

Multimodal Large Language Models in Healthcare: A Structured Narrative Review of Architectures, Datasets, Applications, Evaluation Challenges, and Future Directions

Salman ASHRAF^{1,*}, Rafia TEHSEEN², and Hira ASHRAF³

¹ Department of Computer Science, The University of Lahore, 1-Km Defence Road, Near Bhupatian Chowk, Lahore 54792, Punjab, Pakistan.

² Department of Computer Science, University of Sargodha, University Road, Sargodha 40100, Punjab, Pakistan.

³ Malaysian Institute of Information Technology, Universiti Kuala Lumpur, 1016 Jalan Sultan Ismail, 50250 Kuala Lumpur, Malaysia.

E-mails: salmanashraf.12@gmail.com, rafiатеhseen007@gmail.com, hira.ashraf@s.unikl.edu.my

* Author to whom correspondence should be addressed;

Received: 14 June 2026 / Accepted: 24 September 2026 / Published online: 30 September 2026

Abstract

This structured narrative review examines multimodal large language models (MLLMs) in healthcare, focusing on architectures, datasets, clinical applications, evaluation methods, deployment barriers and future research priorities. A systematic search was conducted across Scopus, Web of Science, PubMed, IEEE Xplore, ACM Digital Library, Google Scholar, arXiv and medRxiv. Sixty-four sources were included after duplicate removal, title and abstract screening, and full-text eligibility assessment. Evidence indicates that healthcare MLLMs are most mature in image-text tasks, particularly radiology report generation and medical visual question answering, while signal-text, audio-text, video-text and longitudinal electronic health record applications remain less validated. Domain-adapted encoders, explicit cross-modal alignment, instruction tuning and grounded output constraints appear more clinically reliable than general-purpose prompting alone. Current systems remain unsuitable for autonomous clinical deployment because of hallucination, limited external validation, bias, privacy constraints and weak prospective evidence.

Keywords: Multimodal Large Language Models (MLLM); Medical Artificial Intelligence; Healthcare Diagnostics; Clinical Decision Support (CDS); Clinical Validation.

Introduction

Artificial intelligence has evolved from rule-based expert systems to machine learning, deep learning, large language models, vision-language models, and more recently, multimodal large language models [1-11]. Early clinical expert systems such as MYCIN demonstrated that formalized medical knowledge could support diagnosis and treatment recommendations, but such systems relied on manually encoded rules and had limited scalability [1]. Machine learning and deep learning later enabled models to learn patterns directly from large datasets, improving performance in imaging, natural language processing, biosignal interpretation, and prediction tasks [1-5]. The transformer architecture further transformed sequence modeling by enabling attention-based representation learning, leading to modern large language models (LLMs) capable of text generation, summarization, reasoning, and clinical question answering [1-5].

Despite their value, text-only LLMs are incomplete for healthcare because clinical decision-making rarely depends solely on text. In routine care, clinicians combine symptoms, medical history, laboratory values,

radiological images, pathology slides, electrocardiogram (ECG) signals, procedural videos, speech patterns, and longitudinal patient records. This multimodal reasoning is essential for complex tasks, such as differentiating pneumonia from heart failure, interpreting cancer progression, triaging emergency patients, and assessing mental health risks.

Multimodal large language models (MLLMs) are systems that connect non-textual data encoders to a language-model backbone so that the model can interpret, align and generate clinically meaningful outputs across modalities. General models such as GPT-4, Gemini, Flamingo and LLaVA demonstrate broad multimodal capability [7-11]. Medical-domain models, including LLaVA-Med, Med-Flamingo, Med-PaLM M, BiomedGPT, BiomedCLIP, MAIRA-2, RadGPT and OmniV-Med, adapt these ideas to biomedical image-text data, radiology reporting, medical visual question answering, pathology, genomics and three-dimensional imaging [12-19].

Several gaps in the literature justify this review. First, many papers emphasize benchmark scores without explaining why some model families perform better than others in clinical tasks. Second, model comparisons are often fragmented by specialty, modality or dataset, making it difficult to judge clinical readiness. Third, evaluation remains inconsistent: lexical report-generation scores, visual question answering accuracy and expert review measure different dimensions of safety and usefulness. Fourth, most discussions of healthcare MLLMs do not sufficiently connect architecture choices with deployment risks such as hallucination, bias, privacy, interpretability, latency and accountability. The aim of this structured narrative review was to synthesize current evidence on healthcare MLLMs by comparing architectural approaches, dataset types, clinical applications, evaluation practices and deployment limitations, with explicit attention to temporal trends, model readiness and clinically meaningful future directions.

Methodology

The review was informed by transparent search and reporting principles used in evidence synthesis, including the PRISMA 2020 reporting framework, while recognizing that the objective was broad conceptual synthesis rather than quantitative pooling of study outcomes [20, 21]. The aim was to map model families, architectures, datasets, clinical applications, evaluation metrics, and deployment barriers relevant to MLLMs in healthcare settings [22].

The search covered peer-reviewed articles, conference papers, preprints, and authoritative technical reports because several influential MLLM systems are first released as arXiv papers or institutional technical reports before journal publication. The databases and sources included Scopus, Web of Science, PubMed, IEEE Xplore, ACM Digital Library, Google Scholar, arXiv, medRxiv, and official model or dataset documentation where needed. Searches combined terms such as “multimodal large language model healthcare,” “medical MLLM,” “medical vision-language model,” “biomedical foundation model,” “LLaVA-Med,” “Med-Flamingo,” “Med-PaLM Multimodal,” “radiology report generation,” “medical visual question answering,” “audio text healthcare AI,” “ECG language model,” “biosignal text multimodal,” “surgical video language model,” “clinical hallucination,” and “AI clinical validation.” Searches covered literature published from 2017 to 2026, with earlier seminal sources retained for foundational artificial intelligence and medical informatics context.

Studies were included when they addressed at least one of the following areas: healthcare MLLM architecture, medical vision-language modeling (VLM), multimodal medical datasets, medical visual question answering, radiology report generation, signal-text or audio-text reasoning, video-text clinical modeling, multimodal clinical decision support, hallucination, interpretability, fairness, privacy, evaluation metrics or clinical deployment. Studies were excluded if they were purely unimodal without implications for multimodal reasoning, unrelated to healthcare, duplicated, inaccessible, or lacked sufficient technical or clinical relevance. Table 1 summarizes the review protocol and data extraction structure used to guide the synthesis.

Table 1. Review protocol and data extraction framework.

Component	Specification	Purpose
Review design	Structured narrative review	Broad conceptual synthesis across architecture, datasets, applications, evaluation, and deployment barriers.
Sources searched	Scopus, Web of Science, PubMed, IEEE Xplore, ACM Digital Library, Google Scholar, arXiv, medRxiv, official documentation	Important MLLM literature includes journal articles, conference papers, preprints, and technical reports.
Core search terms	MLLM, medical VLM, biomedical foundation model, radiology report generation, medical Visual Question Answering (VQA), biosignal-text, audio-text, surgical video, hallucination, clinical validation	Terms were combined with model names and clinical specialties.
Inclusion criteria	Healthcare relevance; multimodal architecture, dataset, benchmark, clinical task, or evaluation focus; English-language sources	Priority was given to sources with clear technical or clinical contribution.
Exclusion criteria	Purely unimodal papers without multimodal relevance; non-healthcare AI; duplicated or insufficiently described sources	Exclusion reduced scope drift and unsupported claims.
Data extraction fields	Model, year, modality, architecture, dataset, clinical domain, task, metrics, limitations, validation status	Fields were used to create Tables 2-7.
Synthesis approach	Narrative comparison and thematic grouping	No meta-analysis was conducted because datasets, tasks, and metrics were heterogeneous.

The synthesis in the following sections is therefore interpretative and comparative. It does not claim that all MLLMs outperform all unimodal models uniformly. Instead, it identifies domains where multimodal models show evidence of improved task capability and domains where the evidence is still preliminary or insufficient for clinical deployment.

The search and screening process used for this structured review is summarized in Figure 1, which reports the number of records identified, duplicates removed, records screened, full-text sources assessed, exclusions with reasons, and sources finally included in the narrative synthesis.

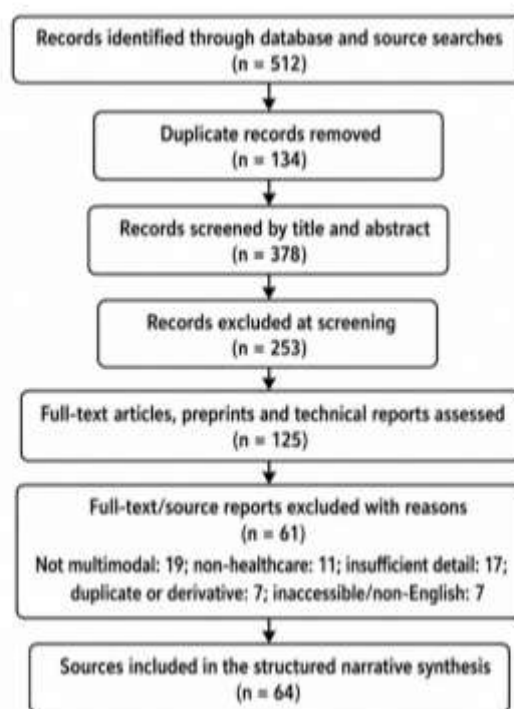


Figure 1. PRISMA-style flow diagram for source identification, screening, eligibility assessment

The search identified 512 records or source documents. After removing 134 duplicates, 378 records were screened by title and abstract. Of these, 253 were excluded because they did not address healthcare, multimodal reasoning, model architecture, dataset development, evaluation or deployment. One hundred twenty-five full-text articles, preprints and technical reports were assessed. Sixty-one were excluded with documented reasons, leaving 64 sources for the final synthesis.

Comparative Findings and Temporal Trends

The evidence does not support a single universal best architecture. The strongest current evidence is for domain-adapted image-text systems in radiology and biomedical visual question answering, where paired images and reports provide a natural training signal [12-18, 23-30]. General-purpose MLLMs are useful for broad reasoning and flexible interaction, but medical adaptation is usually necessary for clinical vocabulary, anatomical grounding, report factuality and error control.

Across model families, better performance is usually associated with four design choices: domain-specific pretraining on biomedical image-text or report data; explicit cross-modal alignment through contrastive learning, projection layers or cross-attention; instruction tuning on clinically framed tasks; and output constraints that encourage factual, localized and auditable responses [12-19, 41-46]. Models tend to underperform when non-text encoders are weakly aligned to the language backbone, when training data are derived from noisy captions or synthetic instructions alone, or when evaluation uses only generic language metrics.

The temporal trend is clear. From 2017 to 2021, the field was dominated by transformer methods, image-text pretraining and dataset construction. In 2022 and 2023, general vision-language models and medical adaptations accelerated the development of biomedical visual question answering and few-shot reasoning. From 2024 onward, the literature increasingly emphasizes grounded radiology generation, unified medical visual encoders, three-dimensional imaging, clinical factuality and deployment safety. Despite this progress, prospective validation and post-deployment monitoring remain rare compared with retrospective benchmark studies. Table 2 compares the main development periods of healthcare multimodal large language models.

Table 2. Comparative trends in healthcare multimodal large language model development and readiness.

Period	Dominant development	Best-supported approach	Reason for stronger performance	Remaining gap
2017-2021	Transformers, contrastive image-text learning and public dataset construction	Task-specific encoders and early image-text alignment	Paired reports and images enable supervised or weakly supervised representation learning	Limited clinical workflow validation
2022-2023	General vision-language models and first medical adaptations	Biomedical pretraining plus instruction tuning	Clinical vocabulary and task framing improve medical visual question answering and report-related tasks	Reliance on curated benchmarks and synthetic instruction data
2024-2025	Grounded radiology generation, three-dimensional imaging and unified medical visual understanding	Domain-specific grounding, factuality checks and multi-task biomedical training	Localization, structured output and benchmark diversity improve interpretability and error detection	External validation, fairness auditing and prospective trials remain insufficient
Deployment phase	Human-in-the-loop decision support rather than autonomous diagnosis	Constrained generation with retrieval, uncertainty and clinician review	Safety controls reduce risk from hallucination and unsupported recommendations	Regulatory pathways and post-market monitoring remain immature

Note: Representative sources include transformer and vision-language foundations [4, 7-11], medical adaptations [12-19], datasets [23-40], evaluation studies [41-46] and clinical deployment guidance [55-62].

Table 2 shows how the field has moved from task-specific image-text alignment toward biomedical instruction tuning, grounded radiology generation, unified medical visual understanding, and human-in-the-loop clinical decision support.

Architecture of Multimodal Large Language Models

Healthcare MLLMs are typically built from five interacting components: input modality streams, modality-specific encoders, alignment or connector modules, LLM backbone, and decoder or output-generation layer. Figure 2 shows the generalized architecture. The same model family may support different inputs depending on training data and design: for example, a radiology MLLM may accept chest X-rays and reports, whereas a surgical model may require video, instrument tracking, and procedural text.

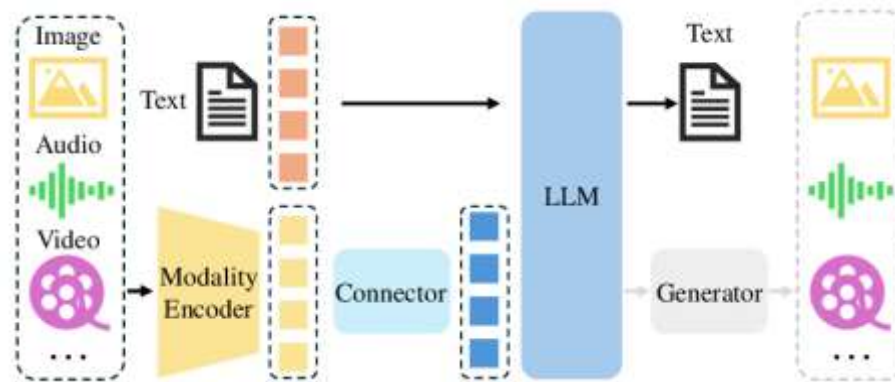


Figure 2. General architecture of healthcare multimodal large language models.

Modality-Specific Encoders

Encoders convert raw modality inputs into numerical embeddings that are suitable for alignment and reasoning. Vision encoders process X-rays, computed tomography (CT) scans, magnetic resonance imaging (MRI) scans, ultrasound images, fundus images, dermoscopic images, and pathology slides. Common architectures include convolutional neural networks, Vision Transformers, Swin Transformers, CLIP-style visual encoders, and domain-specific variants such as BiomedCLIP [16]. Text encoders and LLM backbones process radiology reports, discharge summaries, clinical guidelines, patient messages and biomedical literature. Audio encoders, such as Whisper and wav2vec 2.0 process speech, respiratory sounds, heart sounds, coughs, and other clinically informative audio streams. Biosignal encoders process ECG, EEG, EMG, PPG, and ICU waveform data using one-dimensional CNNs, recurrent networks, transformers, or signal-specific architectures. Video encoders process surgical procedures, rehabilitation movements, endoscopy, and behavioral observations.

As shown in Table 3, the choice of encoder shapes performance, interpretability, computational cost, and clinical suitability. In clinical contexts, encoder selection should be guided not only by benchmark accuracy but also by data availability, modality quality, temporal alignment, noise characteristics, device variability, and the possibility of clinical audit.

Alignment and Fusion Mechanisms

Fusion combines representations from several modalities into a coherent representation for reasoning. Early fusion combines modality embeddings before deep modeling, allowing for early cross-modal interactions but requiring careful synchronization and balanced representation. Late fusion processes each modality independently and combines predictions or generates outputs at a later stage, improving modularity but reducing deep interaction. Hybrid fusion combines both approaches and is common in MLLMs because it permits modality-specific processing while enabling cross-attention and shared reasoning [9-19].

Alignment ensures that embeddings from different modalities correspond to the same clinical meaning. Contrastive image-text learning aligns radiographs or biomedical figures with reports or captions; projection layers, Q-Former modules and cross-attention mechanisms connect non-text encoders to language backbones; and instruction tuning adapts models to clinical tasks [9, 12-18]. Weak alignment can produce fluent but ungrounded outputs, such as a plausible radiology sentence that is not supported by the image. For this reason, alignment should be evaluated through grounding metrics, expert review, cross-modal consistency checks and clinical factuality tests [41-46].

Table 3. Modality encoders, clinical suitability, and major limitations.

Modality	Typical inputs	Representative encoders	Clinical suitability	Main limitations
Text	Clinical notes, reports, guidelines, patient messages	BERT, BioBERT, PubMedBERT, LLM tokenizers/backbones	High value for summarization, QA, coding, guideline support	Sensitive to incomplete notes, ambiguous abbreviations, hallucination, outdated knowledge
Images	X-ray, CT, MRI, ultrasound, OCT, pathology, dermoscopy	CNNs, ViT, Swin Transformer, CLIP/BiomedCLIP-style encoders	Mature evidence in radiology and pathology; supports report generation and VQA	Needs high-quality labels, external validation, visual grounding and skin-tone/device diversity
Audio	Speech, respiratory sounds, heart sounds, cough, clinical conversation	Whisper, wav2vec 2.0, SpeechT5, spectrogram-based CNNs	Useful for telemedicine, mental health, respiratory and cardiac monitoring	Privacy-sensitive, language/accent bias, noise and recording variability
Biosignals	ECG, EEG, EMG, PPG, ICU waveform data	1D-CNN, transformers, temporal encoders, signal-specific networks	Important for cardiology, neurology, ICU and remote monitoring	Temporal misalignment, artifacts, device variability, missing signals
Video	Surgery, rehabilitation, endoscopy, behavioral assessment	3D CNNs, temporal transformers, video-language encoders	Supports procedural guidance, step recognition and movement assessment	High annotation cost, privacy issues, latency and high compute demand
Structured data	Labs, medications, demographics, vital signs, diagnoses	Tabular transformers, embedding layers, feature projection modules	Supports risk prediction and patient stratification	Coding inconsistency, missingness, confounding and fairness concerns

Decoders and Clinical Output Generation

The decoder or language-generation component transforms fused embeddings into outputs such as free-text reports, structured summaries, diagnostic suggestions, international classification of diseases (ICD) codes, systematized nomenclature of medicine clinical terms (SNOMED CT) concepts, visual grounding explanations, triage labels, or patient-facing educational messages. Autoregressive decoders based on LLMs are flexible, but this flexibility also increases the risk. A fluent decoder may produce clinically unsafe text if the input evidence is incomplete, the model is poorly calibrated, or the output is not constrained by the clinical knowledge.

Therefore, clinical deployment requires output control. These include template-constrained generation for reports, terminology grounding through unified medical language system (UMLS) or SNOMED CT, uncertainty estimates, retrieval-augmented grounding with current guidelines, safety filters, audit logs, and mandatory clinician reviews. In high-risk settings, MLLM outputs should be treated as decision-support suggestions, rather than autonomous diagnostic decisions.

Representative Healthcare Multimodal Large Language Models

Healthcare MLLMs differ in terms of modality coverage, training strategy, dataset source, task focus, and clinical maturity. General-purpose MLLMs, such as GPT-4, Gemini, Flamingo, and LLaVA, show broad multimodal capabilities, but they may lack domain-specific medical grounding [7, 8, 10, 11]. Medical-domain MLLMs attempt to address this gap through biomedical pre-training, curated image-text pairs, instruction tuning, clinical benchmarks, and expert evaluation. Table 4 compares the representative systems relevant to healthcare.

Table 4. Representative MLLMs and medical vision-language models relevant to healthcare.

Model	Year	Modalities	Architecture / strategy	Main use case	Strength	Clinical readiness
GPT-4 / GPT-4V [7]	2023	Text + image	Transformer-based multimodal model	Broad reasoning and image-text tasks	Strong general capability; limited transparency and clinical specialization	Not sufficient for autonomous clinical use without validation
Gemini [8]	2023	Text, image, audio, video	General multimodal model family	General multimodal reasoning	Broad modality scope; not primarily a medical model	Requires domain-specific safety evaluation
Flamingo [10]	2022	Image + text	Vision-language few-shot learner	Few-shot image-text reasoning	Influential architecture for interleaved image-text learning	General-domain model; clinical use requires adaptation
LLaVA [11]	2023	Image + text	Vision encoder + Vicuna/LLM with visual instruction tuning	General visual instruction following	Open architecture and strong VQA baseline	Medical performance depends on adaptation
LLaVA-Med [12]	2023	Biomedical images + text	Biomedical image-text alignment and instruction tuning	Biomedical VQA and dialogue about images	Trained from biomedical figure-caption data and GPT-4-generated instructions	Requires careful validation for diagnostic use
Med-Flamingo [13]	2023	Medical images + text	Medical adaptation of OpenFlamingo	Few-shot medical VQA	Supports few-shot generative medical VQA and rationale generation	Early-stage; limited deployment evidence
Med-PaLM M [14]	2023	Clinical text, imaging, genomics and other biomedical tasks	Generalist biomedical multimodal model	MultiMedBench tasks; radiology report generation and summarization	Shows broad biomedical task capability	Proof-of-concept; real-world validation still needed
BiomedCLIP [16]	2023	Biomedical images + text	CLIP-style contrastive pretraining on PMC image-text pairs	Classification, retrieval and VQA	Large-scale biomedical image-text foundation model	Primarily representation model, not a full clinical assistant
BiomedGPT [15]	2024	Biomedical image + text	Generalist vision-language foundation model	Classification, captioning, VQA, summarization	Open-source and computing-friendly biomedical VLM	Clinical readiness depends on task-specific validation
MAIRA-2 [17]	2024	Chest X-ray + report context	Radiology-specific grounded report generation	Grounded radiology report generation	Focuses on factuality and spatial localization	Mainly chest radiography; needs broader validation
RadGPT [18]	2025	3D CT + text	Anatomy-aware vision-language pipeline	3D tumor report generation and dataset construction	Supports structured tumor reporting from CT	Newer model; external clinical validation required
OmniV-Med [19]	2025	2D, 3D, video + text	Unified medical visual encoder + LLM	Universal medical visual understanding	Extends beyond 2D image-text tasks	Preclinical evidence; implementation cost and validation remain barriers

Note: Rows summarize primary model papers and technical reports [7-19]. Clinical readiness was interpreted conservatively according to validation and deployment evidence rather than benchmark performance alone.

The comparison in Table 4 indicates that healthcare MLLMs are not a single, homogeneous technology. Some are broad general-purpose systems, some are medical image-text systems, some are representation models, and some are domain-specific report generators. This distinction is important for journal reporting and clinical

interpretation. A model that performs well on biomedical figure visual question answering (VQA) is not automatically suitable for emergency triage or treatment recommendations. Conversely, a radiology-specific model may be clinically valuable within its narrow workflow but is limited outside that context.

Multimodal Healthcare Datasets and Benchmarks

The quality of the dataset is one of the main determinants of MLLM performance. Healthcare datasets are difficult to build because they require de-identification, expert annotation, secure governance, standardized coding, and careful handling of demographic and institutional biases. Image-text datasets are the most mature because radiology and pathology often produce paired images and reports. Audio-text, signal-text, and video-text datasets are less mature but are increasingly important for telemedicine, cardiology, neurology, mental health, emergency care, and procedural guidance.

Table 5 summarizes the major dataset categories and examples. The table highlights that most available resources remain modality-specific or are weakly paired. Truly comprehensive patient-level datasets that integrate notes, labs, images, genomics, signals, audio, and video are rare because of privacy, interoperability, and annotation constraints.

Table 5. Major multimodal healthcare dataset categories and their relevance for MLLMs.

Dataset category	Examples	Modalities	MLLM relevance	Limitations
Image-text: radiology	MIMIC-CXR; CheXpert; PadChest; IU X-ray	X-ray images with reports or labels	Report generation, disease classification, VQA, grounding	Richest public evidence base; labels may be noisy or institution-specific
Image-text: pathology/biomedical figures	PathVQA; PMC-15M; ROCO	Pathology images, figures and captions or QA pairs	Biomedical VQA, retrieval, image-text alignment	Broad biomedical coverage; may differ from routine clinical data
Signal-text	PhysioNet; MIT-BIH; PTB ECG; MIMIC waveform; CHB-MIT	ECG, EEG, ICU signals with labels or notes	Arrhythmia, seizure detection, ICU monitoring, explanation generation	Temporal noise, missingness and device variability are major issues
Audio-text	DAIC-WOZ; ICBHI respiratory sounds; clinical speech/transcription datasets	Speech, respiratory or auscultation audio with transcripts or labels	Mental health screening, respiratory assessment, telemedicine documentation	Privacy, accent/language bias and recording conditions reduce generalizability
Video-text	Cholec80; M2CAI; endoscopy/surgery and rehabilitation datasets	Procedural or movement video with annotations	Surgical step recognition, procedural guidance, rehabilitation feedback	High annotation burden, patient identifiability and computational cost
Synthetic or weakly labeled multimodal data	Simulator outputs, generated captions, pseudo-labels, instruction data	Generated labels or paired examples	Pretraining, rare disease augmentation, instruction tuning	Risk of unrealistic labels, propagated errors and hallucination

Note: Dataset examples and limitations are synthesized from public radiology, biomedical image, signal, audio and video resources [23-40].

The MIMIC-CXR dataset contains hundreds of thousands of de-identified chest radiographs paired with free-text reports, making it central to radiology image-text research [23]. CheXpert provides a large chest radiograph dataset with uncertainty labels and expert comparisons, supporting disease classification and report-derived label evaluation [24]. PadChest expands chest X-ray resources with reports from a Spanish hospital, enabling cross-language and external setting comparisons [25]. PathVQA supports pathology image question answering and demonstrates the difficulty in building medically meaningful VQA resources [28].

The scarcity of fully paired multimodal clinical datasets means that the reported benchmark gains should be interpreted cautiously. Many models have been evaluated using retrospective datasets, curated benchmarks, or

synthetic instruction data. For clinical translation, future datasets should include multi-institutional sampling, demographic metadata for fairness analysis, device and acquisition metadata, longitudinal temporal structure, expert adjudication, and explicit documentation of intended use.

Clinical Applications of Multimodal Large Language Models

The clinical applications of MLLMs can be grouped by specialty and task. The strongest current evidence is in image-text reasoning, especially radiology and biomedical VQA. Applications involving audio, biosignals, video and longitudinal electronic health records are promising but less mature. The distinction between promise and readiness is important: a model can be useful for drafting or summarization while still being unsafe for autonomous diagnosis. Table 6 provides a structured cross-specialty summary, and the subsections that follow discuss the major areas in more detail.

Table 6. Clinical application areas of MLLMs in healthcare.

Clinical area	Typical modalities	Potential tasks	Main opportunity	Key risk
Radiology	X-ray, CT, MRI, reports, prior imaging	Report generation, VQA, urgent finding detection, visual grounding	Mature image-text datasets and clear workflow value	False positives/negatives, hallucinated findings, limited external validation
Pathology and oncology	Whole-slide images, CT/MRI/PET, pathology reports, genomics	Tumor detection, staging support, report summarization, tumor-board support	High value for multimodal cancer reasoning	Very large images, high annotation burden, genomic integration complexity
Cardiology	ECG, echocardiography, cardiac MRI, auscultation, notes	Arrhythmia explanation, heart failure risk, murmur detection, triage	Combines waveforms with clinical context	Noise, device variability, temporal alignment and calibration
Mental health	Speech, text, video, behavioral cues, clinical notes	Depression screening, suicide-risk triage, therapy support	Multimodal signals capture affect beyond text	High ethical risk, privacy, consent and bias concerns
Emergency and ICU care	Vitals, labs, ECG, imaging, notes, waveforms	Triage, early warning, deterioration prediction	Time-critical integration of multiple streams	Latency, missing data, false alarm risk and accountability
Surgery and procedures	Video, tool motion, audio instructions, operative notes	Step recognition, guidance, error detection, documentation	Direct procedural context	Real-time requirements and limited annotated video
Dermatology and ophthalmology	Dermoscopic images, skin photos, fundus images, OCT, patient history	Screening, lesion localization, tele-specialist support	Visual domains suited to multimodal history-image reasoning	Representation bias, device variation, consent for images
Primary care and chronic disease	Symptoms, EHR, labs, wearables, messages	Visit summarization, risk stratification, patient education, monitoring	High workflow utility	Broad scope increases safety and liability risk
Rehabilitation and pediatrics	Video, sensors, EMG, parent reports, speech	Movement assessment, developmental screening, progress reports	Supports remote monitoring and tailored feedback	Small datasets, vulnerable populations and home-environment variation
Rare diseases	Phenotype text, images, genomics, longitudinal notes	Differential diagnosis and phenotype matching	Few-shot reasoning may help data-scarce domains	High hallucination risk and need for specialist confirmation

Note. Clinical domains are summarized from medical model, dataset, benchmark and deployment literature [12-19, 23-40, 55-62].

Radiology

Radiology is the most developed application area for healthcare MLLMs because it naturally pairs images with reports. Models such as LLaVA-Med, Med-Flamingo, Med-PaLM M, BiomedCLIP, and MAIRA-2 demonstrate the value of image-text alignment for radiology VQA, report generation, and grounded explanation [12-17]. As summarized in Table 6, radiology applications include automated report drafting, abnormality detection, patient-history-aware image interpretation, urgent finding prioritization, and report simplification.

However, report generation should not be evaluated solely using generic language metrics. A report can have a high lexical similarity to a reference report while still missing clinically important findings. Therefore, models should be evaluated using clinical label accuracy, error categorization, factuality, grounding, localization, radiologist review, and prospective workflow studies. MAIRA-2 is important in this context because it explicitly emphasizes grounded report generation and factual evaluation [41-46].

Pathology, Oncology, and Precision Medicine

Pathology and oncology require the integration of images, reports, staging information, genomic data, treatment history, and longitudinal follow-up. MLLMs may support tumor grading, biopsy summarization, radiology-pathology correlation, tumor board preparation, and precision oncology. RadGPT illustrates the emerging use of 3D CT information for tumor-report generation and dataset construction [18]. BiomedGPT and Med-PaLM M indicate a broader movement toward generalist biomedical models that combine multiple tasks and data types [14, 15].

Oncology applications remain clinically relevant. A hallucinated tumor location, incorrect stage, or unsupported treatment recommendation could cause harm. Therefore, oncology MLLMs should include source-grounded outputs, explicit uncertainty, structured reporting templates, clinician verification, and a strong separation between summarization support and treatment decision-making.

Cardiology and Physiological Signal Interpretation

Cardiology is a natural domain for signal-text multimodal learning because ECG signals, auscultation, echocardiography, laboratory results, symptoms and medication histories are often interpreted together. Public physiological resources have supported signal analysis and can contribute to signal-text modeling when paired with clinical labels or notes [32-35].

The main barriers are temporal alignment, signal noise, device heterogeneity, artifact detection, calibration, and lack of explainability. Unlike static text, physiological signals are dynamic; therefore, evaluation should include temporal robustness, artifact sensitivity, clinically impossible signal detection, and safety under missing or corrupted data.

Mental Health, Speech, and Behavioral Applications

Mental health assessment is inherently multimodal because language content, speech prosody, pauses, facial expressions, interaction patterns, and clinical notes may all be informative. Audio-text and video-text MLLMs may support depression screening, suicide risk triage, therapy documentation, and telepsychiatry support. Datasets such as DAIC-WOZ have helped establish multimodal depression detection research, but clinical deployment remains ethically sensitive [36].

The risks to mental health are substantial: privacy invasion, overinterpretation of affect, cultural bias, false reassurance, inappropriate crisis response, and lack of accountability. MLLMs should not replace mental health professionals. Instead, they may assist with documentation, symptom tracking, and triage when used within a robust governance, consent, and escalation protocol.

Emergency Medicine, Surgery, Rehabilitation, and Remote Care

Emergency and ICU settings require the rapid interpretation of multimodal streams, including vitals, labs, imaging, ECG, clinical notes, waveforms, and nurse observations. MLLMs can support early warning, triage summarization, and deterioration detection; however, latency and false-alarm consequences are critical. Surgery and procedural care involve videos, instrument tracking, audio instructions, preoperative notes, and operative

reports. Video-language models can support procedural step recognition, surgical education, documentation, and safety monitoring; however, annotated procedural datasets such as Cholec80 remain limited compared with static image datasets [38-40].

Rehabilitation and remote care can benefit from video, sensor, EMG, speech, and patient-reported data. MLLMs can generate progress reports, monitor adherence, and provide tailored exercise feedback. However, home environment variability, privacy, camera positioning, differences in wearable devices, and vulnerable populations, such as children and older adults, require careful validation [38-40].

Evaluation Metrics, Benchmarks, and Clinical Validation

Evaluation is one of the most important weaknesses in the current MLLM literature. Generic NLP metrics such as BLEU, ROUGE, METEOR, and CIDEr may be useful for report-generation comparison but may fail to measure clinical correctness. A clinically dangerous report may be lexically similar to a reference report, whereas a clinically accurate report may use different wording. Therefore, healthcare MLLMs require multi-layer evaluation, including task performance, clinical factuality, grounding, safety, calibration, fairness, latency, and expert assessment [41-46].

Table 7 summarizes the evaluation approaches by modality and task.

Table 7. Evaluation metrics and validation considerations for healthcare MLLMs.

Evaluation area	Example benchmarks	Core metrics	Clinical evaluation need	Limitation
Image-text report generation	MIMIC-CXR, CheXpert, IU X-ray, PadChest, MAIRA-style grounded tasks	BLEU, ROUGE, METEOR, CIDEr, CheXpert label accuracy, RadGraph/RadFact-style factuality, localization accuracy	Radiologist review, clinical error categories, grounding and report completeness	Generic text metrics may miss clinical meaning
Medical VQA	VQA-RAD, PathVQA, SLAKE, biomedical VQA sets	Accuracy, exact match, F1, expert scoring, rationale quality	Open-ended answers require semantic and clinical review	Small datasets and answer ambiguity
Signal-text	MIT-BIH, PhysioNet, PTB ECG, CHB-MIT	Accuracy, sensitivity, specificity, AUROC, F1, calibration, temporal robustness	Artifact sensitivity and clinically impossible signal detection	Noise and temporal misalignment
Audio-text	DAIC-WOZ, ICBHI, speech-transcription datasets	WER, emotion/symptom classification, AUROC, F1, dialogue quality, MOS	Privacy, accent and language fairness	Limited real-world diversity
Video-text	Cholec80, M2CAI, surgical/rehabilitation video datasets	Top-1/Top-5 accuracy, F1, temporal segmentation, latency/FPS	Procedural safety and expert review	Annotation cost and real-time constraints
Clinical deployment	Prospective local validation, silent-mode trials, human-in-the-loop studies	Calibration, uncertainty, alert burden, time saved, clinician acceptance, adverse events	Workflow impact, accountability, auditability	Rarely reported in current literature

Note. Evaluation categories draw on radiology factuality, report labeling, semantic evaluation and clinical artificial intelligence reporting guidance [41-46, 55-62].

Clinical validation should proceed in stages. The first stage is retrospective benchmark testing using standardized datasets. The second stage is external validation across institutions, devices, languages and patient subgroups. The third stage is silent-mode workflow testing, where outputs are generated but not used for clinical decisions. The fourth stage is prospective human-in-the-loop evaluation under controlled clinical conditions. Limited operational deployment should follow only with monitoring, auditing, fail-safe procedures and clear responsibility boundaries [55-62].

Challenges and Limitations

Data Scarcity and Dataset Imbalance

The most important technical barrier is the scarcity of high-quality paired multimodal healthcare data. Many datasets contain only one or two modalities, such as image-report pairs or ECG-label pairs, whereas real clinical reasoning requires several linked streams over time. Data imbalance also affects the performance. Rare diseases, pediatric cases, low-income settings, low-resource languages, darker skin tones in dermatology, unusual imaging devices, and minority populations are often under-represented. This can lead to inequitable performance and unsafe generalizations [52-54].

Hallucination and Clinical Safety

Hallucinations occur when a model produces plausible but unsupported or incorrect clinical content. In healthcare, hallucinations can involve nonexistent lesions, incorrect ECG interpretations, fabricated patient histories, unsupported medication suggestions, or inaccurate guideline statements. Multimodal grounding may reduce some hallucination risks by connecting the output to visual or signal evidence; however, it does not eliminate hallucinations. Clinical safety requires grounding checks, retrieval support, uncertainty estimation, clinician reviews, conservative output design, and post-deployment monitoring.

Privacy, Consent, and Legal Restrictions

Medical images, clinical notes, speech recordings, videos, and physiological signals may contain sensitive personal information. Voice and video are particularly difficult to anonymize because they can identify patients. Legal frameworks such as HIPAA and GDPR restrict data use, transfer, secondary analysis, and model training. Privacy-preserving approaches, such as federated learning, secure enclaves, de-identification, differential privacy, and local deployment, can reduce risk but do not eliminate the need for governance and consent [63]-[65].

Bias, Fairness, and Low-Resource Settings

Healthcare MLLMs may reproduce or amplify bias if the training data are not representative. Bias can arise from demographics, language, institutions, imaging devices, clinical workflows, insurance status, and disease prevalence. Low-resource settings are especially vulnerable because models trained on high-income hospital datasets may not generalize to different disease patterns or equipment types. Fairness evaluation should be reported by subgroup, where possible, and model deployment should include local validation rather than assuming universal transferability [52-54].

Interpretability and Clinician Trust

Clinicians need to understand why a model produces an output, what evidence it uses, and how confident it is. Methods such as Grad-CAM, integrated gradients, SHAP, LIME, attention visualization, evidence highlighting, and structured rationale generation can help, but interpretability must be validated [47]-[51]. A visually persuasive heatmap may not reflect the true causal reasoning. For clinical trust, explanations should be linked to clinically relevant evidence and reviewed by the domain experts.

Computational Cost and Deployment Barriers

Large multimodal models require substantial GPU memory, storage, and inference infrastructure resources. Real-time tasks, such as emergency triage or surgical support, require low latency. Cloud-based deployment raises privacy and reliability concerns, whereas local deployment may be expensive for smaller hospitals. Parameter-efficient adaptation methods such as LoRA and adapters may reduce training cost, but deployment still requires integration with electronic health records systems, audit trails, cybersecurity, user training, monitoring, and regulatory compliance [63].

Regulatory and Ethical Challenges

MLLMs used in healthcare may fall under medical device or clinical decision-support regulations, depending on their intended use. Ethical governance should address accountability, informed consent, explainability, patient

safety, bias mitigation, monitoring, and transparency issues. The World Health Organization emphasizes responsible AI governance in health, including safety, transparency, equity, accountability, and human oversight issues [55-62]. For high-risk clinical tasks, model outputs should remain advisory unless rigorous prospective validation and regulatory clearance support broader usage. Table 8 summarizes the main technical, clinical, ethical, and regulatory limitations of healthcare multimodal large language models and links each limitation to practical mitigation strategies that can support safer development, validation, and deployment.

Table 8. Key limitations of healthcare MLLMs and practical mitigation strategies.

Challenge	Description	Clinical risk	Mitigation strategy
Data scarcity	Insufficient paired multimodal data and weak labels	Limited generalization and poor rare-case performance	Multi-institutional datasets, expert annotation, standardized reporting, federated learning
Hallucination	Plausible but unsupported clinical content	Misdiagnosis, missing findings, unsafe recommendations	Grounding, retrieval, uncertainty, clinician-in-the-loop review, conservative templates
Privacy and consent	Sensitive text, images, audio and video	Legal risk and loss of patient trust	De-identification, consent management, secure infrastructure, local or federated learning
Bias and fairness	Demographic, language, device and institutional imbalance	Unequal performance across groups	Subgroup evaluation, bias audits, representative data, local validation
Interpretability	Opaque reasoning and weak explanation validity	Low clinician trust and difficult accountability	Evidence highlighting, visual grounding, SHAP/LIME/Grad-CAM, expert review
Deployment cost	High compute, latency, integration and maintenance burden	Limited adoption, especially in smaller hospitals	Efficient architectures, staged deployment, monitoring and workflow co-design
Regulation	Unclear liability and approval pathway	Delayed or unsafe clinical use	Transparent intended use, documentation, audit trails, compliance and post-market monitoring

Note: Mitigation strategies are synthesized from interpretability, fairness, deployment, regulatory and governance literature [47-65].

Future Research Directions

Future healthcare MLLM research should move from impressive demonstrations to clinically accountable systems. The first priority is the development of datasets. Future datasets should be multi-institutional, multimodal, longitudinal, and demographically documented, and accompanied by detailed data sheets. Dataset documentation should state how the data were collected, de-identified, labeled, and validated, and for which clinical tasks the data are appropriate.

The second priority is a clinically grounded evaluation. Benchmark studies should report diagnostic accuracy, sensitivity, specificity, AUROC, calibration, factuality, grounding, hallucination rate, subgroup performance, latency, and expert assessments. Generic NLP metrics should not be used alone. Researchers should also report failure modes rather than only the average performance. This is especially important for rare diseases, pediatric populations, non-English languages, and low-resource healthcare settings.

The third priority is the safe architecture and deployment. MLLMs should include uncertainty estimation, constrained generation, retrieval from current clinical guidelines, terminology grounding, human-in-the-loop review, fail-safe mechanisms, and audit logs. Lightweight adaptation strategies, such as LoRA, may support local fine-tuning, whereas federated learning can help hospitals collaborate without centralizing sensitive data [63-65].

The fourth priority is prospective validation. Retrospective benchmarks are necessary but insufficient for this purpose. Before routine clinical use, MLLMs should be assessed using external validation, silent-mode trials, controlled workflow studies, and post-deployment monitoring. The final goal should not be the autonomous replacement of clinicians but rather a reliable, explainable, and equitable decision support system that reduces workload and improves patient safety.

Conclusions

Multimodal language models are an important next step in healthcare AI because they align more closely with the multimodal nature of clinical reasoning. By combining clinical text, medical images, audio, biosignals, video, and structured data, MLLMs can support report generation, visual question answering, clinical summarization, triage, risk prediction, patient communication, and specialist decision support. The evidence is strongest in radiology and image-text tasks, whereas applications involving audio, biosignals, video, longitudinal records, and rare diseases remain less mature.

The field should avoid over-claiming clinical readiness. Current MLLMs show promise, but they remain limited by scarce paired datasets, hallucinations, privacy restrictions, bias, weak interpretability, high computational cost, and limited prospective validation. For safe translation into clinical practice, future studies must focus on standardized evaluation, external validation, clinically grounded metrics, human oversight, privacy-preserving learning, fairness auditing, and careful regulatory governance. Therefore, MLLMs should be understood as emerging decision-support technologies with significant potential, not as autonomous diagnostic systems.

List of Abbreviations: AI - artificial intelligence; AUROC - area under the receiver operating characteristic curve; BLEU - bilingual evaluation understudy; CNN - convolutional neural network; CT - computed tomography; EHR - electronic health record; ECG - electrocardiogram; EEG - electroencephalogram; EMG - electromyography; GDPR - General Data Protection Regulation; HIPAA - Health Insurance Portability and Accountability Act; ICU - intensive care unit; ICD - International Classification of Diseases; LLM - large language model; MLLM - multimodal large language model; MRI - magnetic resonance imaging; OCT - optical coherence tomography; PPG - photoplethysmography; RLHF - reinforcement learning from human feedback; SNOMED CT - Systematized Nomenclature of Medicine Clinical Terms; UMLS - Unified Medical Language System; VLM - vision-language model; VQA - visual question answering.

Author Contributions: S.A. conceptualized the study, defined the research aim, designed the review framework, and supervised the overall development of this manuscript. R.T. contributed to the literature search, screening of relevant studies, organization of the reviewed materials, and drafting of the sections on datasets, applications and evaluation metrics. H.A. participated in the design of the review structure, contributed to the analysis and interpretation of the reviewed literature, and helped draft and critically revise the manuscript for important intellectual content. All authors contributed to manuscript revision, approved the final version of the manuscript, and agree to be accountable for all aspects of the work in accordance with the ICMJE authorship criteria.

Funding: The authors declare that this research received no specific grants from any funding agency in the public, commercial, or not-for-profit sectors.

Ethics Statement: Not applicable. This article is a literature review and does not involve new data collection from human participants or animals.

Data Availability Statement: No new datasets were generated or analyzed in this review. All the cited datasets and model resources are available through the corresponding publications or official repositories.

Acknowledgments: AI-assisted tools were used to support the language editing, structural refinement, and formatting of the manuscript. The authors reviewed, edited, and took full responsibility for the final content, references, interpretations, and conclusions.

Conflict of Interest: The authors declare that they have no financial or personal relationships that could have influenced the work reported in this manuscript.

References

1. Shortliffe EH. Computer-Based Medical Consultations: MYCIN. New York: Elsevier; 1976.
2. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436-444. <https://doi.org/10.1038/nature14539>.
3. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60-88. <https://doi.org/10.1016/j.media.2017.07.005>.
4. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems 2017*, pp. 6000 – 6010.
5. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, pp. 4171-4186. <https://doi.org/10.18653/v1/N19-1423>.
6. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems 2020*;Article No.: 159:1877 - 1901.
7. OpenAI: Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. *arXiv Submitted on 15 Mar 2023 (v1), last revised 4 Mar 2024 (this version, v6)* <https://doi.org/10.48550/arXiv.2303.08774>.
8. Gemini Team: Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, Schalkwyk J, et al. Gemini: a family of highly capable multimodal models. [Submitted on 19 Dec 2023 (v1), last revised 9 May 2025 (this version, v5)] <https://doi.org/10.48550/arXiv.2312.11805>.
9. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. *Proceedings of the International Conference on Machine Learning. PMLR*. 2021;139:8748-8763.
10. Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*. 2022;35:23716-23736.
11. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems 2023*, Article No.: 1516, pp.34892 - 34916.
12. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems 2023*, Article No.: 1240, pp. 28541 - 28564.
13. Moor M, Huang Q, Wu S, Yasunaga M, Dalmia Y, Leskovec J, et al. Med-Flamingo: a multimodal medical few-shot learner. *Proceedings of Machine Learning for Health. PMLR*. 2023;225: 353-367.
14. Tu T, Azizi S, Driess D, Schaeckermann M, Amin M, Chang PC, et al. Towards generalist biomedical AI. *NEJM AI* 2024;1(3). <https://doi.org/10.1056/AIoA2300138>.
15. Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, et al. A generalist vision-language foundation model for diverse biomedical tasks. *Nat Med*. 2024;30:3129-3141. <https://doi.org/10.1038/s41591-024-03185-2>.
16. Zhang S, Xu Y, Usuyama N, Xu H, Bagga J, Tinn R, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. [Submitted on 2 Mar 2023 (v1), last revised 8 Jan 2025 (this version, v3)] <https://doi.org/10.48550/arXiv.2303.00915>.
17. Bannur S, Bouzid K, Castro DC, Schwaighofer A, Thieme A, Bond-Taylor S, et al. MAIRA-2: grounded radiology report generation. [Submitted on 6 Jun 2024 (v1), last revised 20 Sep 2024 (this version, v2)] <https://doi.org/10.48550/arXiv.2406.04449>.
18. Bassi PRAS, Yavuz MC, Hamamci IE, Er S, Chen X, Li W, et al. RadGPT: constructing 3D image-text tumor datasets. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, pp. 23720-23730.
19. Jiang J, Wang Y, Song S, Zhang Y, Meng Z, Lei B, et al. OmniV-Med: scaling medical vision-language model for universal visual understanding. [Submitted on 20 Apr 2025] <https://doi.org/10.48550/arXiv.2504.14692>.

20. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. <https://doi.org/10.1136/bmj.n71>.
21. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Syst Rev*. 2021;10:39. <https://doi.org/10.1186/s13643-020-01542-z>.
22. Baethge C, Goldbeck-Wood S, Mertens S. SANRA-a scale for the quality assessment of narrative review articles. *Res Integr Peer Rev*. 2019;4:5. <https://doi.org/10.1186/s41073-019-0064-8>.
23. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*. 2019;6:317. <https://doi.org/10.1038/s41597-019-0322-0>.
24. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proc AAAI Conf Artif Intell*. 2019;33(1):590-597. <https://doi.org/10.1609/aaai.v33i01.3301590>.
25. Bustos A, Pertusa A, Salinas JM, de la Iglesia-Vaya M. PadChest: a large chest X-ray image dataset with multi-label annotated reports. *Med Image Anal*. 2020;66:101797. <https://doi.org/10.1016/j.media.2020.101797>.
26. Demner-Fushman D, Kohli MD, Rosenman MB, et al. Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc*. 2016;23(2):304-310. <https://doi.org/10.1093/jamia/ocv080>.
27. Pelka O, Koitka S, Rückert J, Nensa F, Friedrich CM. Radiology Objects in COntext (ROCO): a multimodal image dataset. In: Stoyanov D., et al. *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. LABELS CVII STENT 2018. Lecture Notes in Computer Science*, vol 11043. Springer, Cham. https://doi.org/10.1007/978-3-030-01364-6_20
28. He X, Zhang Y, Mou L, Xing E, Xie P. PathVQA: 30000+ questions for medical visual question answering. *arXiv [Submitted on 7 Mar 2020]*. <https://doi.org/10.48550/arXiv.2003.10286>
29. Liu B, Zhan LM, Xu L, Ma L, Yang Y, Wu XM. Slake: A Semantically-Labeled Knowledge-Enhanced Dataset For Medical Visual Question Answering. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 2021, pp. 1650-1654. <https://doi.org/10.1109/ISBI48211.2021.9434010>.
30. Lin W, Zhao Z, Zhang X, Wu C, Zhang Y, Wang Y, et al. PMC-CLIP: contrastive language-image pre-training using biomedical documents. *arXiv [Submitted on 13 Mar 2023]* <https://doi.org/10.48550/arXiv.2303.07240>.
31. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10:1. <https://doi.org/10.1038/s41597-022-01899-x>.
32. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*. 2000;101(23):e215-e220. <https://doi.org/10.1161/01.cir.101.23.e215>.
33. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. *IEEE Eng Med Biol Mag*. 2001;20(3):45-50. <https://doi.org/10.1109/51.932724>.
34. Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. *Sci Data*. 2020;7:154. <https://doi.org/10.1038/s41597-020-0495-6>.
35. Shoeb A. Application of machine learning to epileptic seizure onset detection and treatment. PhD thesis, Massachusetts Institute of Technology; 2009.
36. Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, et al. The distress analysis interview corpus of human and computer interviews. In: *Proceedings of LREC*. 2014, pp. 3123-3128.
37. Rocha BM, Filos D, Mendes L, Serbes G, Ulukaya S, Kahya YP, et al. An open access database for the evaluation of respiratory sound classification algorithms. *Physiol Meas*. 2019;40(3):035001. <https://doi.org/10.1088/1361-6579/ab03ea>.
38. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans Med Imaging*. 2017;36(1):86-97. <https://doi.org/10.1109/TMI.2016.2593957>.

39. Jin A, Yeung S, Jopling J, Krause J, Azagury D, Milstein A, et al. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: IEEE Winter Conference on Applications of Computer Vision. 2018, pp. 691-699. <https://doi.org/10.1109/WACV.2018.00081>.
40. Stauder R, Ostler D, Kranzfelder M, Koller S, Feußner H, Navab N, et al. The TUM LapChole dataset for the M2CAI surgical workflow challenge. arXiv [Submitted on 28 Oct 2016 (v1), last revised 31 Aug 2017 (this version, v2)] <https://doi.org/10.48550/arXiv.1610.09278>.
41. Jain S, Agrawal A, Saporta A, Truong SQH, Duong DN, Bui T, et al. RadGraph: extracting clinical entities and relations from radiology reports. arXiv [Submitted on 28 Jun 2021 (v1), last revised 29 Aug 2021 (this version, v3)] <https://doi.org/10.48550/arXiv.2106.14463>.
42. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020, pp. 1500-1519. <https://doi.org/10.18653/v1/2020.emnlp-main.117>.
43. Miura Y, Zhang Y, Tsai EB, Langlotz CP, Jurafsky D. Improving factual completeness and consistency of image-to-text radiology report generation. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021, pp. 5288-5304. <https://doi.org/10.18653/v1/2021.naacl-main.416>.
44. Delbrouck JB, Chambon P, Chen Z, Varma M, Johnston A, Blankemeier L, et al. RadGraph-XL: a large-scale expert-annotated dataset for entity and relation extraction from radiology reports. Findings of the Association for Computational Linguistics: ACL 2024, pp. 12902–12915. <https://doi.org/10.18653/v1/2024.findings-acl.765>.
45. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. In: International Conference on Learning Representations. 2020.
46. Yu F, Endo M, Krishnan R, Pan I, Tsai A, Reis EP, Fonseca EKUN, Lee H, Shih G. Evaluating progress in automatic chest X-ray radiology report generation. Patterns. 2023;4(9):100802. <https://doi.org/10.1016/j.patter.2023.100802>.
47. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626. <https://doi.org/10.1109/ICCV.2017.74>.
48. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016, pp. 1135 – 1144. <https://doi.org/10.1145/2939672.2939778>.
49. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, pp. 4768 - 4777.
50. Ancona M, Ceolini E, Oztireli C, Gross M. Towards better understanding of gradient-based attribution methods for deep neural networks. International Conference on Learning Representations. 2018. Available from: <https://openreview.net/pdf?id=Sy21R9JAW>.
51. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv [Submitted on 28 Feb 2017 (v1), last revised 2 Mar 2017 (this version, v2)] <https://doi.org/10.48550/arXiv.1702.08608>.
52. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453. <https://doi.org/10.1126/science.aax2342>.
53. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A. 2020;117(23):12592-12594. <https://doi.org/10.1073/pnas.1919012117>.
54. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27:2176-2182. <https://doi.org/10.1038/s41591-021-01595-0>.
55. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:195. <https://doi.org/10.1186/s12916-019-1426-2>.

56. Sendak MP, D'Arcy J, Kashyap S, Gao M, Nichols M, Corey K, et al. A path for translation of machine learning products into healthcare delivery. *EMJ Innov.* 2020;10:19-00172. <https://doi.org/10.33590/emjinnov/19-00172>.
57. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med.* 2020;26:1364-1374. <https://doi.org/10.1038/s41591-020-1034-x>.
58. The DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med.* 2022;28:186-187. <https://doi.org/10.1038/s41591-021-01643-3>.
59. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell.* 2020;2(2):e200029. <https://doi.org/10.1148/ryai.2020200029>.
60. World Health Organization. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva: World Health Organization; 2021. Available from: <https://www.who.int/publications/i/item/9789240029200>.
61. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg: NIST; 2023. Available from: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>
62. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. *Official Journal of the European Union.* 2024. Available from: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>.
63. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. *arXiv [Submitted on 17 Jun 2021 (v1), last revised 16 Oct 2021 (this version, v2)]* <https://doi.org/10.48550/arXiv.2106.09685>.
64. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, et al. Advances and open problems in federated learning. *Found Trends Mach Learn.* 2021;14(1-2):1-210. <https://doi.org/10.1561/22000000083>.
65. Dwork C. Differential privacy. *ICALP'06: Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II.* 2006, pp. 1-12. https://doi.org/10.1007/11787006_1.