Advances in Medical Sciences and Research
Vol. 1 (2026), No. 2, pp. 101-119
DOI: https://doi.org/10.55578/amsr.2605.008

Intelligent Medical Imaging Systems for Rapid Real-Time Diagnosis: Deep Learning, Edge AI, Hardware Acceleration, and Healthcare Intelligence

Bulus Bali[1,*], Manga Ibrahim[1], Favanza Iliya Kwaha[2]

[1] *Department of Computer Science, Adamawa State University, Mubi, Adamawa State, Nigeria.*

[2] *Department of Pure and Applied Physics, Adamawa State University, Mubi, Adamawa State, Nigeria.*

*Corresponding Author. Email: bali930@adsu.edu.ng

Received: 18 May 2026, Accepted: 25 May 2026, Published: 27 May 2026

Abstract

The rapid evolution of Artificial Intelligence (AI) has transformed medical imaging into an intelligent and data-driven clinical decision-support ecosystem. The growing demand for accurate, low-latency, scalable, and interpretable diagnostics has accelerated the integration of deep learning, Edge AI, and hardware acceleration in healthcare systems. Recent developments in intelligent medical imaging systems for real-time clinical diagnosis are examined in this article. The systematic review was conducted using peer-reviewed literature retrieved from IEEE Xplore, PubMed, Scopus, Web of Science, and ScienceDirect, covering studies published from 2020 to April 2026. 4,912 articles were screened, with 52 high-quality studies included for final synthesis. The review focuses on convolutional neural networks, federated learning, explainable AI, transformer-based architectures, multimodal imaging, and hardware accelerators, including Application-Specific Integrated Circuits (ASICs), Field-Programmable Gate Arrays (FPGAs), Tensor Processing Units (TPUs), Graphics Processing Units (GPUs), and neuromorphic processors. Findings indicate substantial improvements in diagnostic accuracy, computational efficiency, scalability, and low-latency inference across cardiology, oncology, pathology, radiology, and ophthalmology. Main challenges include heterogeneous data sources, high energy demands, limited clinical validation, and lack of interpretability, privacy concerns, interoperability issues, and regulatory barriers. Emerging directions include digital twins, lightweight AI, multimodal foundation models, sustainable edge intelligence, and privacy-preserving federated ecosystems, highlighting the transformative potential of intelligent medical imaging in precision healthcare.

Keywords: Deep learning, Explainable AI, Hardware acceleration, Low-latency diagnosis, Medical image analysis

1. INTRODUCTION

Medical imaging is a cornerstone of modern healthcare, enabling precise visualization of anatomical structures, physiological processes, and pathological abnormalities for accurate diagnosis and treatment planning. Imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), ultrasound, histopathology, and retinal imaging have significantly advanced. These modalities generate vast amounts of complex clinical data that require intelligent and automated analysis. Conventional diagnostic workflows, which largely depend on manual interpretation by radiologists and specialists, are increasingly constrained by diagnostic delays, clinician workload, inter-observer variability, and limited access to expert healthcare professionals in resource-limited settings. Consequently, AI has emerged as a transformative solution for improving diagnostic accuracy, efficiency, and scalability in medical imaging systems [1,2].

Deep learning (DL) has become the dominant paradigm in intelligent medical imaging due to its exceptional capability for automated feature extraction, hierarchical representation learning, and high-

dimensional pattern recognition. Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), transformers, and hybrid deep learning models are examples of advanced architectures that have demonstrated exceptional performance in medical imaging tasks. These include segmentation, disease detection, image registration, multimodal image analysis, and picture classification [3,4,5,6]. Recent advances have enabled AI-driven imaging systems to achieve expert-level diagnostic performance in applications such as cancer detection, diabetic retinopathy screening, pathology analysis, and radiological interpretation.

Despite these advances, centralized cloud-based AI infrastructures remain limited by bandwidth overhead, network latency, privacy concerns, and inadequate real-time responsiveness. Clinical applications such as emergency diagnostics, intensive care monitoring, surgical guidance, and remote telemedicine require ultra-low-latency inference and reliable decision support that cannot tolerate communication delays. As a result, Edge AI has emerged as a critical paradigm for deploying intelligence closer to wearable sensors, imaging devices, and point-of-care systems. By supporting localized inference and decentralized analytics, edge computing significantly improves bandwidth efficiency, privacy preservation, diagnostic responsiveness, and system reliability [7,8,9].

Simultaneously, the computational demands of DL have accelerated the adoption of specialized hardware acceleration platforms optimized for medical AI workloads. ASICs, FPGAs, GPUs, TPUs, neuromorphic processors, and heterogeneous edge architectures now play a pivotal role in accelerating image reconstruction, multimodal fusion, segmentation, and real-time inference under strict latency and energy constraints [10,11,12]. In parallel, hardware-aware optimization techniques, including knowledge distillation, mixed-precision computing, quantization, pruning, and lightweight neural architectures, have facilitated efficient deployment of AI models on edge and embedded healthcare devices [13,14].

This review critically examines the convergence of DL, Edge AI, and hardware acceleration in intelligent medical imaging systems for rapid real-time diagnosis. Unlike prior studies that focus on a single aspect (e.g., only DL or only edge computing), this review unifies DL, Edge AI, and hardware acceleration for medical imaging into a cohesive framework. Special emphasis was placed on real-time clinical deployment, quantifying latency and throughput requirements [15,16,17]. Furthermore, the review explores the convergence of Internet of Medical Things (IoMT) devices, 5G-enabled telemedicine, federated learning, multimodal intelligence, and explainable AI within healthcare ecosystems [18,19,20,21]. The study provides researchers, clinicians, engineers, and policymakers with insights into intelligent, scalable, patient-centric medical imaging for future precision medicine and healthcare delivery.

2. MATERIAL AND METHODS

2.1. Search Strategy

This review followed the PRISMA 2020 reporting guidelines to ensure methodological transparency, reproducibility, and rigorous study selection. A comprehensive literature search was conducted across major scientific databases, including IEEE Xplore, PubMed, Scopus, Web of Science, and ScienceDirect, covering studies published between January 2020 and April 2026. Boolean search expressions were developed using combinations of keywords related to intelligent medical imaging, DL, Edge AI, federated learning, hardware acceleration, and IoMT. The primary search string was formulated as:

("medical imaging" OR "medical image analysis") AND ("hardware acceleration" OR "FPGA" OR "GPU" OR "TPU" OR "ASIC" OR "IoMT" OR "Internet of Medical Things") AND ("deep learning" OR "transformer networks" OR "edge AI" OR "federated learning") AND ("low-latency diagnosis" OR "edge computing" OR "real-time healthcare") AND (Publication Year: 2020-2026).

Database-specific refinements were applied where necessary, including MeSH terms and controlled vocabulary indexing in PubMed. Retrieved records were exported into a reference management system for duplicate removal and screening.

2.2. Inclusion Criteria

Studies that focused on AI-enabled medical imaging applications, such as diagnosis, image classification, registration, multimodal imaging, segmentation, or clinical decision support, and were published in English between 2020 and April 2026, were included in the synthesis. One domain, such as DL architectures, edge AI, explainable AI, federated learning, hardware acceleration, IoMT, or low-latency clinical deployment, was covered by eligible studies. Experimental validation, comparative analysis, review technique, or therapeutically relevant application is necessary for inclusion in the works.

2.3. Exclusion Criteria

Studies were excluded if they were conference papers, book chapters, commentaries, editorials, non-peer-reviewed works, non-English publications, workshops, or studies lacking relevance to medical imaging or intelligent healthcare systems. Purely theoretical ML studies without clinical applicability, duplicate datasets, or methodologically weak studies with insufficient validation were excluded. Figure 1 visualizes the study selection procedure. A total of 52 studies were included, with trends analyzed in Figure 2, accordingly and systematically.

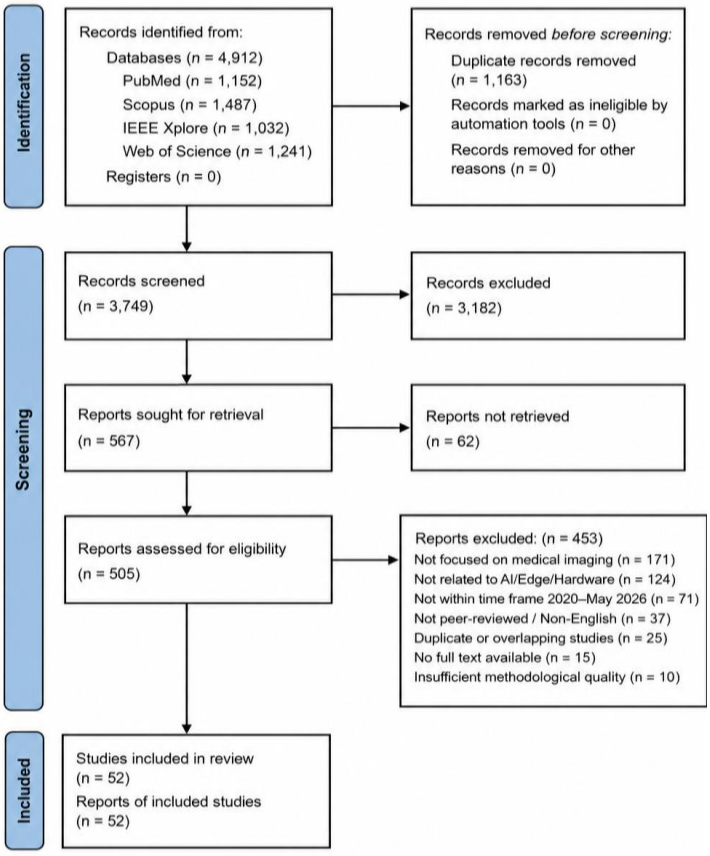

Figure 1. PRISMA flowchart for the selection of included studies

3. RESULTS AND DISCUSSION

3.1. Quality Assessment

Study quality was evaluated using a structured framework aligned with PRISMA 2020. Due to heterogeneity across DL, Edge AI, hardware acceleration, and clinical imaging studies, the Mixed Methods Appraisal Tool (MMAT) was applied to empirical studies, while the Critical Appraisal Skills Programme (CASP) checklist was used for reviews and qualitative research. Assessments covered three dimensions: AI transparency (reproducibility and interpretability), methodological quality (design validity, dataset adequacy, experimental rigor), and risk of bias (selection, evaluation, and publication bias).

Table 1. Quality Assessment approach for Included Studies

Component	Evaluation Focus	Tool/Metric
Methodological Quality	Study design, validation, reproducibility	CASP / MMAT
AI Transparency	Explainability, reporting completeness	ATS (0-5 scale)
Risk of Bias	Dataset, model, evaluation bias	Structured bias checklist

AI Transparency Scoring (ATS) used a 5-point scale (0-5), where 0-1 indicated low transparency, 2-3 moderate transparency, and 4-5 high transparency. Only studies rated as moderate to high quality across all three assessment dimensions were included in the final synthesis (n = 52). This inclusion rule guaranteed the chosen evidence base's medical importance, repeatability, and methodological rigor.

The review variables and data extraction used in this study include AI architecture, which comprises CNNs, Transformers, and GANs; hardware platforms, including FPGA, GPU, and TPU; clinical domains such as radiology and evaluation metrics, oncology; including accuracy, Dice coefficient, and AUC; edge deployment capability (Yes/No); and explainability approaches based on XAI methods.

3.2. Evolution of AI in medical imaging systems

The evolution of AI in medical imaging has progressed from rule-based expert systems and conventional ML algorithms to advanced deep neural architectures capable of intelligent and autonomous clinical analysis. Early medical imaging systems relied heavily on handcrafted features and traditional algorithms such as decision trees, support vector machines, and random forests. Although these approaches achieved moderate success in structured diagnostic tasks, their dependence on manual feature engineering limited adaptability, scalability, and generalization across diverse imaging environments.

DL has transformed medical image analysis by enabling automatic hierarchical feature extraction from raw data. CNNs became central due to their ability to capture complex anatomical patterns and spatial invariances. Architectures such as AlexNet, DenseNet, ResNet, VGGNet, U-Net, and transformers have improved segmentation, classification, detection, disease prediction, and multimodal analysis [3,5,6], establishing DL as the dominant paradigm. Foundational work [2,22,23] showed major gains in automated representation learning and diagnostic accuracy. Recent studies demonstrate expert-level performance in radiology, pathology, ophthalmology, oncology, and cardiology, including cancer detection and diabetic retinopathy screening [4,24]. Multimodal AI integrating imaging, genomics, and clinical data further advances precision medicine, supported by AI, big data, cloud computing, and open-source ecosystems [25].

The growing integration of AI into healthcare has accelerated the development of edge-enabled and low-latency diagnostic systems. Edge AI frameworks now support real-time medical imaging inference by moving computational intelligence closer to imaging devices, wearable sensors, and point-of-care platforms. This improves responsiveness, enhances privacy preservation, and increases bandwidth efficiency [7,8,26]. Simultaneously, specialized hardware accelerators, including ASICs, FPGAs, GPUs, TPUs, and neuromorphic processors, have enhanced the scalability and computational efficiency of DL workloads in clinical environments [10,12,27]. Recent advances in portable MRI

acceleration technologies have further highlighted the growing importance of hardware-aware optimization for real-time, low-latency intelligent medical imaging systems [28].

Despite advances, challenges in fairness, interpretability, robustness, privacy, and clinical trust limit large-scale adoption. The black-box nature of DL raises concerns about transparency and reliability in clinical use. Therefore, XAI, trustworthy AI, federated learning, and fairness-aware methods are essential for ethical and reliable medical imaging systems [9,18, 29].

3.3. Convolutional Neural Networks

CNNs remain the foundational architecture for intelligent medical imaging due to their strong hierarchical feature extraction, spatial representation learning, and pathological pattern recognition capabilities. CNN-based systems achieve state-of-the-art performance in tumor detection, lesion localization, organ segmentation, radiological interpretation, and disease classification across CT, MRI, ultrasound, and histopathological imaging modalities. Architectures including DenseNet, ResNet, EfficientNet, and lightweight edge-oriented CNNs significantly improve diagnostic accuracy, scalability, and computational efficiency in clinical environments [2,3]. Transfer learning has accelerated healthcare deployment by enabling adaptation to limited annotated datasets, while pruning, quantization, and mixed-precision inference support efficient deployment on embedded and edge healthcare devices [11,14]. However, CNNs still face challenges related to annotation scarcity, data imbalance, limited interpretability, and hardware bottlenecks, including memory bandwidth limitations, I/O latency, and high energy consumption in large-scale, real-time MRI and CT reconstruction systems.

3.4. Segmentation Networks

Medical image segmentation is essential for disease monitoring, radiotherapy planning, surgical navigation, and precision diagnosis. Encoder-decoder models such as U-Net excel in tumor delineation, organ extraction, histopathology, lesion localization, and volumetric MRI segmentation due to skip connections and multiscale feature fusion [11,12]. Recent residual, attention, and CNN-transformer hybrids improve accuracy, but real-time deployment remains challenging due to high computational and memory demands. Edge computing integration supports low-latency, real-time clinical imaging in resource-constrained settings [30].

3.5. Recurrent and Temporal Learning Models

Dynamic medical imaging and physiological signal analysis require architectures capable of modeling temporal and sequential dependencies. Long Short-Term Memory (LSTM) networks, Recurrent Neural Networks (RNNs), and Gated Recurrent Units (GRUs) have demonstrated effectiveness in cardiac motion analysis, patient monitoring, video endoscopy interpretation, and longitudinal disease progression modeling [31]. These temporal learning models enhance predictive analytics by capturing spatiotemporal relationships within sequential imaging and biosignal datasets.

Edge-assisted temporal learning systems have further improved intelligent telemedicine, remote patient monitoring, and latency-sensitive healthcare analytics through distributed inference [32,33]. Nevertheless, maintaining robustness in continuous monitoring systems remains challenging because missing temporal information, noisy physiological signals, and heterogeneous multimodal inputs can degrade prediction reliability and clinical interpretability.

3.6. Transformer-Based Medical Imaging systems

Vision Transformers (ViTs) and multimodal transformer architectures offer superior global contextual learning and long-range dependency modeling compared to CNNs, enabling applications in pathology, radiology, radiomics, multimodal fusion, and AI-assisted diagnosis [21]. They integrate imaging with EHRs, genomics, and biosensor data, advancing precision medicine. However, their large model size, high computational cost, and memory-intensive attention hinder edge deployment. Lightweight variants like Swin Transformers and distillation help but may reduce performance, while multimodal systems also raise challenges of privacy, heterogeneity, and interpretability.

3.7. Generative and Self-Supervised Learning

Generative adversarial networks (GANs) and self-supervised learning (SSL) are improving intelligent medical imaging by addressing annotation shortages, class imbalance, privacy concerns, and various clinical datasets. GAN-based methods are widely used for synthetic medical image generation, anomaly detection, multimodal image synthesis, low-dose image enhancement, and reconstruction, thereby improving dataset diversity, robustness, and diagnostic performance in healthcare systems [4,34]. In parallel, SSL enables representation learning from large-scale unlabeled medical data, reducing dependence on costly manual annotations while enhancing scalability and generalization across imaging modalities. These approaches are particularly valuable in edge-oriented healthcare ecosystems, where adaptive SSL supports low-latency inference, intelligent automation, and efficient learning under constrained computational resources [30,35].

Despite strong progress in transformer-based and multimodal AI systems, real-world deployment remains limited by interpretability issues, high computational complexity, and hardware constraints. ViTs improve long-range contextual modeling beyond CNNs but require substantial memory bandwidth due to global self-attention operations [21]. Lightweight variants such as Swin Transformers and distilled models reduce complexity but may sacrifice accuracy. Multimodal systems integrating imaging, biosignals, and EHRs enhance clinical intelligence but face challenges in alignment, missing data, privacy, and interpretability despite XAI methods.

Table 2. Comparative analysis of DL architectures for medical imaging systems

Architecture	Core Characteristics	Major Advantages	Limitations	Representative Applications	Edge Suitability
CNN	Hierarchical spatial learning and convolutional feature extraction [2,3]	Efficient inference and strong performance on limited datasets	Limited global contextual awareness [5]	Tumor detection, CT/MRI analysis, lesion segmentation	High
ViT	Self-attention and global contextual [21]	Superior long-range dependency learning and multimodal fusion [36]	High computational and memory cost [21]	Organ localization, radiomics, multimodal diagnosis	Moderate
GAN	Generator-discriminator adversarial learning framework [2,37]	Synthetic image generation and data augmentation	Training instability and limited interpretability	Image reconstruction, low-dose enhancement	Moderate
Hybrid CNN-Transformer	Integrated local convolutional and global attention mechanisms [12,36]	Balanced feature extraction with improved contextual understanding	Increased architectural complexity and hardware demand	Precision medicine and multimodal analytics	Moderate

Table 2 presents a comparative indication of major DL architectures used in intelligent medical imaging systems. It highlights their computational strengths, structural characteristics, limitations, and clinical applications. It also assesses their suitability for edge-based deployment in healthcare settings with limited facilities.

Table 3. Edge AI Frameworks for Intelligent Medical Imaging Systems

Framework	Supported Hardware	Optimization Features	Typical Latency	Clinical Relevance
TensorFlow Lite [27]	ARM, mobile GPUs, CPUs	Pruning, quantization, lightweight model conversion	30 ms	Portable diagnostics and embedded medical devices
ONNX Runtime [10]	CPU, GPU, TPU (cross-platform accelerators)	Graph optimization, hardware acceleration, TensorRT integration	20-50 ms	Real-time clinical inference systems
PyTorch Mobile [14]	ARM processors, mobile GPUs	Quantization-aware training, on-device deployment	50 ms	Flexible edge AI healthcare applications
TensorRT [11]	NVIDIA GPUs	INT8 precision optimization, kernel fusion, layer merging	<20 ms	Ultra-low latency medical imaging analytics
Vitis [38]	FPGA platforms	Hardware-aware compilation, FPGA acceleration	Low latency	Energy-efficient IoMT and edge inference

Table 3 summarizes widely adopted edge AI frameworks for intelligent medical imaging systems, focusing on supported hardware platforms, optimization strategies, inference latency, and clinical deployment relevance in resource-constrained environments.

Table 4. Comparative Analysis of Hardware Accelerators for Intelligent Medical Imaging Systems

Hardware	Performance Characteristics	Power Efficiency	Advantages	Limitations	Healthcare Applications
GPU	Extremely high parallel throughput [2,3]	Moderate	Excellent for large-scale DL training and inference	High energy consumption	Cloud-based imaging analytics and model training
FPGA	Reconfigurable, low-latency acceleration [27,38]	High	Energy-efficient, customizable architecture	High design complexity	Real-time edge imaging and IoMT systems
TPU	Optimized tensor computation [12,21]	Very high	High-speed inference performance	Limited flexibility outside tensor operations	Radiology AI and inference systems
ASIC / Edge TPU	Fixed-function ultra-low-power acceleration [11,15]	Highest	Minimal latency and energy consumption	Non-reconfigurable architecture	Portable diagnostic and embedded systems

Table 4 presents a comparative analysis of major hardware accelerators used in intelligent medical imaging, emphasizing computational performance, energy efficiency, architectural advantages, limitations, and healthcare applicability.

Table 5. Clinical Performance Benchmarks in Intelligent Medical Imaging Systems

Clinical Domain	AI Architecture	Dataset/Application	Evaluation Metric	Reported Performance
Brain tumor segmentation	CNN / U-Net [5]	BraTS	Dice coefficient	0.92-0.94
Retinal disease screening	CNN / ViT [21,36]	EyePACS	AUC	0.93-0.97
Lung cancer detection	CNN [2,37]	Chest CT / X-ray	AUC	0.86-0.94

Breast cancer diagnosis	CNN [2,3]	Mammography	Accuracy / AUC	0.87-0.91
COVID-19 detection	Deep CNN [17,39]	Chest imaging datasets	Accuracy	0.95

Table 5 summarizes clinical performance benchmarks of DL models across major medical imaging tasks, including classification, segmentation, and disease detection using widely adopted datasets and evaluation metrics.

3.8. Integrated Quantitative Performance Analysis

Recent advances in intelligent medical imaging demonstrate consistently strong empirical performance across diagnostic accuracy, computational efficiency, segmentation quality, and energy-aware deployment. DL models achieve high diagnostic accuracy across diverse imaging modalities, with reported AUC values exceeding 0.90 in multiple clinical classification tasks [3,37,39]. In classification-based applications, CNN-driven COVID-19 detection systems using CT and chest X-ray data achieve accuracies between 95% and 99%, while retinal disease screening systems report AUC values ranging from 0.93 to 0.99 [2]. Lung nodule detection systems similarly achieve AUC values between 0.86 and 0.94, demonstrating robustness in radiological interpretation tasks [5]. For segmentation tasks, U-Net and its variants consistently achieve Dice coefficients between 0.90 and 0.94 on the BraTS dataset [6]. Transformer-enhanced architectures such as TransUNet further improve segmentation accuracy by approximately 1-4% through enhanced global contextual modeling and long-range dependency capture [21].

From a hardware perspective, edge AI acceleration significantly improves latency and energy efficiency. Similarly, optimized frameworks such as ONNX Runtime, TensorRT, and TensorFlow Lite consistently achieve sub-50 ms inference latency for deployed medical imaging models [11,40]. These results collectively highlight the importance of hardware-aware optimization, lightweight model design, and edge deployment strategies for enabling scalable, real-time intelligent healthcare systems.

4. EDGE AI FOR RAPID REAL-TIME DIAGNOSIS

4.1. Foundations of Edge AI in Healthcare

Edge AI refers to the deployment of intelligent models directly on decentralized computing devices located near data acquisition sources. In healthcare environments, this paradigm enables localized processing of medical imaging and physiological signals without continuous reliance on centralized cloud infrastructures. This approach reduces inference latency, minimizes bandwidth consumption, improves system reliability, and enhances patient data privacy. Edge AI has therefore become a critical enabling technology for latency-sensitive applications such as emergency diagnostics, robotic-assisted surgery, wearable monitoring systems, and real-time clinical decision support [9,26].

The combination of DL, the IoMT, and distributed edge computing accelerated the development of intelligent diagnostic ecosystems suited efficiently in dynamic, resource-constrained clinical settings. According to recent research, edge intelligence is necessary to enable healthcare delivery systems that are responsive, scalable, and energy-efficient [8,17,41].

4.2. Edge-Cloud Collaborative Architectures

Modern intelligent healthcare systems increasingly employ edge-cloud collaborative architectures that distribute computational workloads across hierarchical infrastructures. In these frameworks, computationally intensive model training and large-scale data analytics are typically performed in cloud environments, while real-time inference and latency-sensitive tasks are executed at the network edge. This distributed paradigm optimizes computational efficiency, scalability, and diagnostic responsiveness. Edge-cloud continuum architectures have demonstrated substantial potential for supporting scalable healthcare analytics and remote diagnostic services, particularly in underserved and

resource-constrained regions [34]. By balancing workloads between centralized and decentralized infrastructures, these systems reduce communication delays, mitigate network congestion, and improve the reliability of AI-driven clinical decision-making. Recent studies also emphasize the importance of scalable AI orchestration and low-latency edge computing for real-time medical imaging applications [42,43].

4.3. Edge AI in Telemedicine and Remote Diagnostics

The integration of 5G communication networks, IoMT devices, and edge intelligence has significantly transformed telemedicine and remote healthcare infrastructures. Ultra-low-latency communication enables real-time medical image transmission, intelligent patient monitoring, remote diagnostics, and AI-assisted emergency response. Fog and edge computing architectures have demonstrated remarkable effectiveness in reducing diagnostic response times and enhancing healthcare accessibility in distributed clinical environments [19,44]. Edge-enabled imaging systems are particularly valuable in rural and resource-limited settings where cloud connectivity could be unstable or unavailable. Localized inference capabilities ensure continuous diagnostic operation even under intermittent network conditions, thereby improving healthcare resilience and patient outcomes. Emerging edge-aware healthcare frameworks integrating sensor fusion, AI-driven emergency response, and decentralized analytics further strengthen real-time monitoring and intelligent clinical intervention systems [45,46].

4.4. Federated Learning and Privacy Preservation

Medical imaging data are highly sensitive and governed by strict privacy, security, and regulatory requirements. Federated learning has emerged as a powerful privacy-preserving distributed learning paradigm that enables collaborative AI model training without transferring raw patient data across healthcare institutions. Instead, only model parameters or gradients are exchanged, significantly reducing privacy risks and regulatory concerns. Edge-based federated learning frameworks facilitate secure cross-institutional collaboration while maintaining patient confidentiality, data ownership, and compliance with healthcare governance standards. These frameworks are increasingly integrated into intelligent medical imaging systems to support multi-hospital AI training, decentralized diagnostics, and sustainable edge intelligence [18,47]. Furthermore, federated and edge-enabled architectures improve scalability, trustworthiness, and interoperability in next-generation healthcare ecosystems.

5. HARDWARE ACCELERATION FOR INTELLIGENT MEDICAL IMAGING

5.1. Need for Hardware Acceleration

DL workloads in medical imaging are computationally intensive due to high-resolution imaging data, complex neural architectures, multimodal analytics, and stringent real-time inference requirements. Conventional CPU-based systems often fail to satisfy the latency and throughput demands of critical clinical applications such as emergency diagnostics, intensive care monitoring, intraoperative imaging, and remote healthcare services. Consequently, hardware acceleration platforms have become indispensable for enabling efficient execution of parallel tensor operations, neural inference, image reconstruction, and multimodal data processing with reduced latency and energy consumption [10,12]. The convergence of Edge AI and hardware-aware optimization has further accelerated the deployment of intelligent medical imaging systems in portable, embedded, and resource-constrained healthcare environments.

5.2. Graphics Processing Units (GPUs)

GPUs remain the dominant acceleration platform for DL training and inference due to their massively parallel computational architecture. GPUs efficiently execute convolutional operations, image reconstruction pipelines, matrix multiplications, and multimodal learning frameworks, thereby substantially improving throughput in radiology, pathology, and large-scale clinical analytics [27].

GPU-enabled infrastructures have significantly accelerated medical image segmentation, disease classification, radiomics analysis, and real-time diagnostic workflows. Their scalability and compatibility with modern DL frameworks continue to make GPUs central to smart health systems.

5.3. Field Programmable Gate Arrays (FPGAs)

FPGAs provide reconfigurable hardware architectures optimized for low-power and low-latency AI inference. FPGA-based accelerators are highly suitable for real-time medical imaging applications requiring deterministic processing, energy efficiency, and hardware flexibility [27,38]. FPGAs are increasingly integrated into portable diagnostic systems, embedded imaging platforms, wearable healthcare devices, and edge-enabled clinical infrastructures. Recent FPGA-based multimodal edge intelligence systems have demonstrated remarkable efficiency for IoMT applications and decentralized healthcare analytics [45].

5.4. Application-Specific Integrated Circuits (ASICs) and TPUs

ASICs and tensor processing units (TPUs) provide highly optimized acceleration for deep neural network inference and training. These specialized processors enable ultra-fast medical image analytics while minimizing energy consumption and computational overhead. Healthcare-oriented ASICs and TPUs have facilitated the deployment of sophisticated AI imaging models within compact, energy-constrained, and portable clinical devices [12]. Recent advances in edge-oriented AI accelerators have further improved the feasibility of deploying intelligent diagnostic systems in real-time healthcare environments.

5.5. Neuromorphic and Heterogeneous Computing

Neuromorphic computers use event-driven designs with ultra-low power and adaptive processing to mimic organic neural computation. Neuromorphic systems show great promise for intelligent biosignal processing, real-time diagnostic inference, and continuous physiological monitoring in edge healthcare settings, even though they are still in their infancy [48]. Concurrently, modern medical imaging infrastructures are increasingly characterized by heterogeneous computing architectures that combine CPUs, FPGAs, GPUs, TPUs, and edge processors, offering greater scalability, flexibility, and adaptability to diverse workloads. In complex multimodal healthcare analytics, such hybrid designs enable more efficient and optimized resource allocation [27].

5.6. Hardware-Aware AI Optimization

Hardware-aware optimization techniques have become essential for deploying DL models on edge and embedded healthcare devices. Model compression approaches, including pruning, sparse computation, quantization, knowledge distillation, and mixed-precision computing, significantly reduce computational complexity, memory usage, and energy consumption while preserving diagnostic performance [13]. Lightweight neural architectures such as EfficientNet, MobileNet, TinyML frameworks, and sustainable edge AI models further enhance the feasibility of low-latency clinical deployment. Recent studies have emphasized carbon-aware and energy-efficient medical AI frameworks designed to support sustainable healthcare computing infrastructures [49,50].

6. CLINICAL APPLICATIONS OF INTELLIGENT MEDICAL IMAGING SYSTEMS

6.1. Radiology and Diagnostic Imaging

AI-driven radiology systems have demonstrated remarkable performance in chest, neuro, musculoskeletal, and abdominal imaging. DL models support automated lesion detection, image reconstruction, workflow prioritization, and radiological interpretation with improved efficiency and accuracy. ML is increasingly transforming radiology by reducing clinician workload and enhancing diagnostic consistency [2]. Edge-enabled radiological systems further support real-time inference and low-latency diagnostics in emergency and remote healthcare environments.

6.2 Oncology and Cancer Imaging

Cancer imaging represents one of the most impactful applications of intelligent medical imaging. AI systems support tumor segmentation, radiomics analysis, prediction of therapeutic response, treatment planning, and precision oncology. DL-based imaging frameworks have significantly improved early cancer detection and personalized treatment strategies across lung, breast, brain, and prostate cancers [34]. Advanced multimodal imaging intelligence integrating pathological, genomic, and radiological data is further accelerating precision medicine initiatives in oncology.

6.3. Ophthalmology

Retinal imaging and optical coherence tomography (OCT) analysis have become highly successful domains for AI-assisted diagnosis. Intelligent imaging systems enable automated detection of diabetic retinopathy, macular degeneration, glaucoma, and retinal abnormalities with high diagnostic accuracy. AI-assisted ophthalmic screening significantly improves early disease detection and accessibility to preventive healthcare services.

6.4. Histopathology and Digital Pathology

Digital pathology combines high-resolution whole-slide imaging with DL for automated cancer grading, cellular analysis, biomarker identification, and tissue characterization. AI-assisted pathology systems improve diagnostic consistency, reduce interobserver variability, and accelerate pathological interpretation workflows [1]. Recent transformer-based and multimodal pathology frameworks are further advancing precision diagnostics and computational pathology research.

6.5. Cardiology

DL models have demonstrated substantial effectiveness in cardiac imaging interpretation, arrhythmia detection, echocardiographic analysis, and cardiovascular risk prediction. Intelligent cardiac imaging systems support early disease diagnosis, personalized therapeutic planning, and continuous cardiovascular monitoring. Edge-enabled cardiology platforms further facilitate real-time analysis of physiological signals and wearable biosensor data for proactive healthcare management.

6.6. Remote Monitoring and IoMT

Wearable imaging devices, mobile diagnostic systems, and IoMT infrastructures increasingly integrate Edge AI for continuous patient monitoring and remote healthcare delivery. Intelligent wearable sensing systems have become essential for pandemic response, chronic disease management, and decentralized healthcare services [20,51]. Recent edge-aware healthcare frameworks integrating sensor fusion, AI-driven emergency response, and decentralized analytics are strengthening real-time remote monitoring capabilities and intelligent telemedicine systems [45].

As shown in Figure 2, the total number of publications increased more than tenfold during this period, demonstrating the rapidly growing global interest in the field. The most significant growth occurred after 2022, driven by the convergence of Edge AI, federated learning, hardware acceleration, and transformer-based architectures. Recent studies increasingly focus on real-time clinical deployment, explainability, energy efficiency, and privacy-preserving healthcare intelligence. In addition, sustainable AI and multimodal foundation models emerged as dominant research directions during 2025-2026, clinically deployable, reflecting a clear transition toward scalable, and environmentally responsible intelligent medical imaging systems.

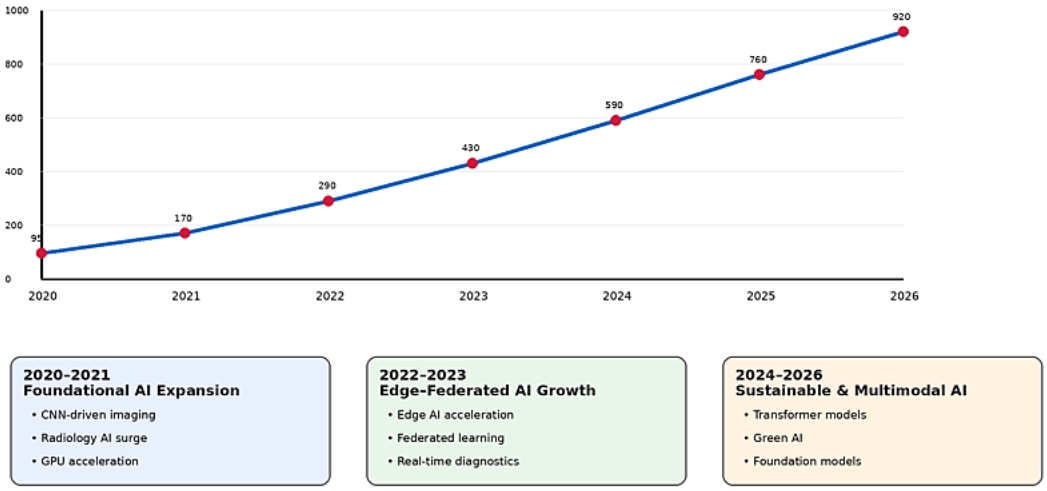

Figure 2 presents publication trends and research growth patterns in intelligent medical imaging systems, 2020-2026.

Furthermore, Figure 3 illustrates the rapid increase in global research activity across DL, Edge AI, federated learning, transformer architectures, and hardware-accelerated healthcare systems, reflecting a clear shift toward scalable, low-latency, and sustainable precision medicine. This growth is driven by the rapid expansion of DL applications in radiology and pathology, increasing demand for real-time, low-latency diagnostic systems, and the growing adoption of Edge AI and IoMT ecosystems. Advances in FPGA, GPU, TPU, and ASIC accelerators, federated AI, privacy-preserving healthcare systems, precision medicine, and telemedicine initiatives are further motivators. Furthermore, increased emphasis on explainable, trustworthy, and sustainable AI, alongside the expansion of multimodal imaging and foundation model research, continues to accelerate global innovation in intelligent medical imaging systems.

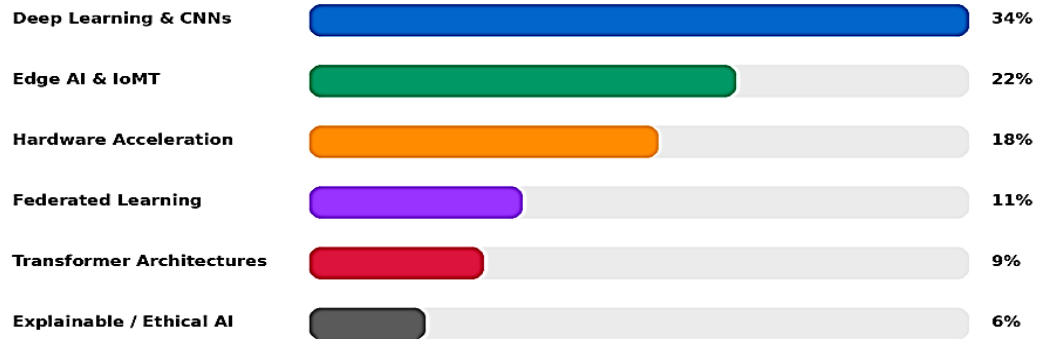

Figure 3. Distribution of Research Focus Areas in Intelligent Medical Imaging Systems (2020-2026)

7. EXPLAINABLE, TRUSTWORTHY, AND ETHICAL AI IN MEDICAL IMAGING

The integration of AI into clinical decision-making requires transparency, accountability, reliability, and trustworthiness. Black-box diagnostic technologies have the potential to erode doctors' trust and raise ethical questions about patient safety, privacy, justice, and bias. Consequently, XAI has emerged as a critical component of intelligent medical imaging systems. Explainable AI frameworks enable clinicians to interpret model predictions through saliency maps, attention visualization, feature attribution techniques, and interpretable decision pathways. Trustworthy AI mechanisms significantly

improve transparency, reduce algorithmic bias, and enhance clinical reliability in healthcare environments [18].

Algorithmic bias remains a major challenge due to demographic imbalance, heterogeneous datasets, institutional disparities, and unequal healthcare representation. Biased imaging models disproportionately affect underrepresented populations, thereby exacerbating healthcare inequities and diagnostic disparities. Regulatory compliance with frameworks such as GDPR, HIPAA, and emerging AI governance policies further necessitates secure, privacy-preserving, and ethically responsible healthcare AI infrastructures. Federated learning, privacy-preserving analytics, and fairness-aware machine learning are increasingly adopted to strengthen clinical trust and regulatory acceptance.

8. CHALLENGES AND LIMITATIONS

Despite significant advances, AI/ML-driven intelligent educational systems still encounter key technical, pedagogical, and operational constraints that limit scalable deployment. A major challenge is the scarcity of high-quality labeled educational data, as expert annotation of learner interactions, assessments, and behavioral logs is costly, time-intensive, and difficult to standardize across institutions. Furthermore, data heterogeneity, inconsistent collection protocols, and demographic imbalance reduce model robustness and generalization across diverse learner populations [3,37]. Modern transformer and multimodal architectures used in learning analytics offer strong representational capacity but introduce substantial computational and energy demands due to large parameter sizes and global attention mechanisms [21]. Real-time adaptive learning systems are additionally constrained by infrastructure and streaming bottlenecks. Multimodal educational data also suffers from missing modalities and limited interpretability [52]. While federated learning enhances privacy, it remains susceptible to bias propagation and security vulnerabilities. Overall, achieving an optimal balance among accuracy, efficiency, privacy, and explainability remains a central challenge.

Table 6. Challenges in trustworthy medical AI

Challenge	Clinical Impact	Existing Mitigation Strategies	Remaining Research Gaps
Dataset bias and heterogeneity [3,37]	Reduced generalizability and performance disparity across populations	Data augmentation, transfer learning, federated learning [18,29]	Lack of globally standardized, high-quality annotated medical imaging datasets
Privacy and security risks [29,32]	Exposure of sensitive patient data and reduced patient trust	Federated learning, encryption, secure aggregation protocols	Vulnerability to model poisoning, membership inference, and adversarial attacks
Limited explainability [18,53]	Reduced clinician trust and limited regulatory approval	Saliency maps, XAI frameworks, attention visualization methods	Persistent accuracy-interpretability trade-off in DL models
Computational complexity [12,21]	High latency, increased energy consumption, and deployment constraints	Quantization, pruning, model distillation, edge acceleration	Efficient transformer and multimodal model deployment on resource-limited edge devices
Hardware bottlenecks [15,27]	Delayed real-time MRI/CT processing and limited scalability	FPGAs, GPUs, TPUs, ASIC/Edge accelerators	Memory bandwidth, I/O constraints, and lack of unified hardware-software co-design
Regulatory and interoperability issues [9,52]	Slower clinical adoption and fragmented system integration	AI governance frameworks and standardization initiatives	Lack of unified global standards for healthcare AI interoperability and validation

Table 6 outlines the primary technical, clinical, and regulatory challenges facing intelligent medical imaging systems, along with their clinical implications, current mitigation strategies, and remaining research gaps that hinder widespread and reliable deployment in real-world healthcare settings.

Furthermore, Figure 4 presents an integrated framework for developing safe, reliable, and clinically deployable medical imaging AI systems, organized around six core pillars: explainability, fairness, governance, privacy, robustness, and clinician trust, all centered on a unified Trustworthy Medical Imaging AI core designed to ensure human-centered and ethically aligned decision support. The *fairness* component focuses on reducing dataset and sampling bias, ensuring demographic equity, and maintaining consistent performance across diverse patient populations. The *explainability* module emphasizes interpretable AI through saliency maps, attention mechanisms, and rationale generation to improve clinician understanding and support transparent decision-making. The *robustness* pillar ensures resilience against adversarial attacks, domain shifts, and noise, while supporting cross-site generalization and continuous model validation in real-world clinical environments.

The *governance* pillar ensures regulatory compliance (FDA, CE, and HIPAA/GDPR-aligned frameworks), documentation transparency, auditability, and accountability throughout the AI lifecycle. The *privacy* dimension incorporates data de-identification, encryption, federated learning, and secure communication protocols to protect sensitive patient information and comply with healthcare data regulations. The *clinician trust* component highlights human-AI collaboration, usability, workflow integration, and transparent communication of model confidence and limitations to promote adoption in clinical practice. The framework is supported by multimodal *inputs*, including medical imaging data (CT, MRI, X-ray, ultrasound, PET), clinical and EHR data, IoMT-generated distributed datasets, and high-quality annotated repositories. It is further strengthened by *foundation enablers* such as advanced AI algorithms, edge-cloud infrastructure, clinician training, interdisciplinary collaboration, and sustainable green AI technologies.

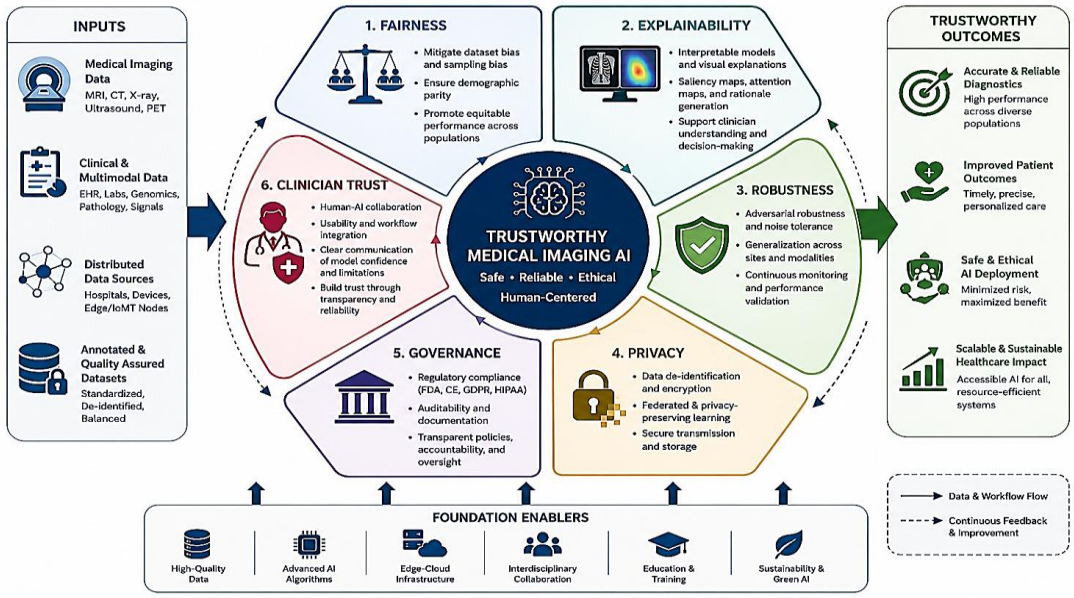

Figure 4. Proposed Framework for Trustworthy Medical Imaging AI

9. FUTURE RESEARCH DIRECTIONS

Sustainability has become a key concern in AI-driven healthcare, with carbon-aware and green computing approaches aimed at minimizing the energy cost of training and deploying large-scale

diagnostic models in hospital environments [49]. At the same time, continual learning under regulatory constraints remains challenging, as adaptive medical AI systems must integrate new clinical data, such as updated imaging devices, while complying with strict validation and re-certification standards. Privacy-preserving real-time imaging is also still emerging, particularly methods based on homomorphic encryption and secure enclaves, which struggle to meet ultra-low-latency demands in intraoperative applications. In addition, next-generation hardware paradigms, including neuromorphic architectures, photonic computing, and 6G-enabled edge intelligence, are expected to significantly transform distributed diagnostic systems and real-time clinical decision support [15,44,54].

10. CONCLUSION

Intelligent medical imaging systems that integrate DL, Edge AI, federated learning, and hardware acceleration are reshaping modern healthcare diagnostics and precision medicine. This review synthesizes these AI subfields into a unified framework to enhance clinical decision support systems that are accurate, low-latency, scalable, and patient-centric across diverse healthcare environments. DL models such as CNNs, transformers, segmentation networks, and self-supervised approaches have advanced image interpretation, disease detection, and multimodal clinical analytics. Concurrently, Edge AI and IoMT enable decentralized, privacy-preserving diagnostics, and real-time, particularly in resource-constrained settings. Hardware accelerators, including ASICs, FPGAs, GPUs, and TPUs, further improve computational efficiency for portable and high-throughput medical AI deployment. Despite progress, challenges persist, including data heterogeneity, annotation scarcity, privacy risks, interoperability issues, and explainability-performance trade-offs. Federated learning and explainable AI offer promising solutions, though issues such as model poisoning and regulatory uncertainty remain. Sustainable, energy-efficient AI design is increasingly critical for scalable deployment. Future systems require interdisciplinary collaboration to ensure robustness, security, ethical governance, and clinical trust. By integrating multimodal intelligence, edge-cloud learning, and federated frameworks, intelligent medical imaging can improve diagnostic accessibility, reduce disparities, and advance global precision medicine.

Funding: Not applicable.

Author Contributions: All authors contributed to the research and manuscript preparation, and have reviewed and approved the final version for publication.

Informed Consent Statement: Not applicable.

Conflict of Interest: The authors declare no competing interest.

Acknowledgements: During the preparation of this manuscript, the authors used grammarly for grammar and language editing, duplichecker for similarity checking, and chatgpt to improve clarity. All outputs were reviewed by the authors, who take full responsibility for the final content of the manuscript.

REFERENCES

[1] Ahmad, Z., Rahim, S., Zubair, M., & Abdul-Ghafar, J. (2021). Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on its potential and promise in pathology. *International Journal of Pathology and Clinical Research*, *7*(1), 1-11. https://doi.org/10.23937/2469-5807/1510138

[2] Wang, J., Zhu, H., Wang, S. H., & Zhang, Y. D. (2021). A review of deep learning on medical image analysis. *Mobile Networks and Applications*, *26*(1), 351-380. https://doi.org/10.1007/s11036-020-01672-9
[3] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., … & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. https://doi.org/10.1186/s40537-021-00444-8
[4] Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., & Yang, X. (2020). Deep learning in medical image registration: A review. *Physics in Medicine & Biology*, *65*(20), 20TR01. https://doi.org/10.1088/1361-6560/ab843e
[5] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3523-3542. https://doi.org/10.1109/TPAMI.2021.3059968
[6] Zhou, S. K., Greenspan, H., & Shen, D. (2021). *Deep learning for medical image analysis*. Academic Press. https://doi.org/10.1016/C2019-0-03742-9
[7] Hayyolalam, V., Aloqaily, M., Özkasap, Ö., & Guizani, M. (2021). Edge-assisted solutions for IoT-based connected healthcare systems: A literature review. *IEEE Internet of Things Journal*, *9*(12), 9419-9443. https://doi.org/10.1109/JIOT.2021.3128787
[8] Wang, K., Kong, S., Chen, X., & Zhao, M. (2024). Edge computing empowered smart healthcare: monitoring and diagnosis with deep learning methods. *Journal of Grid Computing*, *22*(1), 30.
[9] Gill, S. S., Golec, M., Hu, J., Xu, M., Du, J., Wu, H., … & Uhlig, S. (2025). Edge AI: A taxonomy, systematic review and future directions. *Cluster Computing*, *28*(1), 18. https://doi.org/10.1007/s10586-024-04785-2
[10] Deng, C., Fang, X., Wang, X., & Law, K. (2022). Software orchestrated and hardware accelerated artificial intelligence: Toward low latency edge computing. *IEEE Wireless Communications*, *29*(4), 110-117. https://doi.org/10.1109/MWC.001.2100483
[11] Ghani, A., Aina, A., & Hwang See, C. (2024). An optimised CNN hardware accelerator applicable to IoT end nodes for disruptive healthcare. *IoT*, *5*(4), 901-921. https://doi.org/10.3390/iot5040045
[12] Liu, Z., Bi, Z., Liang, C. X., Song, J., Wang, T., Zhang, Y., … & Song, X. (2025). Hardware accelerated foundations for multimodal medical AI systems: A comprehensive survey. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2502.12345
[13] Prasad, N. S., & Sundar, S. (2025). Comprehensive review on the exploitation of advanced memory optimization strategies to improve performance for convolutional and spiking neural networks in medical imaging using hardware accelerators. *IEEE Access*.
[14] Wang, T., Guo, J., Zhang, B., Yang, G., & Li, D. (2025). Deploying AI on edge: Advancement and challenges in edge intelligence. *Mathematics*, *13*(11), 1878. https://doi.org/10.3390/math13111878
[15] Corral, J. M. R., Civit-Masot, J., Luna-Perejón, F., Díaz-Cano, I., Morgado-Estévez, A., & Domínguez-Morales, M. (2024). Energy efficiency in edge TPU vs. embedded GPU for computer-aided medical imaging segmentation and classification. *Engineering Applications of Artificial Intelligence*, *127*, 107298. https://doi.org/10.1016/j.engappai.2024.107298
[16] Li, M., Jiang, Y., Zhang, Y., & Zhu, H. (2023). Medical image analysis using deep learning algorithms. *Frontiers in Public Health*, *11*, 1273253. https://doi.org/10.3389/fpubh.2023.1273253
[17] Sadeghi, A., Sadeghi, M., Fakhar, M., Zakariaei, Z., Sadeghi, M., & Bastani, R. (2024b). A deep learning-based model for detecting Leishmania amastigotes in microscopic slides: A new approach to telemedicine. *BMC Infectious Diseases*, *24*(1), 551. https://doi.org/10.1186/s12879-024-08551
[18] Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., … & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, *96*, 156-191. https://doi.org/10.1016/j.inffus.2023.03.008

[19] He, Q., Xi, Z., Feng, Z., Teng, Y., Ma, L., Cai, Y., & Yu, K. (2024). Telemedicine monitoring system based on fog/edge computing: A survey. *IEEE Transactions on Services Computing*, *18*(1), 479-498. https://doi.org/10.1109/TSC.2024.3362323
[20] Putra, K. T., Arrayyan, A. Z., Hayati, N., Damarjati, C., Bakar, A., & Chen, H. C. (2024). A review on the application of Internet of Medical Things in wearable personal health monitoring: A cloud-edge artificial intelligence approach. *IEEE Access*, *12*, 21437-21452. https://doi.org/10.1109/ACCESS.2024.3365402
[21] Xu, Y., Khan, T. M., Song, Y., et al. (2025). Edge deep learning in computer vision and medical diagnostics: A comprehensive survey. *Artificial Intelligence Review*, *58*, 93. https://doi.org/10.1007/s10462-024-11033-5
[22] Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, *19*, 221-248. https://doi.org/10.1146/annurev-bioeng-071516-044442
[23] Tsuneki, M. (2022). Deep learning models in medical image analysis. *Journal of Oral Biosciences*, *64*(3), 312-320. https://doi.org/10.1016/j.job.2022.04.001
[24] Gao, X., He, P., Zhou, Y., & Qin, X. (2024). A smart healthcare system for remote areas based on the edge-cloud continuum. *Electronics*, *13*(21), 4152. https://doi.org/10.3390/electronics13214152
[25] Siripurapu, S., Darimireddy, N. K., Chehri, A., Sridhar, B., & Paramkusam, A. V. (2023). Technological advancements and elucidation gadgets for healthcare applications: An exhaustive methodological review—Part I (AI, big data, blockchain, open-source technologies, and cloud computing). *Electronics*, *12*(3), 750. https://doi.org/10.3390/electronics12030750
[26] Bourechak, A., Zedadra, O., Kouahla, M. N., Guerrieri, A., Seridi, H., & Fortino, G. (2023). At the confluence of artificial intelligence and edge computing in IoT-based applications: A review and new perspectives. *Sensors*, *23*(3), 1639. https://doi.org/10.3390/s23031639
[27] Alcaín, E., Fernández, P. R., Nieto, R., MonteAprilor, A. S., Vilas, J., Galiana-Bordera, A., … & Torrado-Carvajal, A. (2021). Hardware architectures for real-time medical imaging. *Electronics*, *10*(24), 3118. https://doi.org/10.3390/electronics10243118
[28] Al Habsi, O., Sali, S. M., Meribout, A., Meribout, M., Almazrouei, S., & Seghier, M. (2025). Hardware acceleration in portable MRIs: State of the art and future prospects. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3624072
[29] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., … Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine*, *3*(1), 119. https://doi.org/10.1038/s41746-020-00323-1
[30] Mashmool, A., Delzanno, G., Saadatfar, H., Ahmad, A., Koschke, R., Alizadehsani, R., … & D'Agostino, D. (2026). Edge computing in healthcare using machine learning: A systematic literature review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *16*(1), e70069. https://doi.org/10.1002/widm.70069
[31] Alqudah, A. M., & Moussavi, Z. (2025). A review of deep learning for biomedical signals: Current applications, advancements, future prospects, interpretation, and challenges. *Computers, Materials & Continua*, *83*(3). https://doi.org/10.32604/cmc.2025.062099
[32] Rahman, M. A., Hossain, M. S., Alrajeh, N. A., & Guizani, N. (2020). B5G and explainable deep learning-assisted healthcare vertical at the edge: COVID-19 perspective. *IEEE Network*, *34*(4), 98-105.
[33] Younas, M. I., Iqbal, M. J., Aziz, A., & Sodhro, A. H. (2023). Toward QoS monitoring in IoT edge devices driven healthcare: A systematic literature review. *Sensors*, *23*(21), 8885. https://doi.org/10.3390/s23218885
[34] Gao, X., He, P., Zhou, Y., & Qin, X. (2024). Artificial intelligence applications in smart healthcare: A survey. *Future Internet*, *16*(9), 308. [无 DOI]
[35] Jamshidi, M., Moztarzadeh, O., Jamshidi, A., Abdelgawad, A., El-Baz, A. S., & Hauer, L. (2023). Future of drug discovery: The synergy of edge computing, internet of medical things, and deep learning. *Future Internet*, *15*(4), 142. https://doi.org/10.3390/fi15040142
[36] Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., … & Zhou, Y. (2024). TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of

transformers. *Medical Image Analysis*, *97*, 103280. https://doi.org/10.1016/j.media.2024.103280

[37] Sahiner, B., Pezeshk, A., Hadjiiski, L. M., Wang, X., Drukker, K., Cha, K. H., … & Giger, M. L. (2019). Deep learning in medical imaging and radiation therapy. *Medical Physics*, *46*(1), e1-e36. https://doi.org/10.1002/mp.13264

[38] Gaikwad, N. B., Khare, S. K., Mendhe, D., Mir, H., Kosta, S., & Acharya, U. R. (2025). FPGA SoC implementation of adaptive deep neural network based multimodal edge intelligence for internet of medical things. *IEEE Access*.

[39] Sadeghi, A., Sadeghi, M., Sharifpour, A., Fakhar, M., Zakariaei, Z., Sadeghi, M., … Hajati, F. (2024a). Potential diagnostic application of a novel deep learning-based approach for COVID-19. *Scientific Reports*, *14*(1), 280. https://doi.org/10.1038/s41598-024-00280

[40] Aryendu, I., & Wang, Y. (2024). Raider: Rapid AI diagnosis at edge using ensemble models for radiology. *IEEE Access*, *12*, 115546-115560.

[41] Makina, H., & Ben Letaifa, A. (2023). Bringing intelligence to edge/fog in Internet of Things-based healthcare applications: Machine learning/deep learning-based use cases. *International Journal of Communication Systems*, *36*(9), e5484. https://doi.org/10.1002/dac.5484

[42] Manduva, V. C. (2024). Scalable AI: Leveraging cloud and edge computing for real-time analytics. *International Journal of Scientific Research and Management (IJSRM)*, *12*(11), 1788-1813.

[43] Thota, R. C. (2024). Optimizing edge computing and AI for low-latency cloud workloads. *International Journal of Science and Research Archive*, *13*(1), 3484-3500. https://doi.org/10.30574/ijsra.2024.13.1.1987

[44] Vishweshwara, A., & Ramya, R. (2026). Transforming telemedicine: Reducing latency through edge computing and 5G—A review. *Biomedical Materials & Devices*, *4*(2), 1161-1174. https://doi.org/10.1007/s44174-026-00231-8

[45] Alshuhail, A., Alshahrani, A., Mahgoub, H., Ghaleb, M., Darem, A. A., Aljehane, N. O., … & Alzahrani, F. (2025). Machine edge-aware IoT framework for real-time health monitoring: Sensor fusion and AI-driven emergency response in decentralized networks. *Alexandria Engineering Journal*, *129*, 1349-1361. https://doi.org/10.1016/j.aej.2025.1349-1361

[46] Batool, I. (2025). Real-time health monitoring using 5G networks: Deep learning-based architecture for remote patient care. *JMIRx Med*, *6*, e70906. https://doi.org/10.2196/70906

[47] Ranganathan, R., Annamalai, A., Ruban, S., Mythili, S., Shuriya, B., & Balajishanmugam, V. (2025). Sustainable AI for medical imaging: A federated, edge-based approach to chest X-ray triage. In *2025 International Conference on Modern Sustainable Systems (CMSS)* (pp. 945-952). IEEE.

[48] Zhu, B., Shin, U., & Shoaran, M. (2021). Closed-loop neural prostheses with on-chip intelligence: A review and a low-latency machine learning model for brain state detection. *IEEE Transactions on Biomedical Circuits and Systems*, *15*(5), 877-897.

[49] Shaukat, F., Parwez, K., Ashraf, Z., Alhabeeb, A., Alotabi, F. A., & Alnfiai, M. M. (2026). MobileNet-Lite-Health: A sustainable edge AI framework for medical image classification and carbon-aware computing. *IEEE Access*. https://doi.org/10.1109/ACCESS.2026.000000

[50] Badawy, W. (2026). Low-power AI and signal processing for the edge: Tools, techniques, and applications. *Neural Computing and Applications*, *38*(9), 346. https://doi.org/10.1007/s00521-026-11936-0

[51] Benjumea, A., Ropero, J., Rivera-Romero, O., Dorronzoro-Zubiete, E., & Carrasco, A. (2020). Assessment of mobile monitoring and wearable sensors for continuous management of COVID-19 and chronic diseases. *Sensors*, *20*(18), 5202. https://doi.org/10.3390/s20185202

[52] Mulo, J., Liang, H., Qian, M., Biswas, M., Rawal, B., Guo, Y., & Yu, W. (2025). Navigating challenges and harnessing opportunities: Deep learning applications in Internet of Medical Things. *Future Internet*, *17*(3), 107. https://doi.org/10.3390/fi17030107

[53] Wang, S., Summers, R. M., & Yao, J. (2021). Machine learning and radiology. *Medical Image Analysis*, *71*, 102017. https://doi.org/10.1016/j.media.2021.102017

[54] Letaief, K. B., Shi, Y., Lu, J., & Lu, J. (2021). Edge artificial intelligence for 6G: Vision, enabling technologies, and applications. *IEEE Journal on Selected Areas in Communications*, *40*(1), 5-36.