*International Journal of Recent Research and Review, Vol. XVIII, Issue 3, September 2025*
ISSN 2277 – 8322

# Large Language Models for Image-Based Disease Diagnosis: A Systematic Review

Nabil Walid Rafi[1], Md. Sazzadul Islam Prottasha[2], Sharmeen Jahan Seema[3], Prithviraj Chowdhury[4]
[1,2,3,4]*Department of Information and Communication Technology,*
*Bangladesh University of Professionals, Dhaka, Bangladesh*
[1] nabilwalidrafi@gmail.com, [2] sazzadulislamprottasha@gmail.com, [3] seema@bup.edu.bd, [4]prithviraj.chowdhury.pc@gmail.com

***Abstract*—The integration of artificial intelligence (AI) has revolutionized medical image diagnostics. While Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have significant image classification contributions, their clinical utility is restricted due to a lack of medical specialization and natural language processing features. Multimodal Large Language Models (MLLMs) address these gaps with agentic capabilities in specializations like Radiology, Dermatology, Pathology, and Ophthalmology. This systematic review focuses on Novel Taxonomy of MLLM architectures, categorized by data fusion paradigm: Early Fusion, Late Fusion, and the highly effective Cognitive Integration method. This taxonomy highlights the efficiency of the Intermediate/Cognitive Integration approach, which is essential for effectively aligning features from separate modalities to support complex tasks like Radiology Report Generation (RRG) and Whole Slide Imaging (WSI) analysis. This paper further examines key barriers to clinical deployment, specifically data heterogeneity and hallucination risks. To transition MLLMs into trustworthy clinical assistants, the review proposes a Roadmap for Future Research. This roadmap recommends core, high-impact tasks, including frameworks for verifying results, implementing efficient architectural scaling, and addressing patient data security through privacy preserving architectures.**

***Keywords*—*Large Language Model (LLM)*; Machine Learning, Artificial Intelligence; Disease Diagnosis; Image Processing; MedGemma;**

## I. INTRODUCTION

The problem of diseased image analysis, including CT, MRI, X-ray, or Whole Slide Images, is an intensive cognitive process involving much more than visual pattern recognition [1]. Since modern imaging technology produces enormous files (gigapixels), human experts can't keep up. The massive workload leads to delays and, eventually, diagnostic errors due to burnout. [2]. The early success of deep learning technology was predominantly in pattern recognition, leading to the development of extremely strong but narrow AI models [3]. The breakthroughs in dermatological imaging tasks, although impressive, consisted of purely visual pattern recognition tasks, excluding the incorporation of clinical data, creating a bottleneck of translation into practical, clinically applicable technology [4]. The lack of explanation, or even the explanation itself, is the root of failure, described by the clinician as required 'justification, rather than confidence level [5], [6]. The recent Transformer models or Large Language Models represent an entirely novel paradigm with advanced capabilities of deep semantic analysis, data retrieval, or generative analytics described in [7].

MedGemma-4b-it is a medically specialized model developed by Google, capable of accurately identifying diseases in Radiology, Dermatology, Digital Pathology, and Ophthalmology. MedGemma-4b-it is capable of being shrunken for even more medically specialized tasks, such as skin cancer, genomic analysis, brain tumor classification, etc. [8]. Extending the application of the model to the processing of visual input, creating MLLMs, provides the models with the possible advantage of recreating the holistic diagnostic process followed by human diagnostic personnel, wherein visual cues are coupled with textual information to synthesize the diagnostic process, leading to the creation of a justified diagnosis [9], [10]. MLLMs promise the possible advancement of AI from the predictive stage to the explanatory stage regarding diagnoses [11]. The combination approach, taking advantage of the inherent language capabilities of LLMs, is rapidly gaining popularity in the entire gamut of the medical fraternity, recognized as the future frontier for augmented intelligence support in the diagnostic process [12], [13], [14].

Despite the fast growth of MLLM literature in the field of medicine, there still lacks a comprehensive

review that critically examines their design, their range of performance, and their inherent dangers [15]. This systematic review fills the gap in the literature by offering the following crucial contributions: Novel Taxonomy, categorizing MLLM architectures according to their paradigm on data fusion techniques (Early, Late, & Cognitive Integration methods, Modality-Specific Synthesis, encapsulating the state-of-the-art implementations in Radiology, Pathology, & Dermatology/Ophthalmology domains [16], [17], [18], Critical Challenge Analysis, thoroughly scrutinizing the central challenges to successful integration, dealing with the challenges of data heterogeneity [30] and the vital risk of LLM hallucinations with respect to computational resources [19], [20], and lastly, roadmap for future research, recommending core, high-impact tasks necessary for the successful translation of MLLMs into trustworthy healthcare assistants.

## II. LITERATURE REVIEW

The application of AI in the clinical context has enhanced the timely prediction and efficient reasoning. The previous LLMs have been extended with multimodal features for multi-disease prediction and then classification works. The diseases are not limited to dermatology, brain tumors, Pneumonia, etc. The model's performance on the disease recognition tasks has been compared with state-of-the-art medical LLM models in the related studies.

### *A. Foundations of Medical AI and Multimodal LLM*

Recent progress in research with large language models (LLMs) within health care has uncovered some remarkable new developments. MedGemma, an AI model, shows great promise in addressing questions about medicine, visual question answering, chest X-ray classifications, and report generation. In particular, model performance appears to outperform previous models, both in general health areas and in more specialized time medicine like pathology and dermatology. Nonetheless, although MedGemma has the potential to show great promise, there are limitations, including the narrow range of health conditions in its training data, which could affect the reliability in limited situations [8]. Although Gemini has promise, it does not come close to the capabilities of Med-PaLM 2 or GPT-4 regarding diagnostic reasoning and visual question answering. For example, Gemini achieved just 61.45% accuracy on the Medical VQA benchmark when GPT-4 achieved a remarkable 88%. Though it may have capabilities in subjects such as biostatistics and cellular biology, it seems to struggle with specialties such as cardiology and dermatology [23]. Like other advanced AIs, GPT-4 is a surprisingly competent machine learning tool with a considerable potential range of uses. But, It can give some believable misinformation and may not always use consistent rationale or statements [24]. While the foundation models for generalist medical AI are intended to represent better generalizability across multiple relevant data sources (i.e., imaging, electronic health records (EHR), laboratory results, etc), they also face real challenges with regard to validation, bias, and privacy [25]. The evolution from early clinical systems like MYCIN to the latest LLMs, for example, MedGemma, shows progress, but there are neither any standardized empirical measures nor any systematic performance evaluations based on real-world patient data across diverse generations of AI [26]. The recent arrival of models such as DeepSeek-R1 highlights the value of cross-model evaluations, as it may perform well with reasoning tasks, yet poorly in report summarization and tumor classification [27]. As previous studies focused less on skin condition detection, multimodal frameworks such as SkinGPT-4 will begin to develop and combine visual and textual data with dermatological diagnoses in zero-shot situations with remarkable accuracy [28].

### *B. Medical Visual Question Answering (VQA) and Domain-Specific Benchmarks*

Over the years, Medical Visual Question Answering (VQA) and domain-specific benchmarks have significantly advanced AI capabilities in medical imaging, especially in radiology and dermatology. The ReXVQA benchmark introduced a large-scale test for generalist chest X-ray understanding, focusing on presence/negation, spatial localization, differential diagnosis, and geometric analysis. MedGemma, the best model in this current investigation, achieved an accuracy of 83.24% when considering how accurately a medical robot can be compared against radiologists, and found the importance of task type niche training, and benchmarking [29]. In contrast, the Expert Knowledge-Aware Image

Difference Graph model proposed a new task for "difference VQA," designed to assess disease progression by comparing current and reference chest X-ray images. Finally, MM-Skin, the first open-access dermatology dataset, developed the SkinVL model, which outperformed baseline models and significantly improved dermatology VQA tasks. The findings from all studies support the promise of studying explicitly domain-focused VQA datasets and models, especially as it relates to skin disease detection and classification tasks.

### *C. Medical Imaging and Diagnostic AI Applications*

The combination of large language models (LLMs) and computer-aided diagnosis (CAD) networks represents a substantial advancement in medical imaging, with significant improvements in diagnostic support through multimodal pretraining. KAD combines a knowledge-enhanced visual-language model with a knowledge graph of medical knowledge to guide pre-training. It ranked highest for zero-shot diagnosis of novel diseases for chest X-ray [30]. BiomedGPT, a generalist vision-language model, integrates diverse biomedical data to enhance diagnostic performance, demonstrating state-of-the-art results in VQA and report generation [31]. New developments in dermatology multimodal models have added to the possibilities of AI for skin disease diagnosis. One notable multimodal example is MM-Skin, which is a large-scale dermatology vision-language dataset that can decode skin disease by merging clinical, dermoscopic, and pathological images into long-form specialized captions. The dataset supports models like SkinVL, which demonstrates improved diagnostic accuracy and generalization for skin disease classification tasks, outperforming existing general-purpose models in dermatology-specific tasks [32]. These specialized models, built on a wealth of dermatology-specific data, are critical in overcoming the challenges of skin disease detection and offer substantial improvements over general medical AI models.

### *D. MedGemini and Gemini-Based Systems in Medicine*

Med-Gemini is a multimodal, long-context Gemini-based family promoting meaningful multimodal dialogue in a clinically relevant context. Med-Gemini for dermatology is fine-tuned from the original Gemini with custom encoders after a first epoch on PAD-UFES-20 for the six-class skin-lesion classification task by leveraging augmentation and addressing imbalances in classes [33]. Building on this, Advancing Multimodal Medical Capabilities of Gemini modifies Gemini into a range of Med-Gemini variants, showing that Med-Gemini-2D is on par for skin lesion classification using only images on PAD-UFES-20 (Weighted-AUC 92.1%, Weighted-F1 71.4%, Accuracy 73.3%) [34]. In summary, both fine-tuning Med-Gemini in a dermatological context on PAD-UFES-20 published with augmentation/class-imbalance handling allows for competitively [34], [35].

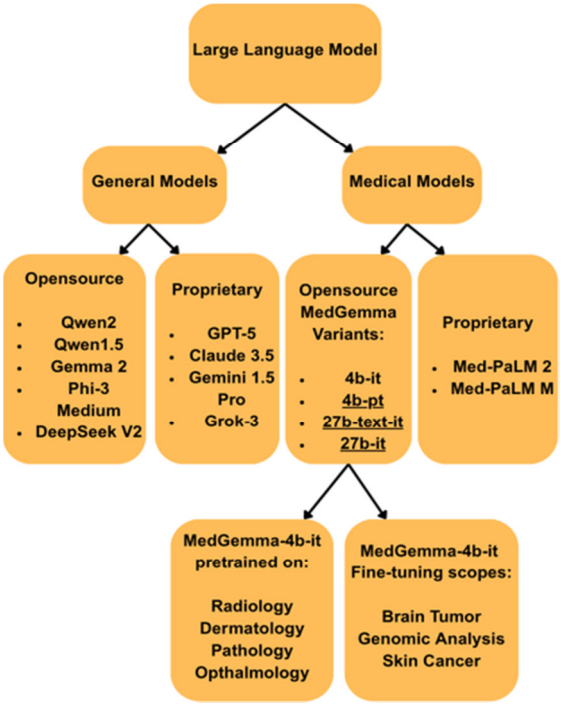


Fig. 1. Generalist and Clinical Large Language Models

Fig. 1 shows the classification of LLMs into generalist and clinically specialized models. Then they are further classified into open-source and proprietary models. The open-source medical model MedGemma-4b-it has been expanded to show its pretrained datasets and the datasets it can be fine-tuned on. Fundamental research works outlined the crucial components of the current state of MLLMs. The works ranged from core transformer structure to medically specialized frameworks, models, datasets, and evaluation techniques.

Table I shows the datasets used in MLLMs in medical specializations like Radiology, Neurology, Dermatology, Ophthalmology, etc., with the number of classes and

images. The current state of MLLMs is defined by several landmark papers that established critical components from the core transformer structure to domain-specific datasets and evaluation paradigms. The datasets in Table I have been extensively used to fine-tune models.

**Table I.** Clinical Datasets used in LLM models

| Dataset | Specialization | Patients / Subjects | Images / Samples | Classes / Labels |
|---|---|---|---|---|
| **MIMIC-CXR** [36] | Radiology | 227,835 | 377,110 | 14 |
| **CheXpert** [37] | Radiology | 65,240 | 224,316 | 14 |
| **NIH Chest-xray14** [38] | Radiology | 30,805 | 112,120 | 14 |
| **Open-I (IU X-Ray)** [39] | Radiology | ~3,955 | 7,470 | ~115 |
| **LIDC-IDRI** [40] | Radiology (CT) | 1,010 | 1,018 | 4 |
| **DeepLesion** [41] | Radiology (CT) | 4,427 | 32,735 | 8 |
| **BraTS** [42] | Neurology (MRI) | 2,040+ | 2,040+ | 4 |
| **ADNI** [43] | Neurology | 1,500+ | 60,000+ | 3 |
| **OASIS-3** [44] | Neurology | 1,098 | 2,168 | 3 |
| **TCGA** [45] | Patho-Genomics | 20,000+ | 30,000+ | 33 |
| **HAM10000** [46] | Dermatology | ~7,000 | 10,015 | 7 |
| **ISIC Challenge** [47] | Dermatology | 15,000+ | 25,000+ | 8 |
| **Pad-UFES-20** [48] | Dermatology | 1,373 | 2,298 | 6 |
| **IDRiD** [49] | Ophthalmology | 516 | 516 | 10 |
| **MIMIC-IV** [50] | Clinical (EHR) | 40,000+ | 0 | ~100+ |
| **MedQA** [51] | Clinical (Text) | 0 | 12,723 | 5 |
| **VQA-RAD** [52] | Radiology (VQA) | ~315 | 315 | 11 |

## III. MLLM ARCHITECTURE

Fusion Strategy explains the architecture of any MLLM and its resulting efficiency in clinical applications. Fusion Strategy is the integration approach of medical images and supporting textual background [9]. This strategy classifies the architectures based on integration point into three classes: Early Fusion (EF), Late Fusion (LF), and Intermediate/Cognitive Integration (IF) [53].

### *A. Early Fusion (EF) Architectures*

Early Fusion maximizes inter-modal interactions by involving the direct concatenation of features at the input level before processing by the primary learning model. Here, raw signals or features extracted by initial encoders are unified into a single, high-dimensional vector, which

is then fed into a comprehensive network. While this mechanism is theoretically optimal for learning the deepest, most complex correlations between modalities, achieving superior accuracy when training data is vast, it faces severe practical constraints. EF is highly susceptible to the curse of dimensionality with high-resolution medical images, leading to intractably large parameter spaces, high computational costs, and a fundamental lack of flexibility when a modality is missing. Exemplary models include early adaptations of BERT to Visual Data (e.g., VisualBERT) [54].

### *B. Late Fusion (LF) Architectures*

The architecture of Late Fusion is fundamentally different. It keeps the modalities separate until the final prediction stage. The independent inputs are processed through specialized networks [55]. The separately trained models give results for averaging or weighting for an aggregated final decision [56]. This design is computationally efficient [57], and is often preferred when training data is limited [58]. However, LF has a lot of data loss during the primary feature extraction. It limits the LLM's ability to accurately analyze the aggregation of the separate set of input. [50].

### *C. Intermediate/Cognitive Integration (IF) Architectures*

Intermediate Fusion, or Cognitive Integration, is the most robust among the 3 fusion types. It is the intersection of deep interaction and architectural flexibility [60]. Modality-specific features are extracted separately, but then processed for merging through a dedicated fusion module. The module can be a cross-attention or query transformer. It is fed into a shared representational space. Then the features are sent to the final generative LLM core [61]. The speciality of IF is its aid in the text generation operations, the central function of an LLM [62], [63]. ClinicalBLIP bridges the semantic gap. The bridging operation is performed through alignment modules like the Query Transformer [64]). The system supports Differential Diagnosis and generates detailed narratives [65], [66]. BLIP utilizes noisy data for visual-language tasks [67]. These models are highly effective, but their multi-stage pipeline is resource-intensive in terms of cost and optimization [58].

Since the models are highly parameterized heavy models, training such heavy billion parameter models is resource obstructive. Withing possible resources, training on these models require specific optimizations. LoRA (Low-Rank Adaptation) fine tunes the high parameter models by inserting low rank matrices into the framework of the model. It freezes the complex model's weights. Then it trains only the new, smaller matrices. Training on these smaller matrices significantly reduces the number of trainable parameters. Due to minimized parameters, the model performance would otherwise degrade. LoRA ensures the performance of the optimized training stays on par with the performance of full fine-tuning [69].

The forward pass given in equation (I):

$$h = W_0 x + \Delta W_x = W_0 x + BA_x \qquad \text{(I)}$$

where $W_0$ and $\Delta W_x$ are the weights of Pretrained and LoRA processed models and A and B are the trainable parameters of LoRA. Instead of training on the high billion parameters of heavy models with their pretrained weight, LoRA trains on its parameters A and B in resource intensive training operations like training MedGemma-4b-it model.

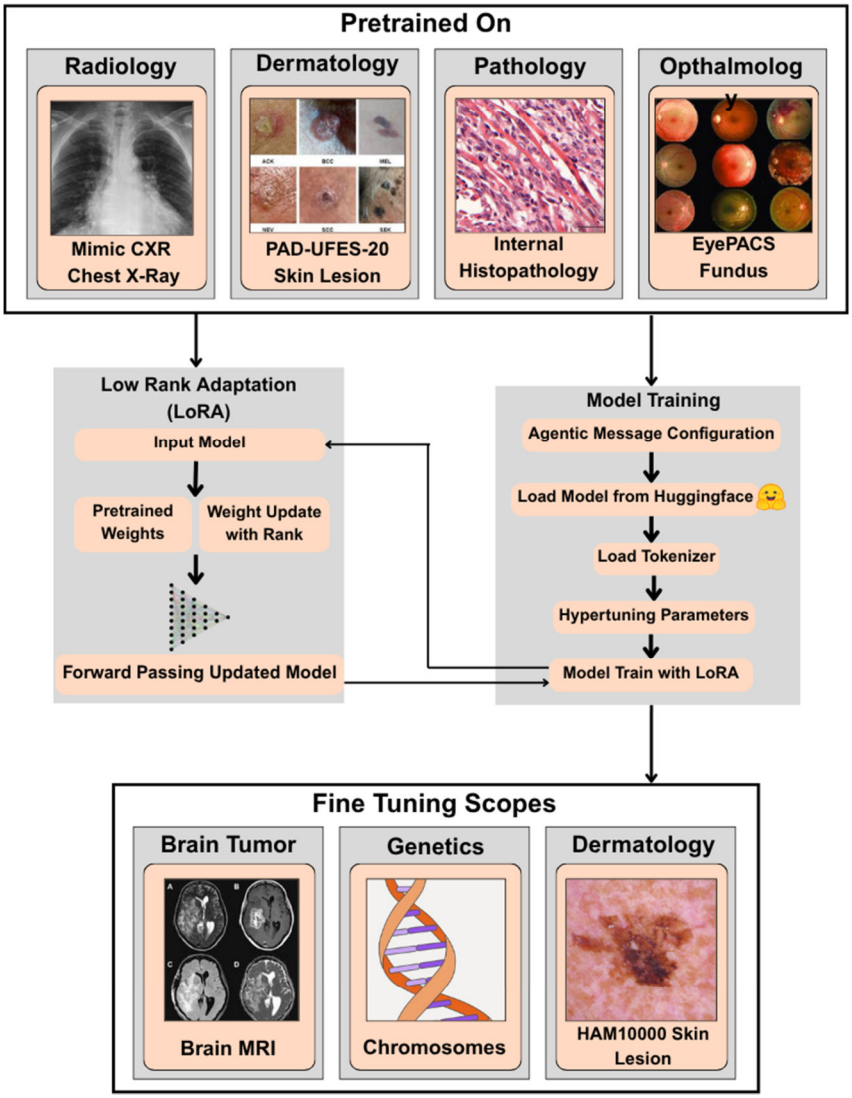


Fig. 2. Fine-tuning Pretrained Clinical Models

Fig. 2 illustrates the training procedure of MLLM's in a possible scale using Low Rank Adaptation (LoRA). The models are pretrained in a number of specializations, whereas they have fine-tuning scopes for more diseases.

## IV. Clinical Applications

The benefit of MLLM is its efficacy in a wide variety of disease diagnoses [70]. Different models are pretrained on different diseases, and there is a wide scope of expansion possible.

### *A. Radiology (CT, MRI, X-ray)*

Radiology includes two-dimensional images like X-ray projections, and three-dimensional CT and MRI volumes [16], [71]. The model analyzes patient history, prior imaging reports, and longitudinal data [36]. Radiology Report Generation (RRG) and Visual Question Answering (VQA) are the MLLM applied to Radiology [71], [72]. Examples of such models are Med-PaLM M [73] and RadFM [74]. 3D MLLMs utilize techniques like spatial pooling perceivers and Masked Image Modeling in recent works [75]. These techniques enable a shift from the current 2D diagnosis to the future 3D diagnosis [76], [77]. *Digital Pathology (Whole Slide Imaging - WSI)* Gigapixel scale of Whole Slide Images (WSIs) is a resource-intensive limitation in the multi-

scale feature analysis works in digital pathology [17], [78]. MLLMs are utilized in tumor grading (e.g., Gleason), integrating visual patterns with molecular markers and genomic data from the EHR [79]. Multi-Instance Learning and Graph Convolutional Networks are advanced feature extraction procedures. They are utilized in accurately modelling complex tissue architecture [80]. WSI-LLaVA is a new framework for bridging the visual-linguistic gap for WSI-level reasoning using diagnosis paths [81], [82].

### *B. Dermatology and Ophthalmology*

Skin and Eye specializations require non-visual contexts like patient demographics or travel history [83], [84]. MLLMs generate Differential Diagnoses (DDx) and detailed image-to-text justification. These background reasons are required for conditions like melanoma and diabetic retinopathy [55], [85]. Frameworks like MICA ensure transparency in skin lesions through explainable concept detection [86]. MLLMs are essential tools for generalized evaluation in these sensitive areas [87].

Table II lists some of the popular MLLMs used in clinical applications with their specializations, base models, parameters, and key training data for fine-tuning.

**Table II.** Clinical MLLMs

| Model | Base Model | Specialization / Task | Parameters | Key Training Data |
|---|---|---|---|---|
| **Med-Gemini** [33] | Gemini | Multimodal reasoning (VQA, text, genomics) | Varies | Medical images, EHRs, text, genomics |
| **Med-PaLM 2** [88] | PaLM 2 | Medical question answering (text) | Varies | Medical domain texts, MedQA dataset |
| **LLaVA-Med** [68] | LLaMA | Multimodal conversational AI (VQA) | 7B | PubMed Central figures & captions |
| **GatorTron** [89] | BERT | Clinical text mining & NLP | 345M - 8.9B | 82B+ words (UF Health clinical notes) |
| **MEDITRON** [90] | Llama 2 | Medical text reasoning & QA | 7B & 70B | PubMed, clinical guidelines, abstracts |
| **ChatDoctor** [91] | LLaMA | Patient-facing conversational AI | 7B | 100k+ real patient-doctor dialogues |
| **BioGPT** [92] | GPT-2 | Biomedical text generation & mining | 1.5B | 15M+ PubMed abstracts & full texts |
| **BioBERT** [93] | BERT | Biomedical text mining (NER, RE, QA) | 110M | PubMed abstracts, PMC full-text articles |
| **ClinicalBERT** [94] | BERT | Clinical note analysis (e.g., readmission) | 110M | MIMIC-III clinical notes |
| **PubMedBERT** [95] | BERT | Biomedical text understanding | 110M | 14M+ PubMed abstracts (from scratch) |

## V. RESEARCH FINDINGS

Large Language Models have been used in clinical prime operations like Clinical Report Narratives, Disease Classification, Question Answering, and Clinical Grounding. The research works have proposed an X-stage Tuning Paradigm containing zero-stage tuning, one-stage tuning, and multi-stage tuning [96], [97]. Fig. 3 shows the 4 medical operations of MLLMs.

Firstly, the models can generate a detailed report of the image provided to them. Secondly, the model can classify the disease of the image, for example, skin lesions, into skin lesion types. Thirdly, the models can answer questions. Finally, the model can localize the disease using a bounding box. These 4 major medical operations of the LLMs allow for efficient clinical applications in real time on a variety of disease prediction, classification, and agentic explanations.

### *5.1 Image-Based Disease Diagnosis Operations*

Clinical MLLMs' wide variety of operations has been classified into four main operations [96].

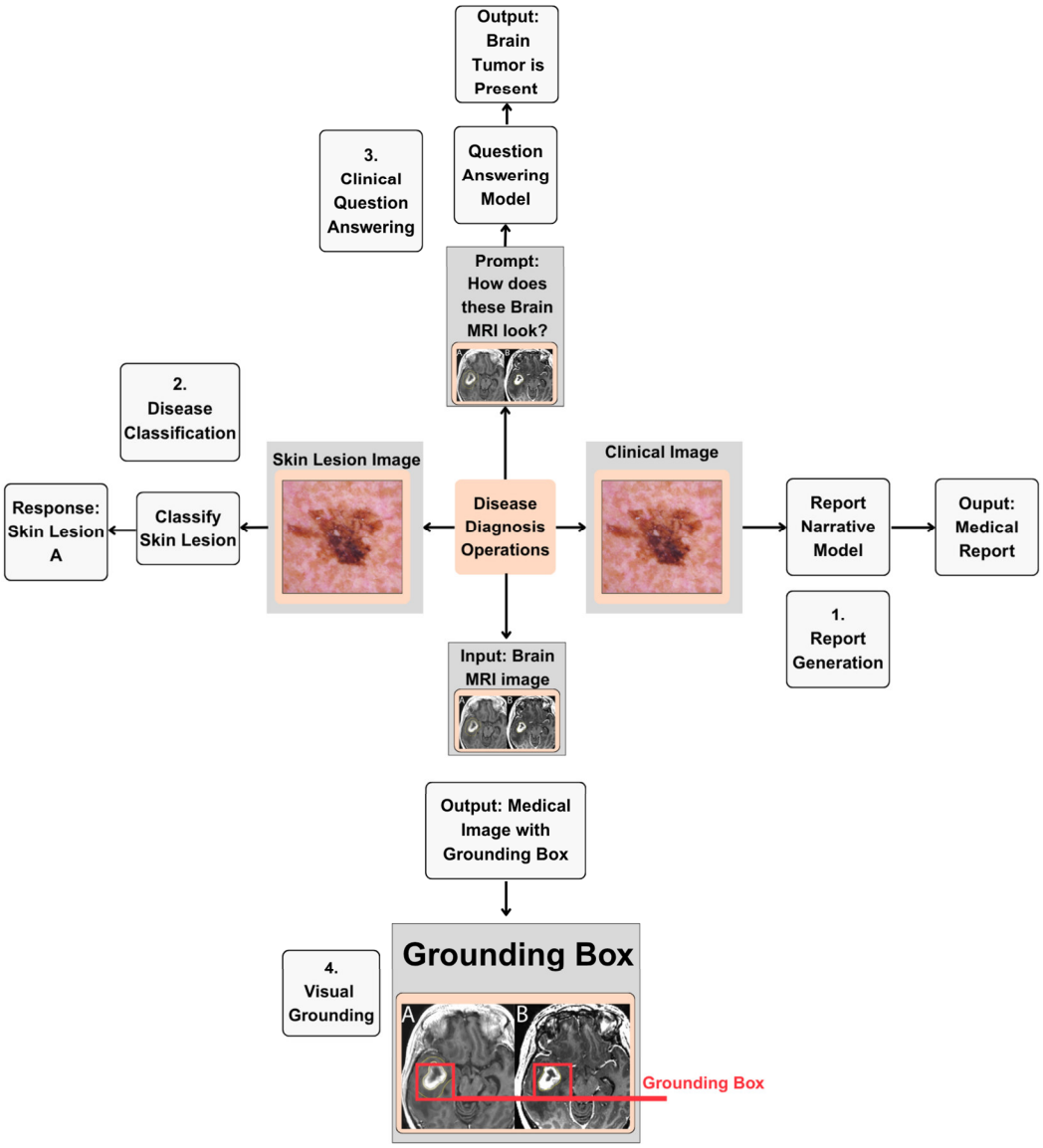


Fig. 3. Clinical Operations of MLLMs

#### 5.1.1 Clinical Report Narratives:

MLLMs are given clinical images of brain MRI, skin lesions, etc., and they give a detailed report on the reasoning for that particular classification. The model also ascertains the logic for not selecting the other classes with a detailed overview. BLEU (Bilingual Evaluation Understudy), ROUGE (Recall Oriented Understudy for Gisting), and CheXpert labels evaluate the generated clinical reports. Reports are evaluated by measuring n-gram overlap between the resulting text and the standard text.

### 5.1.2 Disease Classification:

Classification tasks are focused on attributing a disease image to a specific class of that disease. In Fig. 5, the skin lesion image can be attributed to a class of Skin lesions like melanoma, akiec, nv, etc. In Medical Diagnosis, accuracy and AUC-ROC (Area Under the Receiver Operating Characteristic Curve) values are used for evaluation.

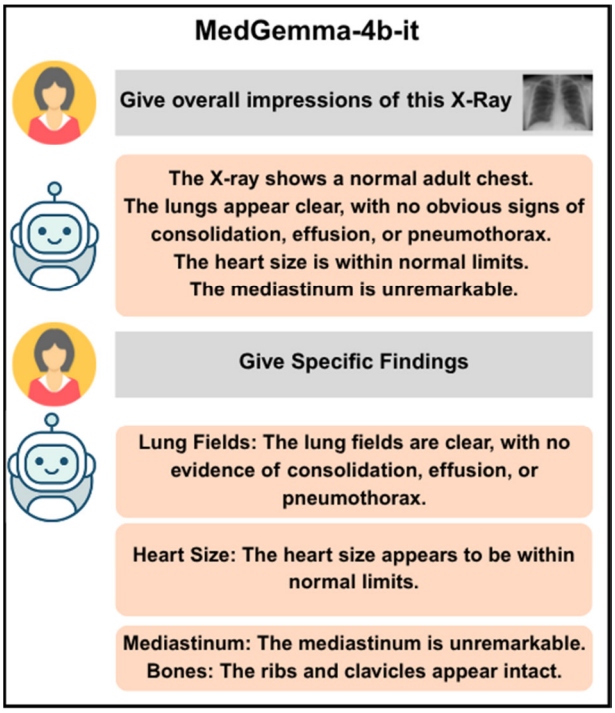


Fig. 4. Med Gemma Question Answering

Fig. 4 shows the question answering capabilities of the open-source model MedGemma, while Fig. 5 shows the comparison of a similar question answer operation on a proprietary model GPT-5. The open-source models can be fine-tuned with datasets of enhanced medical fields.

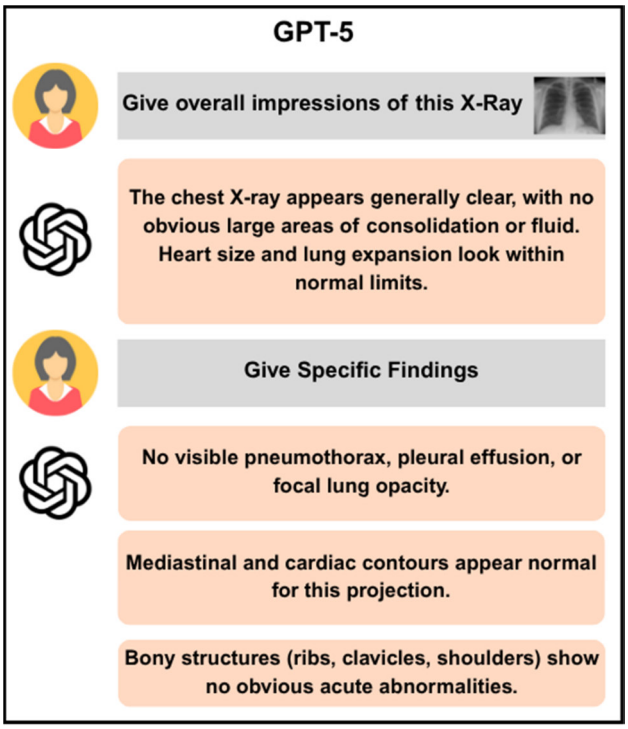


Fig. 5. GPT-5 Question Answering

### 5.1.3 Question Answering:

The MLLM analyzes the textual question and integrates the text and image features for understanding the background and answering correctly. Visual question answering results are evaluated with performance metrics like open and closed-ended accuracy.

### 5.1.4 Clinical Grounding:

Segmentation of regions like tumors in medical images by the MLLMs helps in correct diagnosis. Visual Grounding results in image segmentation tasks are evaluated with IoU (Intersection over Union) and Dice Coefficients.

### *5.2 X-Stage Tuning Paradigm:*

X-stage tuning paradigm utilizes robustness, quick action, and efficiency features of zero-stage, one-stage, and multi-stage fine-tuning; where each of the stage is robust for different use cases. Zero-stage is for standard pretrained model operations while one-stage and multi-stage tuning takes multiple levels of tuning [97].

**5.2.1. Zero Stage Fine-Tuning**: In Zero Stage fine tuning, directly pretrained models like MedGemma-4b-it are used for fast and reliable diagnosis.

$$\gamma = M_{freeze}(X_{img}, X_{Prompt})$$

**5.2.2. One Stage Fine-Tuning:** In one-stage fine-tuning, pretrained models are further fine-tuned on a reliable dataset on pretrained specializations or new fields.

$$\gamma = M_{tune}(X_{img}, X_{Prompt})$$

**5.2.3. Multi-Stage Fine-Tuning:** Multi-stage Fine-Tuning utilizes multiple tuning sets sequentially, ensuring both domain-specific knowledge and generalizability.

$$M_{\text{stage }1}tuning \rightarrow M_{stage\ 2}tuning \rightarrow \cdots \rightarrow M_{stage\ n}tuning$$

$$\gamma = M_{final}(X_{img}, X_{Prompt})$$

The zero-stage fine-tuning is for the quick prediction of diseases. Pretrained models can robustly predict diseases without requiring extensive fine-tuning resources or time. The one-stage tuning fine-tunes the dataset for missing capabilities in a clinical area. The multi-stage fine-tuning approach trains the model on a new clinical field or dataset to enhance its clinical capabilities.

## VI. FUTURE DIRECTIONS

To ensure MLLMs transition ethically and effectively from research prototypes to trustworthy clinical tools, a multi-pronged roadmap is proposed. Future Directions in MLLMs' efficient applications in the clinical context include challenges such as hallucinations, scaling vast data, trustworthiness, security, etc. The areas have been covered with related questions on trustworthiness, architecture scalability, patients' data security, etc.

### *A. How to ensure the trustworthiness of the result?*

The result of an MLLM is reliable if the logical reasoning of its pipeline can be verified [98], [99]. Fact-Grounded Evaluation can be a solution to clinical hallucinations. FAREBIO, a new benchmark, has emerged for biomedical summaries [100]. A Diagnosis Explanation is required to ensure trustworthiness in other instances of hallucinations [83]. Causal discovery (CD) principles can be embedded into MLLM for robust distribution shifts instead of spurious correlations [61]. Furthermore, MLLMs actively signal the clinician toward necessary human intervention during unverified reliance on external knowledge [101].

### *B. Is the architecture efficiently scaled?*

Scalability is required for MLLMs to handle complex data structures [102]. To move MLLMs into prognosis, efficient architectures should be implemented for volumetric 3D data (CT/MRI) and longitudinal tracking of disease progression. These diseases progress across multiple chronological points [103]. Intermediate Fusion strategies can bridge the semantic gap between modalities to utilize even the incomplete data [71].

### *C. Are the patient's data decentralized or secured?*

For global utilization of the MLLMs beyond the local data reserves, the data must be securely processed. Federated Learning allows the data to be securely trained in decentralized local systems. Multiple hospital systems can train the data without requiring sending the data to a global server, an area of data breach [104]. MLLMs can work agentically on their own. To intervene when the MLLM shows unintended operations, a manual deactivation system should be in the pipeline. The MLLM should be integrable with PACs and EHRs [11].

## VII. CONCLUSION

This paper systematically reviews the integration of Large Language Models (LLMs) into medical image diagnosis, illustrating the transformation of artificial intelligence (AI) within the clinical domain. The review emphasizes the evolutionary potential of AI in image-based diagnosis. Earlier Deep Learning (DL) models were limited to unimodal diagnosis, functioning primarily as simple pattern recognizers. Multimodal Large Language Models (MLLMs) represent a significant advance, evolving into cognitive partners capable of integrating both visual features and natural language features to synthesize a robust diagnosis, which is essential for tasks like Radiology Report Generation (RRG) and WSI-level analysis. While MLLMs offer robust capabilities across four major medical operations—Clinical Report Narratives, Disease Classification, Question Answering, and Clinical Grounding yet there exist some critical challenges like LLM hallucination. Minimizing the limitations, MLLMs can complement human expertise and efficiently ensure

evidence-based patient care all around the world. In conclusion, this study provides a blueprint for the next generation of medical AI. Through the fusion of vision and language, MLLMs serve not merely as pattern recognizers but as collaborative partners, essential for complementing clinical judgment and delivering efficient, evidence-based healthcare across diverse populations.

## REFERENCES


[1] X. Li *et al.*, "Vision-Language Models in medical image analysis: From simple fusion to general large models," *Inf. Fusion*, vol. 118, p. 102995, June 2025, doi: 10.1016/j.inffus.2025.102995.

[2] T. Datsi, B. A. Benali, M. Srifi, M. Hachimi, and K. El Kharrachi, "A Short Survey on Multimodal Deep Learning Models in Healthcare," in *Advances in Intelligent Systems and Digital Applications*, vol. 1485, N. Gherabi, J. Kacprzyk, and S. Arezki, Eds., in Lecture Notes in Networks and Systems, vol. 1485. , Cham: Springer Nature Switzerland, 2025, pp. 199–211. doi: 10.1007/978-3-031-95326-2_20.

[3] C. Cui *et al.*, "Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review," *Prog. Biomed. Eng.*, vol. 5, no. 2, p. 022001, Apr. 2023, doi: 10.1088/2516-1091/acc2fe.

[4] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[5] R. AlSaad *et al.*, "Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook," *J. Med. Internet Res.*, vol. 26, p. e59505, Sept. 2024, doi: 10.2196/59505.

[6] T. J. Bradshaw, X. Tie, J. Warner, J. Hu, Q. Li, and X. Li, "Large Language Models and Large Multimodal Models in Medical Imaging: A Primer for Physicians," *J. Nucl. Med.*, vol. 66, no. 2, pp. 173–182, Feb. 2025, doi: 10.2967/jnumed.124.268072.

[7] T. Brown *et al.*, "Language Models are Few-Shot Learners," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, 2020.

[8] A. Sellergren *et al.*, "MedGemma Technical Report," 2025, *arXiv*. doi: 10.48550/ARXIV.2507.05201.

[9] M. Nishino and D. H. Ballard, "Multimodal Large Language Models to Solve Image-based Diagnostic Challenges: The Next Big Wave is Already Here," *Radiology*, vol. 312, no. 1, p. e241379, July 2024, doi: 10.1148/radiol.241379.

[10] Q. Liu *et al.*, "A Review of Applying Large Language Models in Healthcare," *IEEE Access*, vol. 13, pp. 6878–6892, 2025, doi: 10.1109/ACCESS.2024.3524588.

[11] D. Oniani *et al.*, "Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare," *Npj Digit. Med.*, vol. 6, no. 1, p. 225, Dec. 2023, doi: 10.1038/s41746-023-00965-x.

[12] H. Li, J.-F. Fu, and A. Python, "Implementing Large Language Models in Health Care: Clinician-Focused Review With Interactive Guideline," *J. Med. Internet Res.*, vol. 27, p. e71916, July 2025, doi: 10.2196/71916.

[13] S. A. Rabbani *et al.*, "Generative Artificial Intelligence in Healthcare: Applications, Implementation Challenges, and Future Directions," *BioMedInformatics*, vol. 5, no. 3, p. 37, July 2025, doi: 10.3390/biomedinformatics5030037.

[14] J. M. Dolly and N. A.K., "A Survey on Different Multimodal Medical Image Fusion Techniques and Methods," in *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, Chennai, India: IEEE, Apr. 2019, pp. 1–5. doi: 10.1109/ICIICT1.2019.8741445.

[15] J. Cheng, "Applications of Large Language Models in Pathology," *Bioengineering*, vol. 11, no. 4, p. 342, Mar. 2024, doi: 10.3390/bioengineering11040342.

[16] A. Tariq, I. Banerjee, H. Trivedi, and J. Gichoya, "Multimodal artificial intelligence models for radiology," *BJRArtificial Intell.*, vol. 2, no. 1, p. ubae017, Jan. 2025, doi: 10.1093/bjrai/ubae017.

[17] T. Ding *et al.*, "A multimodal whole-slide foundation model for pathology," *Nat. Med.*, Nov. 2025, doi: 10.1038/s41591-025-03982-3.

[18] M. Zarfati *et al.*, "Exploring the Role of Large Language Models in Melanoma: A Systematic Review," *J. Clin. Med.*, vol. 13, no. 23, p. 7480, Dec. 2024, doi: 10.3390/jcm13237480.

[19] R. Cohen, I. Kligvasser, E. Rivlin, and D. Freedman, "Looks Too Good To Be True: An Information-Theoretic Analysis of Hallucinations in Generative Restoration Models," 2024, *arXiv*. doi: 10.48550/ARXIV.2405.16475.

[20] P. Sui, E. Duede, S. Wu, and R. J. So, "Confabulation: The Surprising Value of Large Language Model Hallucinations," 2024, *arXiv*. doi: 10.48550/ARXIV.2406.04175.

[21] B. Allen *et al.*, "A Road Map for Translational Research on Artificial Intelligence in Medical Imaging: From the 2018 National Institutes of Health/RSNA/ACR/The Academy Workshop," *J. Am. Coll. Radiol.*, vol. 16, no. 9, pp. 1179–1189, Sept. 2019, doi: 10.1016/j.jacr.2019.04.014.

[22] R. J. Chen *et al.*, "Algorithmic fairness in artificial intelligence for medicine and healthcare," *Nat. Biomed. Eng.*, vol. 7, no. 6, pp. 719–742, June 2023, doi: 10.1038/s41551-023-01056-8.

[23] A. Pal and M. Sankarasubbu, "Gemini goes to Med School: exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. arXiv (Cornell University). February

2024. doi: 10.48550," *ArXiv Prepr. Arxiv240207023*, 2023.

[24] J. Achiam *et al.*, "Gpt-4 technical report," *ArXiv Prepr. ArXiv230308774*, 2023.

[25] M. Moor *et al.*, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[26] H. R. Saeidnia and M. Nilashi, "From MYCIN to MedGemma: A Historical and Comparative Analysis of Healthcare AI Evolution," *Infosci. Trends*, vol. 2, no. 6, pp. 18–28, 2025.

[27] M. Tordjman *et al.*, "Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning," *Nat. Med.*, pp. 1–1, 2025.

[28] J. Zhou *et al.*, "Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4," *Nat. Commun.*, vol. 15, no. 1, p. 5649, 2024.

[29] A. Pal *et al.*, "ReXVQA: A Large-scale Visual Question Answering Benchmark for Generalist Chest X-ray Understanding," *ArXiv Prepr. ArXiv250604353*, 2025.

[30] X. Zhang, C. Wu, Y. Zhang, W. Xie, and Y. Wang, "Knowledge-enhanced visual-language pre-training on chest radiology images," *Nat. Commun.*, vol. 14, no. 1, p. 4542, 2023.

[31] K. Zhang *et al.*, "A generalist vision–language foundation model for diverse biomedical tasks," *Nat. Med.*, vol. 30, no. 11, pp. 3129–3141, 2024.

[32] W. Zeng, Y. Sun, C. Ma, W. Tan, and B. Yan, "MM-Skin: Enhancing Dermatology Vision-Language Model with an Image-Text Dataset Derived from Textbooks," *ArXiv Prepr. ArXiv250506152*, 2025.

[33] K. Saab *et al.*, "Capabilities of Gemini Models in Medicine," 2024, *arXiv*. doi: 10.48550/ARXIV.2404.18416.

[34] L. Yang *et al.*, "Advancing multimodal medical capabilities of Gemini," *ArXiv Prepr. ArXiv240503162*, 2024.

[35] G. Team *et al.*, "Gemini: a family of highly capable multimodal models," *ArXiv Prepr. ArXiv231211805*, 2023.

[36] A. E. W. Johnson *et al.*, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," 2019, *arXiv*. doi: 10.48550/ARXIV.1901.07042.

[37] J. Irvin *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 590–597, July 2019, doi: 10.1609/aaai.v33i01.3301590.

[38] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, July 2017, pp. 3462–3471. doi: 10.1109/CVPR.2017.369.

[39] D. Demner-Fushman *et al.*, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Am. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, Mar. 2016, doi: 10.1093/jamia/ocv080.

[40] S. G. Armato *et al.*, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans," *Med. Phys.*, vol. 38, no. 2, pp. 915–931, Feb. 2011, doi: 10.1118/1.3528204.

[41] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *J. Med. Imaging*, vol. 5, no. 03, p. 1, July 2018, doi: 10.1117/1.JMI.5.3.036501.

[42] B. H. Menze *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.

[43] R. C. Petersen *et al.*, "Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, Jan. 2010, doi: 10.1212/WNL.0b013e3181cb3e25.

[44] P. J. LaMontagne *et al.*, "OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease," Dec. 15, 2019, *Radiology and Imaging*. doi: 10.1101/2019.12.13.19014902.

[45] The Cancer Genome Atlas Research Network *et al.*, "The Cancer Genome Atlas Pan-Cancer analysis project," *Nat. Genet.*, vol. 45, no. 10, pp. 1113–1120, Oct. 2013, doi: 10.1038/ng.2764.

[46] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, p. 180161, Aug. 2018, doi: 10.1038/sdata.2018.161.

[47] N. C. F. Codella *et al.*, "Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC)," 2017, *arXiv*. doi: 10.48550/ARXIV.1710.05006.

[48] A. G. C. Pacheco *et al.*, "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data Brief*, vol. 32, p. 106221, Oct. 2020, doi: 10.1016/j.dib.2020.106221.

[49] P. Porwal *et al.*, "Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research," *Data*, vol. 3, no. 3, p. 25, July 2018, doi: 10.3390/data3030025.

[50] A. E. W. Johnson *et al.*, "MIMIC-IV, a freely accessible electronic health record dataset," *Sci. Data*, vol. 10, no. 1, p. 1, Jan. 2023, doi: 10.1038/s41597-022-01899-x.

[51] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "PubMedQA: A Dataset for Biomedical Research

Question Answering," 2019, *arXiv*. doi: 10.48550/ARXIV.1909.06146.

[52] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Sci. Data*, vol. 5, no. 1, p. 180251, Nov. 2018, doi: 10.1038/sdata.2018.251.

[53] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Mach. Vis. Appl.*, vol. 32, no. 6, p. 121, Nov. 2021, doi: 10.1007/s00138-021-01249-8.

[54] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A Simple and Performant Baseline for Vision and Language," 2019, *arXiv*. doi: 10.48550/ARXIV.1908.03557.

[55] Y. Li *et al.*, "A review of deep learning-based information fusion techniques for multimodal medical image classification," *Comput. Biol. Med.*, vol. 177, p. 108635, July 2024, doi: 10.1016/j.compbiomed.2024.108635.

[56] B. Huang, F. Yang, M. Yin, X. Mo, and C. Zhong, "A Review of Multimodal Medical Image Fusion Techniques," *Comput. Math. Methods Med.*, vol. 2020, pp. 1–16, Apr. 2020, doi: 10.1155/2020/8279342.

[57] N. Ardic and R. Dinc, "Emerging trends in multi-modal artificial intelligence for clinical decision support: A narrative review," *Health Informatics J.*, vol. 31, no. 3, p. 14604582251366141, July 2025, doi: 10.1177/14604582251366141.

[58] V. Guarrasi *et al.*, "A systematic review of intermediate fusion in multimodal deep learning for biomedical applications," *Image Vis. Comput.*, vol. 158, p. 105509, May 2025, doi: 10.1016/j.imavis.2025.105509.

[59] C. Liu and F. Ye, "A review of multimodal medical data fusion techniques for personalized medicine," in *Proceedings of the 4th International Conference on Biomedical and Intelligent Systems*, Bologna Italy: ACM, Apr. 2025, pp. 338–347. doi: 10.1145/3745034.3745088.

[60] A. Cesario, M. Gorini, and D. D'Amario, "The Future of Digital Medicine," in *Digital Medicine Starter Guide*, Cham: Springer Nature Switzerland, 2025, pp. 109–122. doi: 10.1007/978-3-032-01272-2_7.

[61] X. Song *et al.*, "Cross-modal attention for multi-modal image registration," *Med. Image Anal.*, vol. 82, p. 102612, Nov. 2022, doi: 10.1016/j.media.2022.102612.

[62] Y. Hu, C. Xu, B. Lin, W. Yang, and Y. Y. Tang, "Medical multimodal large language models: A systematic review," *Intell. Oncol.*, vol. 1, no. 4, pp. 308–325, Oct. 2025, doi: 10.1016/j.intonc.2025.09.005.

[63] T. Syeda-Mahmood *et al.*, Eds., *Multimodal Learning for Clinical Decision Support: 11th International Workshop, ML-CDS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings*, vol. 13050. in Lecture Notes in Computer Science, vol. 13050. Cham: Springer International Publishing, 2021. doi: 10.1007/978-3-030-89847-2.

[64] J. Ji, Y. Hou, X. Chen, Y. Pan, and Y. Xiang, "Vision-Language Model for Generating Textual Descriptions From Clinical Images: Model Development and Validation Study," *JMIR Form. Res.*, vol. 8, p. e32690, Feb. 2024, doi: 10.2196/32690.

[65] Q. Chen, X. Hu, Z. Wang, and Y. Hong, "MedBLIP: Bootstrapping Language-Image Pre-training from 3D Medical Images and Texts," 2023, *arXiv*. doi: 10.48550/ARXIV.2305.10799.

[66] E. H. Houssein, A. M. Gamal, E. M. G. Younis, and E. Mohamed, "Explainable artificial intelligence for medical imaging systems using deep learning: a comprehensive review," *Clust. Comput.*, vol. 28, no. 7, p. 469, Sept. 2025, doi: 10.1007/s10586-025-05281-5.

[67] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," 2022, *arXiv*. doi: 10.48550/ARXIV.2201.12086.

[68] C. Li *et al.*, "LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day," 2023, *arXiv*. doi: 10.48550/ARXIV.2306.00890.

[69] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," 2021, *arXiv*. doi: 10.48550/ARXIV.2106.09685.

[70] S. Bhosekar, P. Singh, D. Garg, V. Ravi, and M. Diwakar, "A Review of Deep Learning-based Multi-modal Medical Image Fusion," *Open Bioinforma. J.*, vol. 18, no. 1, p. e18750362370697, July 2025, doi: 10.2174/0118750362370697250630063814.

[71] Y. Bazi, M. M. A. Rahhal, L. Bashmal, and M. Zuair, "Vision–Language Model for Visual Question Answering in Medical Imagery," *Bioengineering*, vol. 10, no. 3, p. 380, Mar. 2023, doi: 10.3390/bioengineering10030380.

[72] M. Hu, J. Qian, S. Pan, Y. Li, R. L. J. Qiu, and X. Yang, "Advancing medical imaging with language models: featuring a spotlight on ChatGPT," *Phys. Med. Biol.*, vol. 69, no. 10, p. 10TR01, May 2024, doi: 10.1088/1361-6560/ad387d.

[73] T. Tu *et al.*, "Towards Generalist Biomedical AI," 2023, *arXiv*. doi: 10.48550/ARXIV.2307.14334.

[74] W. Xie, C. Wu, X. Zhang, Y. Zhang, and Y. Wang, "Towards Generalist Foundation Model for Radiology," Sept. 18, 2023, *In Review*. doi: 10.21203/rs.3.rs-3324530/v1.

[75] J. Zhuang *et al.*, "MiM: Mask in Mask Self-Supervised Pre-Training for 3D Medical Image Analysis," 2024, *arXiv*. doi: 10.48550/ARXIV.2404.15580.

[76] Q. Chen, X. Yao, H. Ye, and Y. Hong, "Enhancing 3D Medical Image Understanding with Pretraining Aided by 2D Multimodal Large Language Models," *IEEE J. Biomed. Health Inform.*, pp. 1–14, 2025, doi: 10.1109/JBHI.2025.3609739.

[77] M. Baharoon, J. Ma, C. Fang, A. Toma, and B. Wang, "Exploring the Design Space of 3D MLLMs for CT

Report Generation," 2025, *arXiv*. doi: 10.48550/ARXIV.2506.21535.

[78] R. Deng *et al.*, "Cross-scale multi-instance learning for pathological image diagnosis," *Med. Image Anal.*, vol. 94, p. 103124, May 2024, doi: 10.1016/j.media.2024.103124.

[79] L. Tong *et al.*, "Integrating Multi-Omics Data With EHR for Precision Medicine Using Advanced Artificial Intelligence," *IEEE Rev. Biomed. Eng.*, vol. 17, pp. 80–97, 2024, doi: 10.1109/RBME.2023.3324264.

[80] M. Afonso, P. M. S. Bhawsar, M. Saha, J. S. Almeida, and A. L. Oliveira, "Multiple Instance Learning for WSI: A comparative analysis of attention-based approaches," *J. Pathol. Inform.*, vol. 15, p. 100403, Dec. 2024, doi: 10.1016/j.jpi.2024.100403.

[81] Y. Liang *et al.*, "WSI-LLaVA: A Multimodal Large Language Model for Whole Slide Image," 2024, *arXiv*. doi: 10.48550/ARXIV.2412.02141.

[82] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, and J. Shi, "Tracing Diagnosis Paths on Histopathology WSIs for Diagnostically Relevant Case Recommendation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, vol. 12265, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds., in Lecture Notes in Computer Science, vol. 12265. , Cham: Springer International Publishing, 2020, pp. 459–469. doi: 10.1007/978-3-030-59722-1_44.

[83] S. Wang *et al.*, "Advances and prospects of multi-modal ophthalmic artificial intelligence based on deep learning: a review," *Eye Vis.*, vol. 11, no. 1, p. 38, Oct. 2024, doi: 10.1186/s40662-024-00405-1.

[84] H. Hashemian *et al.*, "Application of Artificial Intelligence in Ophthalmology: An Updated Comprehensive Review," *J. Ophthalmic Vis. Res.*, vol. 19, no. 3, pp. 354–367, Sept. 2024, doi: 10.18502/jovr.v19i3.15893.

[85] Z. Zhang *et al.*, "Evaluating Large Language Models in Ophthalmology: Systematic Review," *J. Med. Internet Res.*, vol. 27, p. e76947, Oct. 2025, doi: 10.2196/76947.

[86] Y. Bie, L. Luo, and H. Chen, "MICA: Towards Explainable Skin Lesion Diagnosis via Multi-Level Image-Concept Alignment," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 2, pp. 837–845, Mar. 2024, doi: 10.1609/aaai.v38i2.27842.

[87] J. L. Cross, M. A. Choma, and J. A. Onofrey, "Bias in medical AI: Implications for clinical decision-making," *PLOS Digit. Health*, vol. 3, no. 11, p. e0000651, Nov. 2024, doi: 10.1371/journal.pdig.0000651.

[88] K. Singhal *et al.*, "Towards Expert-Level Medical Question Answering with Large Language Models," May 16, 2023, *arXiv*: arXiv:2305.09617. doi: 10.48550/arXiv.2305.09617.

[89] X. Yang *et al.*, "GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records," 2022, *arXiv*. doi: 10.48550/ARXIV.2203.03540.

[90] Z. Chen *et al.*, "MEDITRON-70B: Scaling Medical Pretraining for Large Language Models," 2023, *arXiv*. doi: 10.48550/ARXIV.2311.16079.

[91] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge," 2023, *arXiv*. doi: 10.48550/ARXIV.2303.14070.

[92] R. Luo *et al.*, "BioGPT: generative pre-trained transformer for biomedical text generation and mining," *Brief. Bioinform.*, vol. 23, no. 6, p. bbac409, Nov. 2022, doi: 10.1093/bib/bbac409.

[93] J. Lee *et al.*, "BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1242, 2019.

[94] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," 2019, *arXiv*. doi: 10.48550/ARXIV.1904.05342.

[95] Y. Gu *et al.*, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Trans. Comput. Healthc.*, vol. 3, no. 1, pp. 1–23, Jan. 2022, doi: 10.1145/3458754.

[96] B. Urooj, M. Fayaz, S. Ali, L. M. Dang, and K. W. Kim, "Large Language Models in Medical Image Analysis: A Systematic Survey and Future Directions," *Bioengineering*, vol. 12, no. 8, p. 818, July 2025, doi: 10.3390/bioengineering12080818.

[97] P. Wang, W. Lu, C. Lu, R. Zhou, M. Li, and L. Qin, "Large Language Model for Medical Images: A Survey of Taxonomy, Systematic Review, and Future Trends," *Big Data Min. Anal.*, vol. 8, no. 2, pp. 496–517, Apr. 2025, doi: 10.26599/BDMA.2024.9020090.

[98] K. Matton, R. O. Ness, J. Guttag, and E. Kıcıman, "Walk the Talk? Measuring the Faithfulness of Large Language Model Explanations," 2025, *arXiv*. doi: 10.48550/ARXIV.2504.14150.

[99] C. Agarwal, S. H. Tanneru, and H. Lakkaraju, "Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models," 2024, *arXiv*. doi: 10.48550/ARXIV.2402.04614.

[100] B. Fang, X. Dai, and S. Karimi, "Understanding Faithfulness and Reasoning of Large Language Models on Plain Biomedical Summaries," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 9890–9911. doi: 10.18653/v1/2024.findings-emnlp.578.

[101] Z. Cai *et al.*, "Exploring Compositional Generalization of Multimodal LLMs for Medical Imaging," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria: Association for Computational Linguistics, 2025, pp. 13057–13079. doi: 10.18653/v1/2025.acl-long.639.

[102] F. Bai, Y. Du, T. Huang, M. Q.-H. Meng, and B. Zhao, “M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models,” 2024, *arXiv*. doi: 10.48550/ARXIV.2404.00578.

[103] M. Kim *et al.*, “Interpretable temporal graph neural network for prognostic prediction of Alzheimer’s disease using longitudinal neuroimaging data,” in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Houston, TX, USA: IEEE, Dec. 2021, pp. 1381–1384. doi: 10.1109/BIBM52615.2021.9669504.

[104] M. H. U. Rehman, W. Hugo Lopez Pinaya, P. Nachev, J. T. Teo, S. Ourselin, and M. J. Cardoso, “Federated learning for medical imaging radiology,” *Br. J. Radiol.*, vol. 96, no. 1150, p. 20220890, Oct. 2023, doi: 10.1259/bjr.20220890.