmia in medical IoT. *Scientific Reports*, 16(1). https://doi.org/10.1038/s41598-026-43578-6

Kwon, S., Kim, H. S., Kwon, K., Kim, H., Kim, Y. S., Lee, S. H., Kwon, Y. T., Jeong, J. W., Trotti, L. M., Duarte, A., & Yeo, W. H. (2023). At-home wireless sleep monitoring patches for the clinical assessment of sleep quality and sleep apnea. *Science Advances*, 9(21). https://doi.org/10.1126/sciadv.adg9671

Lee, C., Yuh, A. H., & Kang, S. J. (2024). Real-time prediction of resident ADL using edge-based time-series ambient sound recognition. *Sensors*, 24(19), 6435. https://doi.org/10.3390/s24196435

Lee, S. H., Kim, Y. S., Yeo, M. K., Mahmood, M., Zavanelli, N., Chung, C., Heo, J. Y., Kim, Y., Jung, S. S., & Yeo, W. H. (2022). Fully portable continuous real-time auscultation with a soft wearable stethoscope designed for automated disease diagnosis. *Science Advances*, 8(21). https://doi.org/10.1126/sciadv.abo5867

Leogrande, E., Luca, E. D., & Dell'Olio, F. (2026). From PPG to blood pressure at the edge: Quantization-aware architecture selection and on-MCU validation. *Sensors*, 26(9), 2674. https://doi.org/10.3390/s26092674

Lewandowski, M., Płaczek, B., & Bernas, M. (2020). Classifier-based data transmission reduction in wearable sensor network for human activity monitoring. *Sensors*, 21(1), 85. https://doi.org/10.3390/s21010085

Lim, W. J., & Abdul Ghani, N. M. (2022). COVID-19 Mandatory self-quarantine wearable device for authority monitoring with edge AI reporting & flagging system. Health and Technology, 12(1), 215–226. https://doi.org/10.1007/s12553-021-00631-w

Lin, L., Dacal, E., Díez, N., Carmona, C., Martin Ramirez, A., Barón Argos, L., Bermejo-Peláez, D., Caballero, C., Cuadrado, D., Darias-Plasencia, O., García-Villena, J., Bakardjiev, A., Postigo, M., Recalde-Jaramillo, E., Flores-Chavez, M., Santos, A., Ledesma-Carbayo, M. J., Rubio, J. M., & Luengo-Oroz, M. (2024). Edge Artificial Intelligence (AI) for real-time automatic quantification of filariasis in mobile microscopy. *PLOS Neglected Tropical Diseases*, 18(4), e0012117. https://doi.org/10.1371/journal.pntd.0012117

Mathan.S, M., & Gupta, S. (2026). An edge AI enabled IoT cardiac decision support system using RF fuzzy feature selection and efficient deep neural networks. *International Journal of Engineering Development and Research*, 14(1). https://doi.org/10.56975/ijedr.v14i1.303588

Mohan H M, Shivaraja Kumara H C, Mallikarjun S H, & Prasad A Y (2023). Edge artificial intelligence-based facial pain recognition during myocardial infarction. *Journal of Automation, Mobile Robotics and Intelligent Systems*. https://doi.org/10.14313/jamris/3-2022/23

Mohit Garg (2025). Adaptive differential privacy in federated edge AI for medical IoT: Energy-efficient, HIPAA-compliant frameworks for distributed real-time diagnostics. *Journal of Information Systems Engineering and Management*, 10(34s), 36–44. https://doi.org/10.52783/jisem.v10i34s.5773

Motie-Shirazi, M., Nikookar, S., Ahmad, M., Rafiei, A., Sameni, R., Rohloff, P., Clifford, G. D., & Katebi, N. (2025). Real-time quality feedback on Doppler data for community midwives using edge-AI. *Machine Learning: Health*, 1(1), 015014. https://doi.org/10.1088/3049-477x/ae1bad

Mule, C. (2025). An edge AI–enabled assistive device for the visually impaired. *International Journal for Research in Applied Science and Engineering Technology*, 13(7), 725–735. https://doi.org/10.22214/ijraset.2025.73070

N., K., Gera, B. M., Sharma, B., & S., D. (2023). TinyML based deep learning model for activity detection. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(11s), 149–158. https://doi.org/10.17762/ijritcc.v11i11s.8081

Nawaz, A., Ramzan, H. H. M., Yu, X., Zou, Z., & Westerlund, T. (2026). Blockchain powered edge intelligence for U-healthcare in privacy critical and time sensitive environment. *IEEE Journal of Biomedical and Health Informatics*, 30(9), 8028–8039. https://doi.org/10.1109/jbhi.2025.3617291

Pande, P., Babu, B. M., Bhargav, P., Roy, T. L. D., Muniyandy, E., El-Ebiary, Y. A. B., & Earshia, V. D. (2025). Attention-driven hierarchical federated learning for privacy-preserving edge AI in heterogeneous IoT networks. *International Journal of Advanced Computer Science and Applications*, 16(5). https://doi.org/10.14569/ijacsa.2025.0160545

Pazienza, A., & Monte, D. (2022). Introducing the monitoring equipment mask environment. *Sensors*, 22(17), 6365. https://doi.org/10.3390/s22176365

Peruzzi, G., Galli, A., Giorgi, G., & Pozzebon, A. (2025). Sleep posture detection via embedded machine learning on a reduced set of pressure sensors. *Sensors*, 25(2), 458. https://doi.org/10.3390/s25020458

Phatcharasaksakol, B., Sittithanon, S., Pianapitham, V., Chantrapanichkul, V., Tang, J., & Chancharoen, R. (2026). Design and validation of a cyber–physical medication dispensing platform integrating edge AI verification, distributed control, and cloud synchronization. *Sensors*, 26(12), 3823. https://doi.org/10.3390/s26123823

Prabha, M., Nandhini, S., Dayanidhy, M., & Pradeep, R. (2025). Edge-AI integrated secure wireless IoT architecture for real time healthcare monitoring and federated anomaly detection. *Scientific Reports*, 16(1). https://doi.org/10.1038/s41598-025-30150-x

Priya, V., Aruna, M., Anitha, S., & Jenith Anisha, S. (2025). AI-enabled wearable system for real-time prediction of asthma exacerbation. *Iconic Research and Engineering Journals*, 9(5). https://doi.org/10.64388/irev9i5-1712170

Rincon, J. A., Guerra-Ojeda, S., Carrascosa, C., & Julian, V. (2020). An IoT and fog computing-based monitoring system for cardiovascular patients with automatic ECG classification using deep neural networks. *Sensors*, 20(24), 7353. https://doi.org/10.3390/s20247353

Saha, S. S., Sandha, S. S., Pei, S., Jain, V., Wang, Z., Li, Y., Sarker, A., & Srivastava, M. (2022). Auritus. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2), 1–34. https://doi.org/10.1145/3534586

Sahafi, A., Wang, Y., Rasmussen, C. L. M., Bollen, P., Baatrup, G., Blanes-Vidal, V., Herp, J., & Nadimi, E. S. (2022). Edge artificial intelligence wireless video capsule endoscopy. *Scientific Reports*, 12(1). https://doi.org/10.1038/s41598-022-17502-7

Samakovlis, D., Albini, S., Álvarez, R. R., Constantinescu, D. A., Schiavone, P. D., Peón-Quirós, M., & Atienza, D. (2025). BiomedBench: A benchmark suite of TinyML biomedical applications for low-power wearables. IEEE Design & Test, 42(5), 45–54. https://doi.org/10.1109/mdat.2024.3483034

Sancheti, Y., Parikh, P., & Shah, P. (2025). Intelligent dual-leg wearable for early arthritis screening via gait analysis and on-device machine learning. *International Journal of Advanced Research*, 13(08), 137–148. https://doi.org/10.21474/ijar01/21497

Shon, A., Vernam, J. T., Du, X., & Wu, W. (2026). Edge AI-based gait-phase detection for closed-loop neuromodulation in SCI mice. *Sensors*, 26(4), 1311. https://doi.org/10.3390/s26041311

Shumba, A. T., Montanaro, T., Sergi, I., Fachechi, L., De Vittorio, M., & Patrono, L. (2022). Leveraging IoT-aware technologies and AI techniques for real-time critical healthcare applications. *Sensors*, 22(19), 7675. https://doi.org/10.3390/s22197675

Sokolov, S., & Georgieva, S. (2021). Using edge AI and language understanding for predictive modeling of acute medical intoxications. *The Journal of CIEES*, 1(2), 18–22. https://doi.org/10.48149/jciees.2021.1.2.3

Suffian, M., Ieracitano, C., Mammone, N., Pascarella, A., Ferlazzo, E., & Morabito, F. C. (2026). An EEG-based edge-AI framework for Alzheimer's and Creutzfeldt–Jakob disease classification. *Sensors*, 26(10), 3274. https://doi.org/10.3390/s26103274

Thaware, P., Khurshid, A. A., Sonkusare, C., & Agrawal, Y. (2025). Leveraging fundus images for on device eye disease diagnosis with AI powered lightweight software hardware framework. *Scientific Reports*, 15(1). https://doi.org/10.1038/s41598-025-20990-y

Trivedi, D., Jain, M., Patel, D., Parve, D., & Pathak, L. (2025). Quantum-assisted secure nano-network traffic framework for real-time medical data transmission in smart hospitals. *International Journal of Innovative Science and Research Technology*. https://doi.org/10.38124/ijisrt/25apr1661

Triwiyanto, T., Caesarendra, W., Purnomo, M. H., Sułowicz, M., Wisana, I. D. G. H., Titisari, D., Lamidi, L., & Rismayani, R. (2022). Embedded machine learning using a multi-thread algorithm on a Raspberry Pi platform to improve prosthetic hand performance. *Micromachines*, 13(2), 191. https://doi.org/10.3390/mi13020191

Varmazyar, M., Jafari, A. H., Karbasi, F., & Garosi, E. (2026). An intelligent wearable insole system for machine learning-based detection of high-risk postures in static symmetrical load-lifting tasks. *Scientific Reports*. https://doi.org/10.1038/s41598-026-60297-0

Verbiest, J. R., Bonnechère, B., Saeys, W., Van de Walle, P., Truijen, S., & Meyns, P. (2023). Gait stride length estimation using embedded machine learning. *Sensors*, 23(16), 7166. https://doi.org/10.3390/s23167166

Wang, G., Liu, X., Wang, Y., Xuan, F., Zhang, B., & Wang, X. (2026). A MXene-bridged triboelectric sensor for tinyML-empowered joint biomechanics. *Nano Research*, 19(3), 94908328. https://doi.org/10.26599/nr.2026.94908328

Wiyanti, D. T., Hsu, M. H., & Nugroho, D. C. (2026). Distillation of a locally deployable 8B language model for lung histology abstraction: A 33-report proof-of-concept study. *Journal of Multidisciplinary Healthcare*, 19, 1–17. https://doi.org/10.2147/jmdh.s619855

Wu, H., Sawada, T., Goto, T., Yoneyama, T., Sasano, T., & Asada, K. (2024). Edge AI model deployed for real-time detection of atrial fibrillation risk during sinus rhythm. *Journal of Clinical Medicine*, 13(8), 2218. https://doi.org/10.3390/jcm13082218

Xie, J., Wu, Q., Dey, N., Shi, F., Sherratt, R. S., & Kuang, Y. (2025). Empowering stroke recovery with upper limb rehabilitation monitoring using TinyML based heterogeneous classifiers. *Scientific Reports*, 15(1). https://doi.org/10.1038/s41598-025-01710-y

Yahyati, C., Lamaakal, I., Makkaoui, K. E., & Ouahbi, I. (2026). A lightweight transformer with uncertainty handling for zero-shot epileptic seizure detection on TinyML edge devices. *Discover Artificial Intelligence*, 6(1). https://doi.org/10.1007/s44163-026-01656-8

Yan, J., He, Z., Ur Rehman Junejo, N., Li, Z., Grau, A., Huang, J., & Wang, C. (2024). EKFNet: edge-based Kalman filter network for real-time EEG signal denoising. *Journal of Neural Engineering*, 21(6), 066034. https://doi.org/10.1088/1741-2552/ad995a

Zaid Salah, O., Selvaperumal, S. K., & Abdulla, R. (2022). Accelerometer-based elderly fall detection system using edge artificial intelligence architecture. *International Journal of Electrical and Computer Engineering (IJECE)*, 12(4), 4430. https://doi.org/10.11591/ijece.v12i4.pp4430-4438

Zalewski, P., Marchegiani, L., Elsts, A., Piechocki, R., Craddock, I., & Fafoutis, X. (2020). From bits of data to bits of knowledge—An on-board classification framework for wearable sensing systems. *Sensors*, 20(6), 1655. https://doi.org/10.3390/s20061655

Zambrano-de la Torre, M., Guzman-Alfaro, S., Acuña-Correa, A., Soto-Murillo, M. A., Guzmán-Fernández, M., Robles-Ortiz, R., Villagrana-Bañuelos, K. E., Arceo-Olague, J. G., Espino-Salinas, C. H., Sánchez-Reyna, A. G., & Cuevas-Rodriguez, E. O. (2026). Real-time cardiac arrhythmia classification using TinyML on ultra-low-cost microcontrollers: A feasibility study for resource-constrained environments. *Bioengineering*, 13(5), 532. https://doi.org/10.3390/bioengineering13050532

Zeynali, M., Alipour, K., Tarvirdizadeh, B., & Ghamari, M. (2025). Non-invasive blood glucose monitoring using PPG signals with various deep learning models and implementation using TinyML. *Scientific Reports*, 15(1). https://doi.org/10.1038/s41598-024-84265-8

Zhu, L., Wang, C., He, Z., & Zhang, Y. (2022). A lightweight automatic sleep staging method for children using single-channel EEG based on edge artificial intelligence. *World Wide Web*, 25(5), 1883–1903. https://doi.org/10.1007/s11280-021-00983-3