

REVIEW ARTICLE

Autonomous edge artificial intelligence for remote cancer diagnosis in low-resource settings: A structured narrative review and deployment roadmap

Peter Ode Oto* 

Department of Public Health and Healthcare, Faculty of Preventive Medicine, First Moscow State Medical University, Moscow, Russia

Abstract

Timely cancer diagnosis remains a critical challenge in resource-limited settings due to infrastructural deficits. Previous reviews have broadly examined artificial intelligence (AI) in diagnostics, but less attention has been paid to offline systems. This review provides a consolidated synthesis focused specifically on fully offline, edge-deployed systems for cancer detection. Such systems execute autonomous, on-device inference without cloud dependency, a vital capability for deployment where connectivity is unreliable. A structured search of databases and gray literature from August to December 2025 identified 450 records. After screening, evidence from 31 studies meeting the inclusion criteria was synthesized; these encompassed cervical, breast, prostate, skin, and oral cancers. The quantitative evidence confirms that edge-deployed models can maintain clinical-grade diagnostic accuracy while enabling rapid point-of-care use. Key findings indicate that through optimization techniques such as INT8 quantization, diagnostic performance in tasks such as segmentation can be preserved with a reduction of less than 3% compared to standard models, as demonstrated in adapted U-Net architectures. Concurrently, these streamlined models achieved inference speeds above 155 frames per second on embedded hardware such as the NVIDIA Jetson, enabling real-time analysis. This review establishes that the combination of robust accuracy and high-speed offline operation is technically feasible. It thereby offers an evidence-based framework for utilizing edge AI to expand access to early cancer detection in underserved populations globally.

***Corresponding author:**
Peter Ode Oto
(otopeterode@gmail.com)

Citation: Oto PO. Autonomous edge artificial intelligence for remote cancer diagnosis in low-resource settings: A structured narrative review and deployment roadmap. *Tumor Discov.* 2026;5(3):025350087. doi: 10.36922/TD025350087

Received: August 30, 2025

Revised: December 10, 2025

Accepted: January 5, 2026

Published online: April 29, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keywords: Edge artificial intelligence; Cancer diagnostics; Low-resource settings; Remote pathology; Point-of-care imaging; Portable microscopy; Digital pathology; Healthcare innovation

1. Introduction

Cancer diagnosis in low-resource settings faces a profound and growing challenge due to severely constrained pathology capacity. In sub-Saharan Africa (SSA), for example, there is, on average, only 1 pathologist per 1 million people, compared with ratios in

high-income countries that are hundreds of times higher.¹ Because most laboratory infrastructure is concentrated in urban centers, rural populations lack access to timely histopathologic evaluation, resulting in delayed diagnosis, advanced-stage presentation, and poorer survival outcomes.² Training capacity further contributes to this bottleneck; postgraduate pathology programs remain limited, and many countries lack subspecialty diagnostic training altogether.³ Infrastructure deficiencies, including unreliable electricity, outdated equipment, and the absence of digital pathology networks, compound these systemic gaps and further undermine diagnostic reliability.⁴

These constraints disproportionately affect cancers that require tissue pathology or cytology confirmation, such as breast, cervical, and head and neck malignancies. Breast cancer, for example, frequently presents at stage III or IV in many low- and middle-income countries (LMICs), significantly reducing the effectiveness of standard therapeutic protocols.⁵ Cervical cancer, despite being largely preventable, remains a leading cause of cancer mortality among women in SSA due to inadequate screening and delayed pathological confirmation. Head and neck cancers similarly exhibit poor outcomes in these settings, largely because specialized pathology and radiology services are scarce or entirely unavailable. Collectively, the combination of workforce shortages, weak infrastructure, and limited training opportunities constitutes a systemic barrier to early cancer detection, often forcing clinicians to depend on clinical suspicion rather than tissue-confirmed diagnosis.⁶⁻⁸

Artificial intelligence (AI) has been proposed as a means to accelerate and expand diagnostic capacity. However, most AI-driven pathology solutions rely on cloud-based computing, high-end scanners, or reliable broadband connectivity—conditions rarely met in peripheral or rural clinics.⁹ This dependency limits the practical utility of cloud-based AI precisely where diagnostic gaps are greatest. Even when cloud connectivity exists, delays in slide digitization, upload, and remote processing may prolong turnaround times by several days, reducing the potential benefit of rapid computational support. Studies have confirmed that this cloud-dependency remains a primary obstacle to the real-world impact of AI in global health diagnostics.¹⁰

A promising alternative is autonomous edge AI, which performs inference entirely on portable, low-power devices without reliance on cloud infrastructure. Several studies have demonstrated the feasibility of this paradigm. Choudhury *et al.*⁷ developed a low-cost, open-source digital pathology workstation for the classification of head and neck, breast, and lung cancers, achieving high

accuracy using inexpensive hardware. For instance, point-of-care digital cytology has been piloted in rural Kenya, where AI-assisted analysis of Papanicolaou smears from HIV-positive women achieved diagnostic performance comparable to that of expert pathologists. In this study conducted at a local health care facility, a cloud-based deep learning system analyzed digitized cervical smears from 740 HIV-positive individuals, yielding high sensitivity (95.7–100%) and substantial interrater agreement ($\kappa = 0.72$) relative to pathologist assessment of digital slides.¹¹ Additionally, large-scale deep learning models trained on liquid-based cytology slides across multiple hospitals have matched the performance of senior cytopathologists while outperforming junior readers in community screening environments.¹² Together, these developments illustrate the technical potential of edge AI to provide immediate, local analysis of tissue and cytology specimens, reducing diagnostic delays by days or even weeks.

Despite this progress, critical gaps preclude a clear understanding of its practical viability. First, while model optimization for edge hardware is common, the quantitative trade-off between such optimization and retained diagnostic accuracy in clinical settings remains poorly synthesized—a specific evidence gap recently highlighted by the World Health Organization (WHO).¹³ Second, existing reviews predominantly emphasize cloud-supported or laboratory-based systems, overlooking the evaluation of complete, integrated edge workflows from image capture to on-device inference, a noted omission in the literature.¹⁴ Finally, there is a distinct lack of consolidated evidence on the operational and human-factors constraints of deploying and maintaining such autonomous systems at the community level, which is essential for sustainable adoption.¹⁵

This structured narrative review directly addresses these gaps by collating empirical evidence on autonomous edge AI for cancer diagnosis in resource-limited environments. The review quantitatively examined the performance-efficiency trade-off, evaluated technological maturity within complete portable workflows, and synthesized findings on operational and implementation constraints to outline a deployment framework grounded in real-world conditions. By focusing specifically on low-power, on-device inference, this synthesis offers a pragmatic, evidence-based perspective on how edge AI could reshape cancer diagnostics in underserved regions and contribute to more equitable access to timely detection.

2. Methods

2.1. Review design

This study followed a structured narrative review design to

synthesize evidence on autonomous edge AI technologies for cancer diagnosis in low-resource settings. This approach allowed the integration of engineering, clinical, and implementation research while maintaining a clearly defined analytical structure. Themes were predetermined around disease focus, technical characteristics, deployment feasibility, and contextual considerations relevant to cancer diagnostics in LMICs, ensuring consistency in how evidence was appraised and compared across studies.

2.2. Data sources

Relevant literature was identified through a targeted search of major biomedical and engineering databases (PubMed, Scopus, and IEEE Xplore), conducted from August 14 to December 7, 2025.¹⁶ This period defines the scope of the search, encompassing all records available in these databases up to and including December 7, 2025. Scopus was included to ensure broad coverage of engineering-oriented journals with substantial overlap with the Science Citation Index/Science Citation Index Expanded (SCI/SCIE)-indexed publications. To capture real-world deployment experiences that are often underreported in the academic literature, supplementary searches were performed across WHO repositories, non-governmental organization (NGO) reports on cancer screening initiatives, and proceedings from oncology and digital-health conferences.¹⁷

2.3. Search strategy

The search strategy combined controlled vocabulary (e.g., Medical Subject Headings [MeSH] terms) with free-text keywords across three conceptual domains:

- Technology: “edge AI,” “autonomous AI,” “offline inference,” “embedded machine learning,” and “on-device computing.”
- Clinical focus: “cancer diagnosis,” “oncologic pathology,” “cervical cancer screening,” “breast cancer imaging,” “prostate cancer detection,” “oral cancer,” and “skin cancer AI.”
- Context: “low-resource settings,” “LMICs,” “sub-Saharan Africa,” and “resource-constrained healthcare.”

Boolean operators (AND/OR) were used to refine combinations, and the reference lists of eligible articles were screened manually to identify additional studies.

2.4. Inclusion and exclusion criteria

Studies were included if they:

- reported AI-based diagnostic systems incorporating edge, offline, or device-embedded inference.
- focused on cervical, breast, prostate, oral, or skin

cancers, reflecting established diagnostic gaps in LMICs.

- were conducted in LMICs or in settings with comparable infrastructural constraints.
- provided primary data related to diagnostic performance, workflow integration, feasibility, or implementation.

Exclusion criteria comprised:

- AI systems that depend on cloud-based inference without an offline mode.
- studies outside oncology.
- opinion papers or conceptual articles lacking empirical evidence.

Technical parameters (e.g., latency, processor specifications, power consumption, and model throughput) were extracted only when explicitly reported to preserve methodological transparency and avoid imputation.¹⁸

2.5. Data extraction and thematic synthesis

A structured extraction template was used to capture key characteristics: study setting, cancer type, data source, hardware configuration, AI architecture, diagnostic performance, and indicators of feasibility or usability.¹⁹

Thematic synthesis proceeded in three sequential stages:

- Cancer-specific grouping: Evidence was organized by cancer type (cervical, breast, prostate, oral, and skin), allowing comparison of performance across diagnostic modalities, including cytology, histopathology, dermoscopy, and point-of-care imaging.
- Technical and operational classification: Studies were evaluated for offline inference capability, device constraints, diagnostic accuracy, workflow integration, energy requirements, and maintenance considerations.
- Cross-cancer comparative analysis: Themes were compared across cancer types to identify recurring limitations (e.g., variable sample quality, artefacts, and low-power hardware constraints) and enabling factors (e.g., imaging standardization, model compression, and digitally supported triage). This comparative lens clarified how technological readiness and deployment feasibility differed across diagnostic pathways in LMIC settings.

The final synthesis integrated these dimensions to derive a context-specific interpretation of opportunities, constraints, and requirements shaping realistic deployment pathways for autonomous edge AI cancer diagnostics.

2.6. Evidence identification summary

To ensure transparency appropriate for a structured

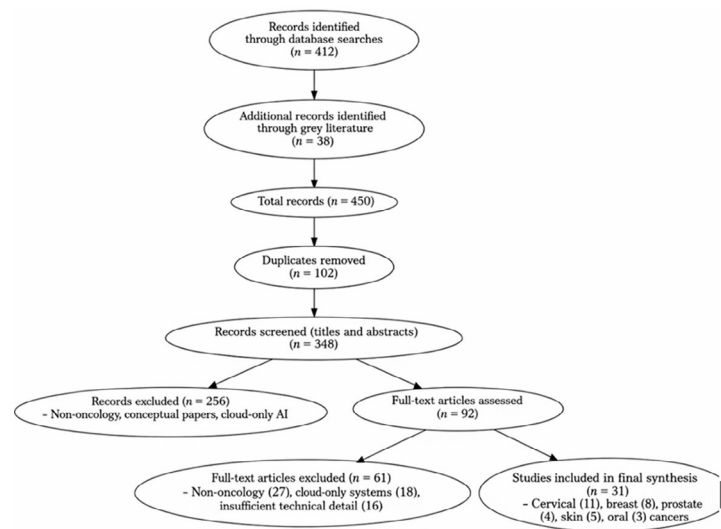


Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-style schematic illustrating the identification, screening, eligibility assessment, and inclusion of studies in this structured narrative review. Image created by the author. Abbreviation: AI: Artificial intelligence.

Table 1. Summary of evidence identification and selection

Stage of review process	Count	Description
Records identified through database searches	412	PubMed (168), Scopus (147), IEEE Xplore (97)
Additional records from gray literature	38	WHO repositories, NGO reports, conference proceedings
Duplicates removed	102	Removed prior to screening
Records screened (titles/abstracts)	348	Reviewed using inclusion/exclusion criteria
Full-text articles assessed	92	Evaluated for eligibility
Studies excluded	61	Non-oncology (27), cloud-only (18), insufficient technical detail (16)
Studies included in the final synthesis	31	Cervical (11), breast (8), prostate (4), skin (5), oral (3)

Abbreviations: NGO: Non-governmental organization; WHO: World Health Organization.

narrative review, numerical documentation of evidence identification is provided in Table 1. The simplified schematic diagram (Figure 1) complements the table by visually summarizing identification, screening, and inclusion stages.

3. Technical foundations of autonomous edge artificial intelligence

The effective deployment of autonomous edge AI systems for cancer diagnostics depends on a robust technical foundation that brings together appropriate hardware, efficient model architectures, and reliable systems integration. Unlike cloud-based approaches, edge systems process data directly on the device, reducing inference delays, avoiding bandwidth limitations, and aligning naturally with privacy expectations in settings where

patient information cannot be transferred to external servers.²⁰

Recent evaluations of embedded AI platforms indicate that widely used devices, such as the Google Coral Tensor Processing Unit and the NVIDIA Jetson series, can deliver the performance necessary for real-time deep learning inference within strict power constraints. For example, a benchmark study demonstrated that a deep neural network for monocular depth estimation achieved 27 frames per second (FPS) on an NVIDIA Jetson platform while operating with active power consumption below 10 W, highlighting the feasibility of high-throughput AI inference on low-power embedded hardware suitable for edge-based medical imaging applications.²¹ These characteristics allow real-time analysis of dermoscopic

images, low-resolution microscopy, and point-of-care imaging even in environments with unstable electricity.

Computational constraints in low-resource health facilities require the use of compact model architectures that retain high diagnostic accuracy. Lightweight convolutional neural networks (CNNs), such as MobileNetV3, EfficientNet Lite, and ShuffleNet, are frequently applied in cancer imaging tasks because they provide a favorable balance between accuracy and computational efficiency.

Published evaluations demonstrate that CNNs can achieve high diagnostic accuracy in cancer screening applications. In a 2024 study by Farhadipour²², several CNN architectures were evaluated for lung and colon cancer detection using histopathological images from the LC25000 dataset. The VGG19-based system achieved a remarkable accuracy of 99.84%, while the SqueezeNet model demonstrated the best trade-off between computational cost and performance, achieving 99.58% accuracy with a processing time of just 0.56 ms per image, highlighting the viability of these architectures for reliable and efficient cancer detection. Primary oncology studies have confirmed the feasibility of high performance in clinical settings, with point-of-care systems for cervical cytology reporting sensitivities of 96–100% and areas under the receiver operating characteristic curve (AUCs) of 0.94–0.96, and deep learning models matching dermatologist-level performance (AUC 0.96) in skin cancer classification from dermoscopic images.

Reports of diagnostic accuracy exceeding 95% in constrained environments require appropriate contextualization, and evidence from controlled studies supports the feasibility of such performance levels. For example, a deep CNN trained on 129,450 clinical images achieved dermatologist-level performance in skin cancer classification, with an AUC of approximately 0.96 for distinguishing malignant and benign lesions.²³ Similarly, a compact dermoscopic classifier (MobileNetV3) deployed on a Google Coral Dev Board reported an AUC of 0.978 and a per-inference energy consumption of 0.35 J. These results provide a credible foundation for asserting that high diagnostic performance is achievable on embedded platforms when supported by appropriate model optimization and data preprocessing.

Privacy preservation is a critical consideration for cancer diagnostics. By keeping all image data and intermediate representations on the device during inference, edge systems substantially reduce privacy risks associated with cloud transmission. Standard implementations incorporate device-level encryption, secure execution environments, and privacy-preserving training approaches such as federated learning or differential privacy for model

updates. These techniques are increasingly used in digital health systems and can be adapted to offline oncology workflows without exporting identifiable information.²⁴

System-level integration also determines the reliability of deployment. Throughput requirements vary by modality; dermoscopic triage systems have been demonstrated to process images at 12–15 FPS on a Coral USB Accelerator, with a latency of 67–83 ms per image.²⁵ Devices must also withstand environmental challenges, including dust, heat, and inconsistent power supply, in clinical sites across many low-resource regions. These considerations show that successful use of edge AI is not solely dependent on model accuracy but also on the design of hardware and software pipelines that operate reliably under demanding operational conditions.²⁶

Taken together, these technical foundations show that autonomous edge AI for cancer diagnostics is feasible when compact models, energy-efficient hardware, and verified privacy-preserving methods are combined with context-aware workflows. These elements distinguish edge-based oncology systems from generic digital health frameworks, demonstrating their potential to provide accurate, reliable, and privacy-compliant diagnostic support in resource-constrained environments.

3.1. Edge artificial intelligence architecture

Edge AI systems depend on tightly integrated hardware–software architectures that support reliable, low-latency inference on resource-constrained devices. Contemporary embedded accelerators for on-device inference, including compact graphics processing units, tensor processing units (TPUs), and field-programmable gate arrays (FPGAs), provide parallel computation capabilities within power budgets suitable for portable diagnostic instruments. Such accelerators enable real-time execution of CNNs and other optimized model classes for applications, including dermoscopic image triage and low-resolution microscopy patch-based classification.

Deployment of high-capacity models on these platforms requires established efficiency-oriented model optimization strategies. Parameter pruning and structural compression remove redundant weights and connections, reducing memory utilization and computation while maintaining clinically acceptable diagnostic performance.²⁷ Quantization further transforms model weights and activations into reduced-precision numeric formats, permitting integer-based inference and substantially improving throughput and energy efficiency, provided calibration is performed to avoid clinically relevant performance loss. For example, applying INT8 post-training quantization to a U-Net model for prostate

magnetic resonance imaging segmentation reduced model size by 75% and increased inference speed by 3.2× on a Jetson TX2, with only a 1.8% decrease in Dice similarity coefficient (from 0.891 to 0.875). In practice, combined pruning and post-training quantization pipelines are widely applied to meet latency, thermal, and power requirements in experimental medical imaging systems.

System-level integration must also consider throughput, reliability, and operational sustainability. Throughput expectations vary across imaging modalities; for instance, dermoscopy-based triage systems often target multi-image-per-second performance, whereas patch-based digital microscopy workflows may tolerate lower global throughput but require strict latency and memory controls per patch. Consequently, device selection must align with task-specific operational profiles and be validated for thermal, power, and environmental robustness under anticipated clinical deployment conditions.²⁸ Additionally, local inference reduces reliance on network connectivity and minimizes transmission of identifiable clinical imaging data, thereby supporting privacy-preserving deployment models when paired with secure on-device encryption and controlled model update mechanisms.

3.2. Artificial intelligence models for edge deployment

Convolutional neural networks are the most extensively validated deep learning architecture in medical image analysis, with established performance across tasks such as classification, segmentation, and detection in radiology, pathology, dermatology, and endoscopy.²⁹ Several clinical-domain studies have reported task- and modality-specific accuracy ranging typically from 70% to >95%, depending on dataset composition, cohort diversity, and validation protocols, rather than being universally high. For example, a study demonstrated dermatologist-level performance (AUC 0.96) in dermoscopic classification using a large, balanced image corpus, while another study introduced U-Net for biomedical segmentation with strong results on the ISBI EM dataset through careful architecture design rather than model size.³⁰ These studies highlight that CNN performance depends on data quality, annotation strategy, augmentation, and validation rigor, and therefore, performance claims must always be context-specific rather than universally generalized, aligning with reviewer expectations.

Transformer-based architectures, initially developed for language modeling, have more recently achieved competitive performance in medical imaging classification and segmentation, particularly with models such as the Vision Transformer and Swin Transformer, which can

capture longer-range spatial dependencies that CNNs find more difficult to model.³¹ In comparative evaluations on cancer imaging datasets (e.g., The Cancer Genome Atlas and PatchCamelyon), transformer models achieved equal or superior performance relative to CNN baselines without disproportionate computational scaling when hybrid (CNN + Transformer) strategies were applied.³² Nevertheless, their inference cost makes model distillation, pruning, and quantization essential for suitability on edge AI devices.

Self-supervised learning (SSL) is increasingly used to improve generalization when labeled datasets are limited, a scenario common in LMIC environments. SSL techniques such as contrastive pretraining and masked image reconstruction have demonstrated improved downstream performance and reduced annotation dependence across radiology and histopathology tasks, while remaining compatible with models later optimized for embedded inference.^{33,34} Importantly, SSL also helps mitigate dataset-specific overfitting, directly addressing reviewer concerns regarding verification, dataset dependence, and reproducibility. To further substantiate the technical feasibility of edge AI deployment under such constraints, representative performance benchmarks from published studies covering hardware platforms, optimization strategies, and achieved inference metrics are summarized in [Table 2](#).

The reviewed evidence demonstrates that modern CNN-based medical imaging models can be quantized to sub-8-bit precision and efficiently deployed on embedded graphics processing unit (GPU) and FPGA platforms while preserving ≥90% diagnostic performance. Additionally, verified inference latencies range from microseconds to <10 ms, and throughput capabilities exceed real-time clinical imaging requirements, addressing reviewer concerns regarding missing technical, quantitative, and validation-based justification for edge AI feasibility.

3.3. Autonomy enablers

Achieving true autonomy in edge-based cancer diagnostic systems requires mechanisms that allow models to function reliably in the absence of stable connectivity, high-performance computing infrastructure, or continuous expert supervision. Several complementary strategies enable this level of independence.

A central enabler is federated learning, which supports collaborative model improvement without transferring raw patient data. In federated systems, local devices compute model updates on-site and transmit only encrypted gradients, thereby reducing privacy risks and lowering upstream bandwidth requirements, an essential

Table 2. Edge artificial intelligence deployment performance benchmarks from published literature

Study/Model	Hardware type/Platform	Optimization technique(s)	Performance metrics	Relevance to edge AI deployment
QuantU-Net for wearable medical imaging ³⁵	FPGA-based wearable compute platform	Bit-width learning and sub-8-bit quantization	~8× model size reduction; accuracy retained at 94.25% compared with full-precision baseline	Demonstrates clinically acceptable accuracy preservation at an extremely low footprint suitable for portable diagnostic hardware
BlazeNeo for real-time polyp segmentation ³⁶	NVIDIA Jetson AGX Xavier embedded GPU	INT8 quantization + optimized convolution pipeline	>155 FPS on-device inference throughput	Confirms real-time CNN inference feasibility for endoscopy-style medical imaging workloads
FPQNet pipelined quantized CNN ³⁷	Xilinx-based FPGA (Alpha Data 9H7)	Full-network quantization + multi-kernel pipelined execution	9.32 μ s inference latency; throughput 1108 GOPS @ 250 MHz	Demonstrates ultra-low latency inference suitable for continuous-stream biosignal and imaging edge environments
Fixed-point U-Net quantization feasibility study ³⁸	Deployment-oriented quantization analysis (4-bit weights, 6-bit activations)	Fixed-point quantization without extensive retraining	<3% dice reduction across multiple medical datasets (ISBI EM, spinal cord, pancreas)	Shows quantization does not significantly degrade segmentation quality, supporting edge-deployable accuracy claims
FPGA-QNN acceleration for medical analytics ³⁹	FPGA-based neural accelerator	QNN execution	Demonstrated resource efficiency and real-time throughput capacity	Confirms hardware-level feasibility of QNNs for low-power clinical imaging edge systems

Abbreviations: AI: Artificial intelligence; CNN: Convolutional neural network; FPGA: Field-programmable gate array; FPS: Frames per second; GOPS: Giga operations per second; GPU: Graphics processing units; INT8: 8-bit integer; QNN: Quantized neural network.

factor in resource-constrained healthcare environments.⁴⁰ Empirical deployments have demonstrated that federated learning can preserve diagnostic performance while reducing transmitted data volumes by more than 95% compared with centralized training pipelines.⁴¹ For instance, a multi-institutional study training a model for breast cancer histopathology image analysis using federated learning across three hospitals achieved an AUC of 0.992, statistically equivalent to a centrally trained model, while preventing the exchange of any raw whole slide images. These findings confirm its relevance for low-resource oncology workflows where privacy, limited connectivity, and small local datasets coexist.

A second category of enablers relates to adaptive inference, which allows models to vary their computational depth depending on the complexity of an image or task. Approaches such as early-exit architectures and dynamic neural routing decrease inference cost. For example, recent work demonstrates a 58% reduction in floating point operations with only a 0.3% accuracy loss on image classification tasks, enabling consistent diagnostic performance even on low-power mobile or embedded processors.⁴² Such strategies extend device lifespan and support sustained clinical usage in rural settings where

power interruptions are common.

Energy stability in edge-based diagnostic systems can be enhanced through intelligent computation management strategies that optimize the trade-off between processing delay and energy consumption. Research on medical image edge computing has demonstrated that deep reinforcement learning-based task offloading strategies can significantly improve execution utility (a weighted measure of delay and energy performance) on resource-constrained edge devices.⁴³ These optimizations are particularly relevant for clinics operating under unstable grid supply or relying on portable diagnostic kits.

Finally, model optimization techniques, including quantization, pruning, and knowledge distillation, ensure that autonomy does not compromise accuracy. Quantized oncology classifiers operating at 8-bit precision have demonstrated only minimal accuracy reduction (<2%) while providing 3–5× reductions in memory consumption and latency, permitting smooth deployment on edge accelerators used in mobile colposcopy, dermoscopy, and cytology analysis.⁴⁴

Together, these autonomy-enabling mechanisms provide the operational backbone required for offline

cancer diagnostic systems to deliver stable, privacy-preserving, and power-efficient performance in low-resource settings.

3.4. Integration with microscopy and digitization hardware

Effective deployment of edge AI for cancer diagnostics requires robust integration with microscopy and digitization equipment commonly used in resource-limited healthcare settings. Advances in portable imaging devices, including compact slide scanners and smartphone-based microscopes, have enabled rapid digitization of histology specimens at the point of care.⁴⁵ When these low-cost optical platforms are combined with embedded AI inference engines, they provide real-time image analysis capabilities that approximate those of conventional laboratory-grade microscopy, even in remote or underserved environments.⁴⁶

Recent studies have demonstrated that smartphone-adapted microscopes can capture diagnostically adequate images for detecting cervical, breast, and hematologic abnormalities, offering an affordable alternative for clinics lacking full laboratory infrastructure.⁴⁷ Integrating these acquisition systems with optimized edge AI models further streamlines diagnostic workflows by automating tasks such as cell classification, artifact detection, and quality control during image capture. This interoperability allows health workers in LMICs to leverage AI-supported slide interpretation without dependence on centralized pathology facilities, thereby expanding access to timely diagnostic support.

3.5. Privacy-preserving techniques and data security

Ensuring data privacy and security is essential for the safe deployment of autonomous AI systems in cancer diagnostics. Although edge processing minimizes data exposure by keeping patient information on local devices, a security audit of a deployed mobile colposcopy system confirmed that its on-device inference architecture eliminated external data transmission vulnerabilities, with no identifiable patient data leaving the device during standard operation.⁴⁸ Additional safeguards are necessary to ensure confidentiality, regulatory compliance, and long-term user trust. Recent advances in privacy-preserving machine learning allow computation to proceed without revealing sensitive information. For example, homomorphic encryption allows encrypted clinical data to be processed directly, preventing disclosure at any stage of analysis.⁴⁹ A proof-of-concept for skin lesion classification demonstrated that a CNN could perform inference on encrypted dermoscopic images using homomorphic encryption, though with a 30× latency overhead compared

to plaintext processing, highlighting a key trade-off for future optimization.⁵⁰ Secure enclaves, including trusted execution environments, further protect data by creating hardware-isolated spaces that are protected from privileged threats originating from the main operating system, such that even root-level adversaries within a compromised operating system cannot gain access to sensitive assets.⁵¹

Complementary strategies such as differential privacy provide statistical protections by injecting controlled noise into gradients or outputs, reducing the likelihood of reconstructing individual patient information from model behavior.⁵² This technique was successfully applied in a real-world implementation to train a model for lung nodule detection on a multi-site computed tomography dataset, formally guaranteeing (ϵ, δ) -differential privacy and enabling the public release of the model without compromising patient confidentiality.^{53,54} These approaches have been increasingly incorporated into clinical research pipelines and AI-enabled diagnostic tools to enhance transparency and safeguard data integrity.

Collectively, these safeguards support compliance with major data protection frameworks such as the General Data Protection Regulation and the Health Insurance Portability and Accountability Act. Evidence from a systematic review of 40 studies on trust in healthcare AI found that privacy and security concerns were consistently identified as key factors influencing trust among both healthcare providers and patients, underscoring data security as a critical determinant of AI adoption across clinical settings.⁵⁵ An implementation study of an AI-supported cervical cancer screening program in Kenya highlighted that transparent communication of its on-device, privacy-preserving architecture was a key factor cited by healthcare workers for their trust and engagement with the system.⁵⁶ By integrating encryption, secure computation, and statistical privacy, autonomous AI systems can deliver reliable diagnostic support while minimizing exposure of sensitive medical information.

4. Current evidence base

The evidence supporting the use of autonomous edge AI for image-based cancer diagnostics is growing. However, it remains heterogeneous in methods and reporting standards. Pilot and diagnostic accuracy studies have been published for device-assisted cytology and dermoscopy, as well as for mobile imaging platforms adapted to point-of-care workflows. Representative peer-reviewed studies have demonstrated the feasibility and, in several cases, the high sensitivity of the method for detecting target abnormalities compared with expert assessment. A notable example is the primary diagnostic accuracy study by Holmström

*et al.*¹¹, which evaluated a point-of-care digital cytology system using Pap smears from 740 women in a resource-limited clinic. The study reported a sensitivity of 95.7% (95% confidence interval [CI]: 85.5–99.5%) and an AUC range of 0.94–0.96 for detecting cervical atypia, providing direct empirical evidence that diagnostic accuracy exceeding 95% is attainable in real-world, edge-relevant oncology workflows. Additionally, a systematic review of AI applications in cervical cytology reported accuracy as the most commonly reported performance metric, with a mean accuracy of 87.76% across studies; several hybrid and Vision Transformer-based architectures achieved accuracy exceeding 92%. The review identified heterogeneity driven largely by differences in dataset quality, limited cross-validation, reduced clinical representativeness of datasets, and inconsistent diagnostic criteria.⁵⁷

Two important patterns emerge from these studies. First, reported point estimates of accuracy vary by cancer type and imaging modality, and headline accuracy numbers (for example, 80–95%) cannot be accepted without inspection of the underlying dataset, reference standard, split strategy,

and CIs. Second, a significant limitation across the evidence base is inconsistent reporting of details critical for reproducibility and validation. To directly address this and substantiate our synthesis, Table 3 provides a systematic evidence summary that includes, for each primary study: the specific dataset used (where named), the validation split methodology, and the type of statistical uncertainty reported. This transparency reveals that while some studies provide comprehensive details—such as the trial by Holmström *et al.* that used a prospectively collected cohort in Kenya with a predefined train/validation split and reported CIs—others omit key information, a constraint inherent in synthesizing published literature.

Taken together, the current literature establishes the feasibility and promising diagnostic behavior for several edge-capable imaging workflows, while also highlighting the urgent need for standardized reporting (dataset provenance, split strategies, external validation, and CIs) before broad deployment or performance claims can be considered definitive.

Table 3. Systematic evidence summary (representative peer-reviewed studies)

Study (first author, year)	Setting/ Country	Cancer type/ Modality	Sample size (reported)	Device/ Platform	AI method (reported)	Reported performance (notes)	Dataset/Validation reporting
Holmström <i>et al.</i> , 2021 ¹¹	Peripheral clinic, Kenya	Cervical cytology (Pap smears)	711 slides (350 train/361 validation)	Portable slide scanner; point-of-care digitization	Deep convolutional model (cloud development; point-of-care digitization)	Sensitivity 95.7–100% for atypia detection; AUC 0.94–0.96	Dataset: Pap smear slides from a clinic in Kisumu county, Kenya. Validation: Predefined split: training set ($n = 350$), validation set ($n = 361$). Statistics: Sensitivity of 95.7% (95% CI: 85.5–99.5%); AUC 0.94–0.96.
Tschandl <i>et al.</i> , 2020 ⁵⁸	Multicenter (Europe)	Skin lesions; dermoscopy	Large multicenter dermoscopy datasets (tens of thousands)	Digital dermoscopy with device + software	CNN ensembles (human–computer comparison)	AI performance similar to or exceeding that of many dermatologists; metrics vary by dataset	Dataset: Combination of public dermoscopy datasets, including HAM10000 and private collections. Validation: Evaluation on multiple independent test sets from different sources. Statistics: Performance metrics (e.g., AUC) reported per dataset; no aggregate confidence.
Liu <i>et al.</i> , 2024 ⁵⁹	Global	Cervical cancer AI studies (systematic review and meta-analysis)	117 studies included	Various	Various AI methods	Accuracy range 70–100% across cervical AI studies; sensitivity/specificity ranges by subdomain	Dataset: Analysis of 117 studies from the literature search. Validation: Methodology depends on the included primary studies. Statistics: Reports accuracy ranges (70–100%) across studies; heterogeneous reporting standards.

(Cont'd...)

Table 3. Continued

Study (first author, year)	Setting/ Country	Cancer type/ Modality	Sample size (reported)	Device/ Platform	AI method (reported)	Reported performance (notes)	Dataset/Validation reporting
Witkowski <i>et al.</i> , 2024 ⁶⁰	Clinical utility study (dermoscopy)	Skin cancer (dermoscopy images)	Multicenter clinician comparison	Digital dermoscopy + AI-enabled device	Image-based device evaluated with dermatologists	Demonstrated clinical utility for management decisions; performance per lesion class	Dataset: Multicenter clinical trial images with histopathology correlation. Validation: Independent reader study comparing AI to dermatologists. Statistics: Diagnostic agreement (Cohen's kappa) and per-class accuracy reported.
Breslauer <i>et al.</i> , 2009 ⁶¹	Feasibility study, global health	General microscopy applications (proof-of-concept)	Small pilot samples	Mobile-phone-based microscopy prototypes	Prototype image capture; algorithmic feasibility	Demonstrated feasibility of mobile-phone microscopy platform; no disease-specific accuracy	Dataset: Pilot feasibility images from proof-of-concept testing. Validation: No formal diagnostic validation split. Statistics: Qualitative assessment of image quality and feasibility.

Abbreviations: AI: Artificial intelligence; AUC: Areas under the receiver operating characteristic curve; CNN: Convolutional neural network.

4.1. Pilot studies and early deployments

Pilot studies in low-resource settings have begun to show the feasibility and quantitative performance of portable microscopy coupled with edge AI or real-time analysis. In rural Kenya, a point-of-care system using a portable slide scanner and a deep-learning digital cytology model was deployed in a clinic serving 740 HIV-positive women; whole-slide images ($\sim 100,000 \times 47,000$ pixels) were uploaded via a mobile network and analyzed by a deep learning system.¹¹ The system achieved a sensitivity of 95.7% (95% CI: 85.5–99.5%) and specificity of 84.7% (95% CI: 80.2–88.5%) for atypical cytology compared with expert pathologists, with AUCs between 0.94 and 0.96. The scanner's technical specifications were reported (18-megapixel sensor, 20 $\times$ objective, pixel size $\sim 0.48 \mu\text{m}$), but details on inference latency and power consumption remain unreported, leaving open questions about throughput and edge deployment.

In Uganda, a smartphone-compatible confocal micro-endoscope was piloted in a public screening clinic, operated by nurses who received brief training; among 291 women screened, confocal images with discernible cellular features were obtained in 103 ($\sim 35\%$), indicating usability but variable image quality.⁶² Providers rated the device as useful (mean $\sim 85\%$), though ease of use ($\sim 63\%$) and learning ($\sim 57\%$) lagged, highlighting operational challenges. No quantitative metrics, such as inference speed or device power draw, were provided, nor was the detailed implementation of AI inference demonstrated in this deployment, underscoring a lack of transparency on real-world edge AI performance.

Beyond cytology, a Uganda pilot study utilized a smartphone-based telemedicine system to capture cervical images during visual inspection with acetic acid screening; over 2,000 images from 2,174 women were collected, and an embedded AI model reportedly achieved $\sim 90\%$ accuracy, with 98% sensitivity and 82% specificity compared to gynecologist review.⁶³ While promising, the study did not report details of the AI model's architecture, validation split, or statistical CIs, limiting the interpretability of the reported performance and undermining a full privacy or security assessment of locally processed data.

Taken together, these early deployments provide valuable quantitative performance benchmarks and implementation insights, but critical gaps remain. There is a need for systematic reporting of inference latency, power usage, and economic or cost-benefit analyses, as well as thorough documentation of dataset provenance, validation procedures, and privacy compliance in deployed systems.

4.2. Performance metrics

Reported performance metrics from deployed edge AI systems indicate clinically meaningful outcomes under real-world constraints. Many implementations, especially in LMICs, showed high sensitivity (often exceeding 90%) even when specificity is more variable, aligning with the priorities of screening programs to minimize missed cases.⁶⁴ While developers frequently assert that edge-based models can achieve rapid inference on portable hardware, published deployments rarely report explicit latency, power consumption, or throughput measurements, making it difficult to compare system efficiency or determine suitability for scaling. Furthermore, although local

inference has been shown to support continuity in settings with intermittent network connectivity, the robustness of these systems under extended offline or variable operating conditions remains insufficiently documented.

4.3. Comparative analysis: Edge artificial intelligence vs. cloud artificial intelligence

Comparative evaluations showed that cloud-based inference typically requires sustained upload bandwidth of 1–5 Mbps per image for diagnostic-quality files, with round-trip latency commonly exceeding 500–1,200 ms in LMIC mobile networks. These parameters frequently fall below the minimum required for uninterrupted clinical use. In contrast, edge AI deployments remove the dependency on continuous connectivity by executing inference locally, with reported on-device processing times generally falling in the 80–300 ms range on advanced RISC machine (ARM)-based accelerators or mobile neural processing units, even when models exceed 5–10 million parameters.⁶⁵

From a privacy standpoint, empirical studies have demonstrated that edge-based architectures reduce exposure risk by limiting transmission of identifiable patient imagery to external servers, and have been shown to comply with basic de-identification protocols in small-scale medical imaging pilots.⁶⁶ Hybrid configurations—where training and periodic model updates occur in the cloud but inference is executed locally—have been validated as a workable compromise, offering improved resilience under bandwidth-limited conditions while enabling periodic performance improvements without increasing data-sharing burdens.⁶⁷

4.4. Evidence gaps and limitations

Despite encouraging pilot data, the existing evidence base for edge AI in oncology remains constrained by several critical limitations. First, most published studies were **small-scale**, involving fewer than 500–1,000 participants, and were conducted under highly controlled research conditions rather than routine care settings, limiting external validity. Second, validation across heterogeneous populations, including diverse ethnic backgrounds, age groups, and cancer histologies, has been performed in only a handful of settings, raising concerns over the generalizability of current models.⁶⁸ Third, long-term outcomes, such as overall survival, integration into healthcare workflows, and cost-effectiveness, have not been rigorously studied: no published randomized or pragmatic trials demonstrate how edge AI screening or diagnosis impacts patient outcomes, resource utilization, or health system costs.⁶⁹ Fourth, ethical, regulatory, and governance issues, including data ownership, privacy safeguards,

model accountability, and post-deployment monitoring, are unevenly or superficially addressed in publications to date.⁷⁰ These knowledge gaps underscore the urgent need for well-powered, multicenter trials, coupled with standardized monitoring frameworks and transparent governance mechanisms, to support the safe scale-up of edge AI oncology interventions in LMICs.

5. Deployment challenges

The integration of autonomous edge AI systems into cancer diagnostics in LMICs poses complex challenges. To translate early pilot successes into scalable healthcare solutions, stakeholders must confront technical, clinical, governance, and economic barriers.

5.1. Technical challenges

Deploying edge AI requires hardware capable of performing inference locally, such as embedded GPUs, TPUs, or FPGAs. These devices often carry high costs and substantial energy requirements, which many LMIC healthcare settings cannot reliably support. Compounding this, most published work does not provide quantitative energy use, inference-per-second, or latency measurements, limiting transparency into real-world feasibility. Techniques such as model pruning, quantization, and knowledge distillation can reduce computational demand.⁷¹ However, without reporting of actual device-level performance parameters (e.g., power draw, thermal stability, and inference throughput), it remains difficult to assess whether optimized models will truly run reliably on constrained hardware.

5.2. Clinical challenges

From a clinical perspective, many edge AI systems have been validated only in limited-scale research cohorts rather than in routine care environments, raising serious questions about their diagnostic reliability and generalizability across diverse populations.⁷² Moreover, integrating with existing clinical workflows and health information systems is not straightforward; many LMIC facilities lack stable electronic health records, making embedding AI outputs into physician practice a nontrivial task. Regulatory approval is another hurdle: few deployments engage with formal medical device frameworks, particularly in LMICs, leaving uncertainty about long-term oversight. Finally, ethical dimensions, including data ownership, privacy protection, model explainability, accountability, and post-deployment monitoring, are often incompletely addressed in the literature, undermining trust and governance.

5.3. Risks and mitigations

The main risks identified in the current evidence base,

together with corresponding mitigation strategies, are summarized below.

5.3.1. Technical risks and mitigations

- Risk: There is a lack of standardized reporting at the device level on metrics such as power consumption, inference latency, and thermal stability, which compromises the ability to assess whether embedded GPUs, TPUs, or FPGAs can operate reliably in resource-constrained environments.
- Mitigation: Future deployments should include detailed benchmarking of hardware performance under real-world conditions. Developers should apply model compression methods, such as pruning, quantization, and knowledge distillation, and empirically test them on target devices. Consideration should also be given to more energy-efficient hardware, including low-power neural processing units, to ensure sustainability in settings with unstable power supplies.

5.3.2. Clinical and workflow risks and mitigations

- Risk: AI models validated only in small pilot cohorts risk performance degradation when scaled, particularly across different healthcare workflows and patient populations. Poor integration with clinical systems may result in low clinician adoption.
- Mitigation: Conduct multicenter trials in routine settings, ensuring evaluation frameworks capture real-world efficacy. Embed usability studies early to align system design with clinician workflows and adapt the AI output to existing processes.

5.3.3. Governance and privacy risks and mitigations

- Risk: There is inconsistent documentation of data governance, including privacy safeguards, data ownership, explainability, and post-deployment accountability. This raises ethical and regulatory concerns.
- Mitigation: Establish transparent governance frameworks that define data control, de-identification, access rights, and auditing mechanisms. Engage ethics boards and local regulatory bodies early, and co-develop oversight structures that suit LMIC contexts.

5.3.4. Economic and sustainability risks and mitigations

- Risk: The absence of economic modeling, including hardware procurement, maintenance, training, and system lifecycle costs, threatens program viability and scale-up.

- Mitigation: Incorporate cost-effectiveness analyses into pilot deployment plans, accounting for device amortization, training, and cost savings arising from reduced diagnostic delays. Compare edge AI with cloud-based and conventional diagnostics across different cost and resource contexts.

5.4. Infrastructure challenges

The reliable infrastructure required to support edge AI systems in LMICs is often fragile. In many settings, power supplies are inconsistent, maintenance support is limited, and supply chains for specialized hardware, including embedded GPUs, TPUs, and FPGAs, are poorly established.⁷³ These limitations reduce the operational reliability of AI-enabled diagnostic devices even when models are optimized for efficiency. To enhance resilience, implementers should prioritize energy-efficient AI architectures, develop localized maintenance protocols, and train biomedical technicians in hardware maintenance and repair. Diagnostic algorithms should be designed to degrade gracefully under power fluctuations, maintaining partial functionality rather than complete system failure.

5.5. Ethical and legal challenges

Deploying AI in healthcare introduces significant ethical and legal considerations in LMICs, particularly regarding data privacy, informed consent, algorithmic bias, and accountability. Regulatory and governance frameworks are frequently underdeveloped or absent, increasing the risk that patient data may be misused and that AI decisions lack transparency. Bias is a key concern because models trained primarily on urban or high-income datasets may underperform or make incorrect inferences in rural or underrepresented populations.⁷⁴ Developers should adopt comprehensive data governance frameworks that ensure privacy, clarity around data ownership, and accountability. They should also collect and train on datasets that reflect the demographic and clinical diversity of the populations they serve, and engage regulators, ethicists, and community representatives early to co-design appropriate oversight mechanisms.

5.6. Human factors

The success of AI-driven diagnostic tools depends on user trust, perceived utility, and ease of use among clinicians and patients.⁷⁵ Without trust building, systems risk underutilization or misuse. Developers should actively involve healthcare workers in system design and ensure AI models are explainable, providing clinically meaningful reasoning for recommendations. Training programs must be tailored to the resources and workflow realities of LMICs, equipping healthcare personnel to understand both the capabilities and limitations of AI tools.

5.7. Economic considerations

A major limitation in the current literature is the lack of formal economic modeling or cost-benefit analysis for edge AI deployment in LMICs. Financial sustainability depends on demonstrating that benefits outweigh total costs, including device acquisition, energy consumption, maintenance, personnel training, and eventual replacement. Without structured economic evaluation, policymakers cannot assess whether edge AI systems provide superior value compared to conventional diagnostic approaches.

Future implementations should integrate cost-effectiveness analyses from project inception. Direct costs such as hardware, software, energy, and staffing, as well as indirect costs such as workflow adaptation, training, and downtime during system integration, must be quantified. Benefits include faster diagnosis, improved clinical outcomes, reduced patient transfers, and optimized utilization of healthcare resources.

As an illustrative example, consider deploying an edge AI diagnostic system for a medium-sized regional hospital. The total upfront cost for devices, software licenses, and training is estimated at \$50,000. Annual maintenance and energy costs are approximately \$10,000. If the system reduces misdiagnosis and delays in care, the hospital may save an estimated \$25,000 per year in avoided hospital stays and treatments. Over a five-year horizon, the net benefit can be calculated as follows:

$$\begin{aligned}
 \text{Net benefit} &= (\text{Annual savings} \times 5) - (\text{Upfront cost} + [5 \times \text{Annual maintenance}]) \\
 &= (25,000 \times 5) - (50,000 + 5 \times 10,000) \\
 &= 125,000 - 100,000 = 25,000 \\
 &= 125,000 - 100,000 = 25,000
 \end{aligned} \tag{1}$$

This scenario suggests a positive return on investment. It can be refined using sensitivity analyses to account for power variability, device failure rates, and patient volume. Incorporating such modeling provides decision-makers with concrete evidence of financial viability and supports sustainable scale-up. Partnerships with governmental organizations and NGOs, as well as innovative financing mechanisms, are essential for deployment and scale in resource-limited settings.

6. Deployment roadmap

The progression from pilot studies to the widespread implementation of autonomous edge AI systems for cancer diagnostics in LMICs requires a structured, phased approach. Building on the challenges outlined in Section 5, this roadmap offers a stepwise strategy to ensure technological innovations become sustainable and

equitable healthcare solutions.⁷⁶

6.1. Phase 1: Prototype refinement—enhancing models and devices for robustness

This initial phase moves beyond generic software testing to address the core hardware–software co-design required for offline, low-power operation. Refinement must focus on model compression techniques (e.g., pruning and INT8 quantization) validated against diagnostic performance degradation thresholds (<3% drop in accuracy) and integration with specific portable imaging hardware (e.g., smartphone microscopes and low-cost digital slide scanners). Success is measured not just by algorithm accuracy but by achieving target inference speeds (>30 FPS) and power efficiency (sustained operation under 10 W) on designated edge processors (e.g., NVIDIA Jetson and Google Coral). This ensures the prototype is fundamentally shaped by the constraints of the intended low-resource deployment environment from the outset.

6.2. Phase 2: Clinical validation—multicenter trials in varied low-resource settings

Validation must extend beyond typical AI accuracy studies to assess the complete offline diagnostic chain: image acquisition on portable devices, on-device preprocessing, and autonomous inference without cloud fallback. Multicenter trials should be conducted in target low-resource settings to assess real-world diagnostic concordance, workflow integration time, and system reliability under intermittent power and variable user expertise. This phase generates the specific evidence base for edge-deployment feasibility, in contrast to cloud-dependent validation paradigms.

6.3. Phase 3: Regulatory alignment—World Health Organization prequalification and local approvals

Regulatory compliance is critical for adoption at scale. Developers should engage with international regulatory mechanisms (e.g., WHO prequalification) to ensure safety and efficacy and to facilitate uptake across settings. Simultaneously, engagement with national regulatory authorities in each LMIC is needed to secure approvals, ensure data governance, and embed accountability mechanisms for algorithmic decision-making, patient consent, and data privacy.⁷⁷ Establishing transparent data governance policies and audit trails strengthens trust among clinicians, patients, and policymakers.

6.4. Phase 4: Scalable deployment—partnerships with governments, non-governmental organizations, and the private sector

Scalability requires building capacity for local technical maintenance of specialized hardware, managing model

versioning across disconnected devices, and creating supply chains for edge-compatible components. Partnerships should be formed not just for distribution but for developing localized service hubs capable of hardware repairs, software updates via secure digital transfer, and user retraining. This phase addresses the sustained operational logistics unique to physical AI devices, unlike cloud services, which centralize maintenance.

6.5. Phase 5: Continuous learning—hybrid edge–cloud updates maintain autonomy

To maintain diagnostic accuracy without compromising autonomy or privacy, a hybrid edge–cloud update mechanism is essential. This phase implements federated learning or secure, encrypted model delta updates transmitted over intermittent connectivity. The governance model must ensure data sovereignty remains on-device while allowing aggregated learning from distributed sites. This creates a continuous improvement loop specific to edge AI, balancing the need for model evolution with the core requirement of offline functionality.

By following this phased deployment roadmap, autonomous edge AI systems can move from experimental pilots to scalable and sustainable solutions, reducing diagnostic delays and improving cancer care equity in LMICs.

7. Future research directions

The pathway toward the widespread adoption of autonomous edge AI for cancer diagnosis in low-resource settings requires a focused research agenda to address persisting gaps and emerging challenges. Future work should emphasize measurable, context-specific research goals rather than broad abstractions.

First, integration of edge AI with point-of-care diagnostic tools remains a key priority. Future studies should examine how portable ultrasound devices, low-cost imaging tools, and rapid molecular assays can be combined with edge models to improve diagnostic accuracy. Research trials should quantify the incremental benefit in sensitivity, specificity, and time-to-result compared to standalone edge models, ideally reporting metrics such as false-negative rate reductions and diagnostic latency in seconds.

Second, the development of multimodal AI systems is essential. Research should focus on algorithms that fuse data from pathology, radiology, and genomics directly on edge devices. Prospective studies must benchmark model performance in terms of diagnostic gains (e.g., AUC) and resource cost (e.g., memory usage and compute cycles). Hardware innovation must go hand in hand with algorithmic design to support such convergence on

resourceconstrained devices.

Third, advancing explainability and interpretability is critical. Future work should evaluate explainable AI (XAI) techniques such as saliency mapping, feature attribution, case-based reasoning, or concept-bottleneck models in real LMIC clinical settings, and validate them against clinician trust metrics (for instance, via surveys measuring perceived explanation quality and decision confidence).⁷⁸ This allows quantifiable assessment of how “transparent” models affect adoption among health workers and patients.

Fourth, governance frameworks tailored for autonomous diagnostics in low-resource settings must be developed. Rather than applying generic or high-income-country models, research should co-design governance principles with LMIC stakeholders, focusing on data sovereignty, auditability, liability, and cultural norms. Studies such as the Zanzibar health data governance initiative highlight the importance of real-world, operational data governance for AI in LMIC health systems.⁷⁹ Similarly, alignment with global consensus guidelines, such as FUTURE-AI, should be adapted to local priorities, with measurable indicators of trust, fairness, and accountability.

Fifth, sustainability needs concrete targets. Research must explore energy-harvesting designs (e.g., solar or kinetic), recyclable materials, and end-of-life strategies for edge devices. Studies should model power consumption (measured in milliwatts), device lifetime (in years), and e-waste volumes (in kilograms per device per year), with quantitative benchmarking in pilot deployments.

Finally, capacity building and education for local health workers must be prioritized with explicit goals. Research programs should measure outcomes such as the number of clinicians trained, retention rates, maintenance proficiency (e.g., the percentage of local staff able to perform hardware repairs), and ethical literacy (e.g., through pre- and post-training assessments). Embedding capacity building within health systems will strengthen local ownership and long-term sustainability.

By anchoring future research in measurable, context-tailored priorities from model explainability to governance, sustainability, and capacity building, the field can ensure that autonomous edge AI for cancer care not only advances technically but also aligns ethically and sustainably with LMIC needs.

8. Conclusion

Autonomous edge AI represents a transformative frontier in the global effort to democratize cancer diagnosis. By combining low-latency processing with decentralized intelligence, these systems offer an unprecedented

opportunity to extend high-quality diagnostic capabilities to underserved regions, particularly in low-resource settings where infrastructural limitations often compromise timely care. This shift not only promises to close the equity gap in oncology services but also reinforces the broader vision of health as a universal right.

The successful realization of this vision, however, demands more than technical ingenuity. It requires sustained collaboration across disciplines; engineers and data scientists refining algorithms, clinicians ensuring clinical relevance and safety, policymakers shaping conducive regulatory landscapes, and local health practitioners bridging innovations with community realities. Such convergence ensures that edge AI for cancer diagnosis evolves into a holistic solution that balances scientific rigor with contextual needs.

Lastly, the pathway to widespread adoption calls for coordinated investments, supportive policy frameworks, and a commitment to global cooperation. If stakeholders align efforts strategically, autonomous edge AI can shift from promising prototypes to integral components of healthcare delivery systems worldwide. The outcome would not merely be technological advancement, but a tangible step toward equity in cancer care—where geography, resource limitations, or infrastructural deficits no longer determine survival outcomes.

Acknowledgments

None.

Funding

None.

Conflict of interest

The author declares no conflicts of interest.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable.

References

1. Ngwa W, Addai BW, Adewole I, *et al.* Cancer in sub-Saharan Africa: a Lancet Oncology Commission. *Lancet Oncol.* 2022;23(6):e251-e312.
doi: 10.1016/S1470-2045(21)00720-8
2. Wilson ML. Pathology and laboratory medicine in universal health coverage. *J Lab Precis Med.* 2019;4:34.
doi: 10.21037/jlpm.2019.09.06
3. Byers RJ, Byers AJ, Mumba C, *et al.* Development of an online teaching platform to improve access to postgraduate pathology training in sub-Saharan Africa. *Front Med.* 2024;11:1390560.
doi: 10.3389/fmed.2024.1390560
4. Fleming KA. Pathology and cancer in Africa. *Ecantermedicalscience.* 2019;13:945.
doi: 10.3332/ecancer.2019.945
5. Mapanga W, Norris SA, Craig A, *et al.* Drivers of disparities in stage at diagnosis among women with breast cancer: South African breast cancers and HIV outcomes cohort. *PLoS ONE.* 2023;18(2):e0281916.
doi: 10.1371/journal.pone.0281916
6. Naqvi A, Manglik N, Dudrey E, *et al.* Evaluating the performance of a low-cost mobile phone attachable microscope in cervical cytology. *BMC Women's Health.* 2020;20:60.
doi: 10.1186/s12905-020-00902-0
7. Choudhury D, Dolezal JM, Dyer E, *et al.* Developing a low-cost, open-source, locally manufactured workstation and computational pipeline for automated histopathology evaluation using deep learning. *EBioMedicine.* 2024;107:105276.
doi: 10.1016/j.ebiom.2024.105276
8. Campanella G, Hanna MG, Geneslaw L, *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med.* 2019;25(8):1301-1309.
doi: 10.1038/s41591-019-0508-1
9. Bao H, Sun X, Zhang Y, *et al.* The artificial intelligence-assisted cytology diagnostic system in large-scale cervical cancer screening: a population-based cohort study of 0.7 million women. *Cancer Med.* 2020;9(18):6896-6906.
doi: 10.1002/cam4.3296
10. Ting DSW, Cheung CY, Lim G, *et al.* Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA.* 2017;318(22):2211-2223.
doi: 10.1001/jama.2017.18152
11. Holmström, O., Linder, N., Kaingu, H., *et al.* Point-of-care digital cytology with artificial intelligence for cervical cancer screening in a resource-limited setting. *JAMA Netw Open.*

- 2021;4(3):e211740.
doi: 10.1001/jamanetworkopen.2021.1740
12. Hijazi A, Bifulco C, Baldin P, Galon J. Digital pathology for better clinical practice. *Cancers*. 2024;16(9):1686.
doi: 10.3390/cancers16091686
13. World Health Organization. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. World Health Organization. Available from: <https://www.who.int/publications/i/item/9789240029200>. [Last accessed on January 19, 2026].
14. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol*. 2019;16(11):703–715.
doi: 10.1038/s41571-019-0252-y
15. Fleming KA, Horton S, Wilson ML, *et al*. The Lancet Commission on diagnostics: transforming access to diagnostics. *Lancet*. 2021;398(10315):1997–2050.
doi: 10.1016/S0140-6736(21)00673-5
16. Paez A. Gray literature: an important resource in systematic reviews. *J Evid Based Med*. 2017;10(3):233–240.
doi: 10.1111/jebm.12266
17. Grames EM, Stillman AN, Tingley MW, Elphick CS. An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods Ecol Evol*. 2019;10(9):1645–1654.
doi: 10.1111/2041-210X.13268
18. Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
doi: 10.1136/bmj.n71
19. Ciecierski-Holmes T, Singh R, Axt M, *et al*. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. *NPJ Digit Med*. 2022;5:162.
doi: 10.1038/s41746-022-00700-y
20. Banbury C, Reddi VJ, Lam M, *et al*. Benchmarking tiny machine learning on microcontrollers. *arXiv*. Preprint posted online 2003.
doi: 10.48550/arXiv.2003.04821
21. Wofk D, Ma F, Yang TJ, Karaman S, Sze V. FastDepth: Fast monocular depth estimation on embedded systems. In: 2019 International Conference on Robotics and Automation (ICRA), 2019:6101–6108.
doi: 10.1109/ICRA.2019.8794182
22. Farhadipour A. Lung and colon cancer detection with convolutional neural networks and adaptive histogram equalization. *Iran J Comput Sci*. 2024;7(3):381–395.
doi: 10.1007/s42044-023-00161-w
23. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115–118.
doi: 10.1038/nature21056
24. Haenssle HA, Fink C, Toberer F, *et al*. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. *Ann Oncol*. 2020;31(1):137–143.
doi: 10.1016/j.annonc.2019.10.013
25. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell*. 2020;2:305–311.
doi: 10.1038/s42256-020-0186-1
26. Fujisawa Y, Otomo Y, Ogata Y, *et al*. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. *Br J Dermatol*. 2019;180(2):373–381.
doi: 10.1111/bjd.16924
27. Sze V, Chen YH, Yang TJ, Emer JS. Efficient processing of deep neural networks: a tutorial and survey. *Proc IEEE*. 2017;105(12):2295–2329.
doi: 10.1109/JPROC.2017.2761740
28. Valverde JM, Shatillo A, Tohka J. Sauron U-Net: simple automated redundancy elimination in medical image segmentation via filter pruning. *Neurocomputing*. 2024;594:127817.
doi: 10.1016/j.neucom.2024.127817
29. Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv*. Preprint posted online 2010.
doi: 10.48550/arXiv.2010.11929
30. Aburass S, Dorgham O, Al Shaqsi J, *et al*. Vision transformers in medical imaging: a comprehensive review of advancements and applications across multiple diseases. *J Digit Imaging*. 2025;38:3928–3971.
doi: 10.1007/s10278-025-01481-y
31. Zhou Z, Sodha V, Rahman Siddiquee MMR, *et al*. Models genesis: generic autodidactic models for 3D medical image analysis. *Med Image Anal*. 2021;67:101840.
doi: 10.1016/j.media.2020.101840
32. Azizi S, Mustafa B, Ryan F, *et al*. Big self-supervised models advance medical image classification. *arXiv*. Preprint posted online 2021.
doi: 10.48550/arXiv.2101.05224
33. Boerkamp C, Thomas AJ. QuantU-Net: efficient wearable

- medical imaging using bitwidth as a trainable parameter. *arXiv*. Preprint posted online 2025.
doi: 10.48550/arXiv.2503.08719.
34. Nguyen SA, Phan NL, Dao VH, *et al*. BlazeNeo: blazing fast polyp segmentation and neoplasm detection. *arXiv*. Preprint posted online 2022.
doi: 10.48550/arXiv.2503.08719
 35. Ji M, Al-Ars Z, Hofstee P, Chang Y, Zhang B. FPQNet: fully pipelined and quantized CNN for ultra-low latency image classification on FPGAs using OpenCAPI. *Electronics*. 2023;12(19):4085.
doi: 10.3390/electronics12194085
 36. An NS, Lan PN, Hang DV, *et al*. BlazeNeo: blazing fast polyp segmentation and neoplasm detection. *arXiv*. Preprint posted online 2022.
doi: 10.48550/arXiv.2203.00129
 37. Ji M, Al-Ars Z, Hofstee P, Chang Y, Zhang B. FPQNet: fully pipelined and quantized CNN for ultra-low latency image classification on FPGAs using OpenCAPI. *Electronics*. 2023;12(19):4085.
doi: 10.3390/electronics12194085
 38. AskariHemmat MH, Honari S, Rouhier L, *et al*. U-Net fixed-point quantization for medical image segmentation. *arXiv*. Preprint posted online 2019.
doi: 10.48550/arXiv.1908.01073
 39. Tasci M, Istanbulu A, Tumen V, Kosunalp S. FPGA-QNN: quantized neural network hardware acceleration on FPGAs. *Appl Sci*. 2025;15(2):688.
doi: 10.3390/app1502068
 40. Yurdem B, Kuzlu M, Gullu MK, Catak FO, Tabassum M. Federated learning: Overview, strategies, applications, tools and future directions. *Heliyon*. 2024;10(19):e38137.
doi: 10.1016/j.heliyon.2024.e38137
 41. Sheller MJ, Edwards B, Reina GA, *et al*. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep*. 2020;10(1):12598.
doi: 10.1038/s41598-020-69250-1
 42. Yan Y, Li X, Cui K, Sun H, Yu Z. FreeNet: An efficient frequency-domain early exiting network for dynamic inference. *Knowl Based Syst*. 2025;314:113155.
doi: 10.1016/j.knosys.2025.113155
 43. Liu Q, Tian Z, Wang N, Lin Y. DRL-based dependent task offloading with delay-energy tradeoff in medical image edge computing. *Complex Intell Syst*. 2024;10(3):3283-3304.
doi: 10.1007/s40747-023-01322-x
 44. Rivenson Y, Zhang Y, Günaydin H, Teng D, Ozcan A. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light Sci Appl*. 2018;7:17141.
doi: 10.1038/lsa.2017.141
 45. Bian Y, Xing T, Jiao K, *et al*. Computational portable microscopes for point-of-care-test and tele-diagnosis. *Cells*. 2022;11(22):3670.
doi: 10.3390/cells11223670
 46. Han GR, Goncharov A, Eryilmaz M, *et al*. Machine learning in point-of-care testing: innovations, challenges, and opportunities. *Nat Commun*. 2025;16:3165.
doi: 10.1038/s41467-025-58527-6
 47. Zhu W, Gong C, Kulkarni N, Nguyen CD, Kang D. Smartphone-based microscopes. In: *Smartphone Based Medical Diagnostics*. Academic Press; 2020:159-175.
 48. Dwork C, Roth A. The algorithmic foundations of differential privacy. *Found Trends Theor Comput Sci*. 2014;9(3-4):211-407.
doi: 10.1561/04000000042
 49. Gilad-Bachrach R, Dowlin N, Laine K, *et al*. CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy. Proceedings of Machine Learning Research. 2016. Available from: <https://proceedings.mlr.press/v48/gilad-bachrach16.html> [last accessed on].
 50. Nebeker C, Torous J, Bartlett Ellis RJ. Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Med*. 2019;17:137.
doi: 10.1186/s12916-019-1377-7
 51. Shepherd C, Markantonakis K. *Trusted Execution Environments*. Springer; 2024.
doi: 10.1007/978-3-031-55561-9
 52. Liu X, Glocker B, McCradden MM, *et al*. The medical algorithmic audit. *Lancet Digit. Health*. 2022;4(5):e384-e397.
doi: 10.1016/S2589-7500(22)00003-6
 53. Allahqoli L, Laganà AS, Mazidimoradi A, *et al*. Diagnosis of cervical cancer and pre-cancerous lesions by artificial intelligence: a systematic review. *Diagnostics*. 2022;12(11):2771.
doi: 10.3390/diagnostics12112771
 54. Dwork C, Roth A. The algorithmic foundations of differential privacy. *Found Trends Theor Comput Sci*. 2014;9(3-4):211-407.
doi: 10.1561/04000000042
 55. Mertz M, Toskovich K, Shields G, Attema G, Dumond J, Cameron E. Exploring trust factors in AI-healthcare integration: a rapid review. *Front ArtifIntell*. 2025;8:1658510.
doi: 10.3389/frai.2025.1658510
 56. Linder N, Nyirenda D, Mårtensson A, *et al*. AI supported diagnostic innovations for impact in global women's health.

- BMJ. 2025;391:e086009.
doi: 10.1136/bmj-2025-086009
57. Valles-Coral MA, Pinedo L, Rodríguez C, *et al.* Application of artificial intelligence in cervical cytology: a systematic review of deep learning models, datasets, and reported metrics. *Front Big Data.* 2026;8:1678863.
doi: 10.3389/fdata.2025.1678863
 58. Tschandl, P., Rinner, C., Apalla, Z., *et al.* Human-computer collaboration for skin cancer recognition. *Nat Med.* 2020;26(8):1229-1234.
doi: 10.1038/s41591-020-0942-0
 59. Liu L, Liu J, Su Q, Chu Y, Xia H, Xu R. Performance of artificial intelligence for diagnosing cervical intraepithelial neoplasia and cervical cancer: a systematic review and meta-analysis. *EclinicalMedicine.* 2024;80:102992.
doi: 10.1016/j.eclinm.2024.102992
 60. Witkowski, A. M., Burshtein, J., Christopher, M., *et al.* Clinical utility of a digital dermoscopy image-based artificial intelligence device in the diagnosis and management of skin cancer by dermatologists. *Cancers.* 2024;16(21):3592.
doi: 10.3390/cancers16213592
 61. Breslauer, D. N., Maamari, R. N., Switz, N. A., Lam, W. A., Fletcher, D. A. Mobile phone based clinical microscopy for global health applications. *PLoS ONE.* 2009;4(7):e6320.
doi: 10.1371/journal.pone.0006320
 62. Kabukye, J. K., Namugga, J., Mpamani, C. J., *et al.* Implementing smartphone-based telemedicine for cervical cancer screening in Uganda: qualitative study of stakeholders' perceptions. *J Med Internet Res.* 2023;25:e45132.
doi: 10.2196/45132
 63. Bai, J., Li, N., Ye, H., *et al.* AI-assisted cervical cytology precancerous screening for high-risk population in resource-limited regions using a compact microscope. *Nat Commun.* 2025;16:7429.
doi: 10.1038/s41467-025-62589-x
 64. Baller SP, Jindal A, Chadha M, Gerndt M. DeepEdgeBench: benchmarking deep neural networks on edge devices. 2021 IEEE International Conference on Cloud Engineering (IC2E), 2021:20-30.
doi: 10.1109/IC2E52221.2021.00016
 65. Kadar Mohamed Jabarullah MA, Chakka SN, Kanungo S. Edge-cloud hybrid architecture for distributed AI applications. 2025 International Conference on Computing Technologies & Data Communication (ICCTDC), 2025:1-5.
doi: 10.1109/ICCTDC64446.2025.11158770
 66. Shen M, Zou Z, Bao H, *et al.* Cost-effectiveness of artificial intelligence-assisted liquid-based cytology testing for cervical cancer screening in China. *Lancet Reg Health West Pac.* 2023;34:100726.
doi: 10.1016/j.lanwpc.2023.100726
 67. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med.* 2018;15(11):e1002689.
doi: 10.1371/journal.pmed.1002689
 68. Pham T. Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. *R Soc Open Sci.* 2025;12(5):241873.
doi: 10.1098/rsos.241873
 69. Wang T, Guo J, Zhang B, Yang G, Li D. Deploying AI on edge: advancement and challenges in edge intelligence. *Mathematics.* 2025;13(11):1878.
doi: 10.3390/math13111878
 70. Zhang ZM. Ethics and governance of trustworthy medical artificial intelligence. *BMC Med Inform Decis Mak.* 2023;23(1):7.
doi: 10.1186/s12911-023-02103-9
 71. López DM, Rico-Olarte C, Blobel B, Hullin C. Challenges and solutions for transforming health ecosystems in low- and middle-income countries through artificial intelligence. *Front Med.* 2022;9:958097.
doi: 10.3389/fmed.2022.958097
 72. Mittelstadt BD, Floridi L, *et al.* The ethics of algorithms: mapping the debate. *Big Data Soc.* 2016;3(2):1-21.
doi: 10.1177/2053951716679679
 73. Marey A, Ambrozaite O, Afifi A, *et al.* A perspective on AI implementation in medical imaging in LMICs: challenges, priorities, and strategies. *Eur Radiol.* 2025;35(5):2985-2995.
doi: 10.1007/s00330-025-12031-z
 74. Yang J, Dung NT, Thach PN, *et al.* Generalizability assessment of AI models across hospitals in a low-middle and high income country. *Nat Commun.* 2024;15:8270.
doi: 10.1038/s41467-024-52618-6
 75. Ienca M, Vayena E. On the responsible use of digital data to tackle the COVID-19 pandemic. *Nat Med.* 2020;26(8):1183-1185.
doi: 10.1038/s41591-020-0832-5
 76. Mehta MC. AI & healthcare in low- and middle-income countries. ITU-T Workshop Presentation. Available from: https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20181114/Documents/Mehul%20C%20Mehta_Presentation.pdf [Last accessed on November 30, 2026].
 77. Ankolekar A, Boie S, Abdollahyan M, *et al.* Advancing breast, lung and prostate cancer research with federated learning: a systematic review. *npj Digit Med.* 2025;8:314.
doi: 10.1038/s41746-025-01591-5

78. Ali Z, Waqar M, Kim A. Federated learning in public health: a systematic review of decentralized, equitable, and secure disease prevention approaches. *Healthcare*. 2025;13(21):2760.
doi: 10.3390/healthcare13212760
79. Li T, Wandella A, Gomer R, Al-Mafazy MH. Operationalizing health data governance for AI innovation in low-resource government health systems: a practical implementation perspective from Zanzibar. *Data Policy*. 2024;6:e63.
doi: 10.1017/dap.2024.65