Review

Multimodal Generative AI in Diagnostics: Bridging Medical Imaging and Clinical Reasoning

Morteza Maleki * and SeyedAli Ghahari

Posted Date: 20 August 2025

doi: 10.20944/preprints202508.1425.v1

Keywords: artificial intelligence; generative AI; multimodal AI; medical diagnostics; clinical reasoning

 doi:10.20944/preprints202508.1425.v1

Review

Multimodal Generative AI in Diagnostics: Bridging Medical Imaging and Clinical Reasoning

Morteza Maleki [1,*] **and SeyedAli Ghahari** [2]

[1] Adjunct Researcher, Emory University, Winship Cancer Institute, Georgia, USA
[2] Researcher, Institute for Advanced Construction and Smart Infrastructure Solutions of America
* Correspondence: mmalek3@emory.edu

Abstract

Multimodal generative artificial intelligence (AI) has emerged as a transformative approach in medical diagnostics, integrating diverse data sources to significantly enhance clinical decision-making and patient care. In this review, we systematically analyze recent advancements and methodologies in multimodal generative AI, focusing particularly on the fusion of medical imaging data with clinical records, genomic information, and textual narratives. We evaluate how these combined modalities closely mimic physician cognitive processes, leading to improved diagnostic accuracy and personalized patient management across various specialties including radiology, pathology, dermatology, and ophthalmology. Specifically, we discuss three key integration s trategies: tool-use approaches, where large language models orchestrate specialized diagnostic modules; grafting techniques, which directly incorporate visual analysis into linguistic frameworks; and unified frameworks, providing simultaneous multimodal data processing within cohesive models. Additionally, we highlight exemplary models, such as PathChat, demonstrating substantial accuracy improvements (e.g., 89.5% in pathological image interpretation) resulting from multimodal integration. We also critically assess ongoing challenges, including technical barriers to data integration, interpretability issues affecting clinical trust, privacy and ethical concerns, and the evolving regulatory landscape surrounding AI-driven diagnostics. Finally, we propose directions for future research, emphasizing the need for large-scale clinical validation studies, standardized evaluation frameworks, advances in explainable AI methods, and privacy-preserving techniques such as federated learning. Ultimately, multimodal generative AI holds significant promise to augment rather than replace clinical expertise, serving as a powerful complement to human decision-making in medicine.

Keywords: artificial intelligence; generative AI; multimodal AI; medical diagnostics; clinical reasoning

1. Introduction

1.1. Context and Motivation

Artificial intelligence (AI), particularly generative AI, has emerged as a transformative technology in healthcare, revolutionizing medical diagnostics, clinical decision-making, and patient care [1–5]. Unlike traditional AI, which primarily classifies, predicts, or analyzes data, generative AI models are uniquely capable of creating novel, high-quality synthetic data, images, and textual narratives that closely mimic real-world examples [6–12]. This capability is especially impactful in medical imaging and clinical diagnostics, where multimodal generative models can integrate complex data from multiple sources, including radiological images, pathology slides, electronic health records, and clinical notes, to produce comprehensive insights and enhanced diagnostic precision. Such multimodal approaches not only assist clinicians in interpreting intricate medical data but also facilitate personalized treatment planning, improve patient-provider communication, and enhance training and educational resources. As generative AI technologies evolve, their application in diagnostics is poised to bridge critical gaps between medical imaging and clinical reasoning, fostering deeper

 doi:10.20944/preprints202508.1425.v1

understanding of diseases and enabling healthcare providers to make more informed, accurate, and patient-centered clinical decisions.

The integration of image and text data in multimodal generative AI systems is becoming increasingly important for improving diagnostic accuracy and clinical decision-making in medicine. This approach allows for a more comprehensive analysis by combining visual information from medical imaging with contextual data from clinical reports and other text sources, creating synergies that address limitations inherent in single-modality approaches. Multimodal AI models that fuse imaging and text data have demonstrated superior performance compared to unimodal approaches in many diagnostic tasks [13,14]. For example, studies have shown that integrating radiological images with corresponding text reports can enhance the detection and classification of diseases like breast cancer [15]. The complementary nature of these data types allows AI systems to capture both visual features and semantic context, leading to more robust and accurate diagnostics. This integration mirrors the clinical reasoning process, where physicians naturally synthesize visual findings with patient history and clinical context. One key advantage of integrating image and text data is the ability to provide more interpretable and explainable AI outputs. While image-only models may struggle to communicate their reasoning, multimodal systems can generate textual descriptions or annotations to accompany visual findings [16]. This improves transparency and helps clinicians better understand and trust AI-assisted diagnoses. The textual component serves as a natural interface for explaining the model's decision-making process, bridging the gap between complex algorithmic outputs and clinical applicability.

Techniques like visual question answering (VQA) leverage the synergy between images and text, allowing models to answer clinically relevant queries about medical images [17]. This capability can augment radiologists' analyses and potentially speed up reporting workflows. Similarly, automated report generation systems that combine image analysis with natural language processing show promise for assisting in clinical documentation [18]. These interactive systems enable more intuitive human-AI collaboration by allowing clinicians to engage with medical images through natural language inquiries, potentially uncovering subtleties that might be missed in traditional workflows. The fusion of imaging and text data also enables more comprehensive patient profiling for precision medicine approaches. By integrating information from radiomics, genomics, pathology reports, and clinical records, AI systems can generate holistic disease profiles to guide personalized diagnosis and treatment planning [19]. This multi-dimensional approach to patient data allows for the identification of complex patterns and relationships that might not be apparent when examining each data type in isolation, potentially revealing novel biomarkers or disease subtypes with clinical significance. However, several challenges remain in effectively integrating multimodal data. These include dealing with missing or incomplete data across modalities, aligning information from heterogeneous data types, and developing models that can handle the high dimensionality and complexity of combined image-text inputs [20]. Ongoing research is exploring advanced deep learning architectures like transformers to better model cross-modal interactions [21]. Additionally, the development of standardized frameworks for multimodal data preprocessing and harmonization is critical for ensuring the robustness and generalizability of these systems across different clinical settings and patient populations.

As multimodal AI continues to advance, it has the potential to significantly enhance diagnostic capabilities across medical specialties. The synergistic use of imaging and text data allows for a more nuanced and contextual understanding of patient cases. However, careful consideration must be given to issues of data privacy, model interpretability, and clinical validation as these systems are developed for real-world medical applications [18]. Ensuring appropriate governance structures and ethical frameworks for multimodal AI systems is particularly important given the sensitive nature of the combined data types and their potential impact on clinical decision-making. In conclusion, the integration of image and text data in diagnostic AI represents a powerful approach for improving accuracy, interpretability, and clinical utility. As algorithms and datasets continue to evolve, multimodal systems that effectively combine visual and textual information are likely to play an increasingly important role

 doi:10.20944/preprints202508.1425.v1

in supporting clinical decision-making and advancing precision medicine. The future development of these systems will require continued collaboration between AI researchers, medical specialists, and healthcare stakeholders to ensure they address genuine clinical needs while maintaining the highest standards of safety and efficacy.

1.2. Audience and Scope

Multimodal generative AI is emerging as a transformative force in medical diagnostics, offering tremendous potential to enhance clinical decision-making, improve diagnostic accuracy, and advance personalized medicine. This review synthesizes recent developments in multimodal generative AI applied to diagnostics, with a focus on applications for clinical AI researchers and specialists across domains. Multimodal generative AI models are designed to process and integrate diverse types of clinical data, including medical imaging, genomics, electronic health records, and clinical notes. By synthesizing information across modalities, these models can provide more comprehensive analyses compared to unimodal approaches. For example, in neurodegenerative disease diagnosis, multimodal models combining MRI, PET, and clinical data have demonstrated superior performance in tasks like predicting Alzheimer's disease progression [19,22]. This integration capability addresses a fundamental limitation in traditional diagnostic approaches that often rely on single data streams, potentially missing crucial correlations between different clinical indicators.

A key capability of multimodal generative AI is the ability to perform modality translation and data augmentation. This includes generating synthetic data to expand limited datasets, such as creating realistic histopathology images from clinical narratives [14]. Modality conversion techniques also enable transforming data between imaging modalities while preserving anatomical structures, which could reduce the need for invasive biopsies in some cases. These techniques are particularly valuable in medical contexts where data collection is often constrained by ethical considerations, patient burden, and resource limitations. In diagnostics, multimodal generative AI is being applied to enhance several key areas. By integrating complementary information across modalities, these models can improve diagnostic precision. For instance, combining radiological imaging with genomic and clinical data has shown promise in enhancing early detection and classification of cancers [15]. Multimodal models can also synthesize patient-specific data to generate more nuanced and individualized risk predictions for disease onset and progression [13]. Furthermore, by simulating personalized treatment trajectories, generative models can assist in optimizing individualized therapeutic strategies [23]. In clinical settings, multimodal AI systems provide interpretable insights to augment decision-making, integrating imaging findings with relevant patient history and lab results [14]. Beyond direct patient care, these technologies enable creation of realistic synthetic case studies and virtual patient simulations for medical education and skill development [23], addressing the need for comprehensive training resources while avoiding privacy concerns associated with real patient data.

Despite the promise, several challenges remain in implementing multimodal generative AI in clinical practice. Data integration and standardization across modalities and institutions presents significant technical hurdles, as medical data often exists in siloed systems with varying formats and protocols. Preserving patient privacy when working with multi-source data requires sophisticated approaches to ensure compliance with regulatory frameworks while maintaining data utility. Ensuring model interpretability and transparency is critical for clinical adoption, as healthcare providers need to understand the reasoning behind AI-generated recommendations. Validating model performance and reliability in real-world clinical settings remains an ongoing challenge, requiring rigorous testing across diverse patient populations. Additionally, addressing potential biases in training data and model outputs is essential to ensure equitable care delivery across demographic groups [15]. As the field advances, researchers are exploring novel architectures like vision-language models to enable more seamless integration of imaging and text data [14]. These approaches allow models to understand the relationships between visual features in medical images and corresponding clinical descriptions, facilitating more nuanced diagnostic interpretations. There is also growing interest in federated learning approaches to enable multi-institutional collaboration while protecting data privacy [15]. This

 doi:10.20944/preprints202508.1425.v1

distributed learning paradigm allows models to be trained across multiple healthcare institutions without sharing raw patient data, addressing both privacy concerns and the need for diverse training datasets. In conclusion, multimodal generative AI represents a powerful new paradigm in medical diagnostics with the potential to significantly enhance clinical care across specialties. However, realizing this potential will require continued interdisciplinary collaboration between AI researchers, clinicians, and domain experts to develop robust, interpretable, and clinically-validated systems. As these technologies mature, they have the potential to augment clinical expertise, leading to more precise, personalized, and effective diagnostic capabilities. The integration of diverse data streams through generative AI approaches may ultimately transform how diagnoses are made, treatments are planned, and patient outcomes are optimized across the healthcare continuum.

1.3. Clinical Reasoning and Multimodal Integration

Multimodal generative AI is revolutionizing diagnostic approaches in medicine by enabling the integration of diverse data types and aligning with physician workflows. This technology facilitates more comprehensive and accurate diagnoses by synthesizing information from multiple sources, including medical imaging, genomics, clinical records, and patient histories. The integration of vision-language reasoning systems into clinical workflows represents a significant advancement in diagnostic capabilities. For example, PathChat, a multimodal AI pathology assistant, demonstrates remarkable proficiency in interpreting pathological images while engaging in natural language interactions [24]. This system combines a specially trained pathological image visual encoder with a large language model, enabling it to answer diagnostic questions conversationally. In multiple-choice diagnostic tasks with pathological images, PathChat achieved an accuracy of 78.1%, significantly outperforming other models. When both images and clinical background information were provided, its accuracy increased to 89.5%, highlighting the synergistic value of combining multiple data modalities in clinical reasoning [12].

These multimodal systems are designed to mimic the cognitive steps in clinical reasoning, including data gathering, hypothesis generation, and diagnostic verification. In radiology, AI models are being developed to integrate information from various imaging modalities (e.g., X-ray, CT, MRI) with clinical data and patient histories [14]. This approach allows for a more nuanced analysis, capturing complex relationships between different data types and extracting clinically relevant features that may not be apparent when examining each modality in isolation. The integration of multiple imaging techniques mirrors the practice of radiologists who routinely synthesize findings across different scan types to formulate comprehensive assessments. The clinical reasoning process is further enhanced by the ability of these systems to engage in interactive dialogues with healthcare professionals. This allows for clarification of doubts, obtaining additional information, and refining diagnoses through multiple rounds of questioning [12]. Such capabilities are particularly valuable in complex cases, such as diagnosing tumors of unknown primary origin, where multiple rounds of testing and analysis may be required. The conversational interface creates a collaborative diagnostic environment that more closely resembles clinical consultations between specialists rather than traditional algorithmic outputs.

Evaluation metrics for these multimodal AI systems in clinical settings are evolving to reflect their real-world performance. Studies have compared AI performance against human benchmarks, with some models demonstrating comparable or superior accuracy in specific tasks [25]. For example, in a study evaluating multimodal large language models (LLMs) in radiology, human accuracy was used as a benchmark (55.2%), and models like Claude 3.5 Sonnet achieved an overall accuracy of 66.2% when combining imaging data with descriptive text [26]. These promising results suggest that multimodal approaches may overcome limitations inherent to single-modality diagnostic systems by leveraging complementary information sources. However, the integration of these AI systems into clinical practice faces several challenges. One significant hurdle is the need for comprehensive, high-quality multimodal datasets that accurately represent diverse patient populations and clinical scenarios [15]. The scarcity of such datasets has led to exploring synthetic data generation techniques to augment existing datasets and provide diversity for training robust machine learning models [27].

 doi:10.20944/preprints202508.1425.v1

Creating synthetic data that maintains clinical relevance while addressing privacy concerns represents an important frontier in developing generalizable multimodal systems.

Another critical challenge is ensuring that AI models can generalize across diverse clinical environments and patient populations. Techniques such as domain adaptation are being explored to enhance the transferability of AI algorithms across different healthcare settings [16]. Additionally, there is a growing emphasis on developing explainable AI (XAI) techniques to provide transparency in the decision-making processes of these models. Methods such as Grad-CAM, SHAP, and LIME are being adapted to multimodal scenarios, allowing clinicians to understand how different data sources contribute to diagnostic conclusions [14]. This transparency is essential for building trust and facilitating adoption in clinical settings where interpretability is often as important as raw performance metrics. As the field progresses, there is a focus on creating AI systems that augment and complement human expertise rather than replace it. This involves designing interfaces and workflows that facilitate seamless collaboration between clinicians and AI tools. Some systems allow for multiple rounds of dialogue, enabling pathologists to refine their assessments and leverage AI insights effectively [24]. This human-AI collaborative approach acknowledges the unique strengths of both parties—machines excel at pattern recognition across vast datasets, while clinicians contribute contextual understanding and judgment informed by years of experience and domain knowledge.

Looking ahead, the integration of even more diverse data types, such as time-series data from continuous monitoring devices, holds promise for creating even more comprehensive diagnostic models [15]. Longitudinal data collection enables the tracking of disease progression and treatment response over time, potentially allowing for earlier intervention and more personalized therapeutic approaches. There is also potential for these multimodal systems to support not just diagnosis but also treatment planning and monitoring of disease progression, creating a continuum of AI assistance throughout the patient care journey. In conclusion, multimodal generative AI represents a significant leap forward in diagnostic capabilities, offering a more holistic and nuanced approach to patient assessment. By aligning closely with physician workflows and leveraging diverse data sources, these technologies have the potential to enhance clinical decision-making and ultimately improve patient outcomes. However, continued research, validation, and careful consideration of ethical implications will be crucial as these systems become more integrated into clinical practice. The future of diagnostic medicine likely lies in this synergistic relationship between human expertise and AI capabilities, with multimodal systems serving as powerful tools that expand rather than replace the clinician's diagnostic repertoire.

2. Emerging Multimodal Medical AI Models

2.1. Overview of Major Models

Multimodal generative AI models are emerging as powerful tools with potential applications in medical diagnostics, including radiology and pathology. This review examines several prominent models in this domain, discusses their real-world implementation, comparative performance, integration with existing workflows, and ethical implications. Med-PaLM M, LLaVA-Med, BiomedGPT, and BioGPT-ViT represent significant advancements in multimodal AI for medical applications. These models can process both text and medical imaging data, allowing for integrated analysis of radiology scans, pathology slides, and clinical notes [14]. Early results suggest these models may assist with tasks like generating radiology reports, answering clinical questions about images, and providing decision support. The ability to simultaneously process multiple data modalities represents a paradigm shift from traditional unimodal approaches that have dominated medical AI development. Real-world clinical implementation of these models is still in early stages, but some promising examples are emerging. For instance, multimodal AI systems have been used to enhance breast cancer detection by combining imaging data with clinical information. In one study, a hybrid deep learning approach combining image and clinical data achieved 90.6% accuracy, outperforming unimodal methods that relied on either imaging (83.6% accuracy) or clinical data (81.5% accuracy) alone [28]. This demonstrates the

 doi:10.20944/preprints202508.1425.v1

potential for multimodal approaches to improve diagnostic accuracy in real clinical settings. The integration of diverse data streams enables these systems to capture complex relationships between visual findings and patient characteristics that might be missed by traditional diagnostic approaches.

Comparative performance metrics between these models are limited, as they are often evaluated on different datasets and tasks. However, some general trends are emerging. Multimodal models consistently outperform unimodal approaches across various diagnostic tasks [29]. For example, in Alzheimer's disease diagnosis, a multimodal transformer network combining MRI and PET imaging data achieved superior performance compared to state-of-the-art unimodal methods on multiple datasets [19]. These performance improvements highlight the complementary nature of different data modalities in capturing the multifaceted aspects of disease presentation and progression. Integration with existing diagnostic workflows remains a significant challenge. Successful implementation requires careful consideration of operational context and existing clinical processes [30]. Some healthcare systems are developing assessment frameworks to evaluate the value of AI deployment and prioritize which algorithms to implement first [30]. Key considerations include the ability to seamlessly integrate with existing electronic health record systems, provide interpretable outputs for clinicians, and adapt to varying levels of data availability across different healthcare settings. The heterogeneity of healthcare IT infrastructure across institutions further complicates large-scale deployment of these sophisticated AI systems.

Point-of-care testing applications represent a promising area for multimodal AI implementation, particularly in resource-limited settings. Mobile-based colorimetric biosensors combined with AI algorithms can enable low-cost, scalable biomarker quantification for early disease detection and real-time health monitoring [31]. When integrated with wearable sensor data and clinical records, these AI-driven approaches can enhance diagnostic robustness and reliability in settings with limited access to centralized laboratory infrastructure [31]. This democratization of advanced diagnostic capabilities could help address healthcare disparities in underserved regions, though careful attention must be paid to contextual factors that might affect model performance across diverse populations. Ethical implications of implementing these models extend beyond bias and privacy concerns. Patient autonomy and informed consent are critical considerations when AI assists in diagnosis. There is a need to develop clear guidelines for communicating the role of AI in clinical decision-making to patients [32]. Additionally, the potential for AI to exacerbate existing healthcare disparities must be carefully monitored and addressed through inclusive development practices and equitable deployment strategies [13]. As these technologies become more integrated into clinical practice, transparency regarding their limitations and the extent of human oversight becomes increasingly important for maintaining patient trust.

To address these challenges, interdisciplinary collaboration between clinicians, AI researchers, and ethicists is essential. Ongoing research should focus on developing standardized evaluation frameworks to enable fair comparisons between multimodal AI models across diverse clinical tasks and datasets. Creating robust integration strategies that allow AI systems to augment rather than replace existing clinical workflows is crucial, with emphasis on improving efficiency and reducing administrative burden [33]. Investigating the impact of AI-assisted diagnostics on patient-clinician relationships and developing best practices for maintaining trust and shared decision-making remains a priority [22]. Exploring federated learning and other privacy-preserving techniques enables collaborative model development while protecting sensitive patient data [29]. Finally, conducting large-scale, prospective clinical trials is necessary to evaluate the real-world impact of multimodal AI on patient outcomes, clinical efficiency, and healthcare costs.

The transition from research prototypes to clinically validated tools requires careful validation across diverse patient populations. Models trained primarily on data from academic medical centers may not generalize well to community hospitals or rural clinics where patient demographics, equipment specifications, and clinical expertise differ substantially. Furthermore, continuous monitoring and updating of deployed models is essential to maintain performance over time as clinical practices

 doi:10.20944/preprints202508.1425.v1

evolve and new evidence emerges [34,35]. In conclusion, while multimodal generative AI shows tremendous potential to transform medical diagnostics, realizing this potential requires careful navigation of technical, clinical, and ethical challenges. Ongoing research and development efforts, coupled with thoughtful implementation strategies, will be crucial to ensuring these powerful tools enhance rather than disrupt the practice of medicine. The promise of improved diagnostic accuracy, increased efficiency, and more personalized care must be balanced against concerns regarding algorithmic transparency, data privacy, and the preservation of human judgment in clinical decision-making.

2.2. Approaches: Tool Use, Grafting, and Unified Systems

Multimodal generative AI has emerged as a promising approach in medical diagnostics, offering the potential to integrate diverse data types and enhance clinical decision-making. This review explores the design strategies and applications of multimodal large language models (LLMs) in diagnostics, focusing on three main approaches: tool use, grafting, and unification. Multimodal LLMs are designed to process and analyze various data modalities, including medical imaging, electronic health records, genomic information, and real-time patient data. The integration of these diverse data sources allows for a more comprehensive and nuanced analysis of patient conditions, potentially leading to improved diagnostic accuracy and personalized treatment plans [21]. The ability to simultaneously interpret multiple data streams represents a significant advancement over traditional single-modality approaches, mirroring the way clinicians synthesize different types of patient information in practice. The tool use approach involves utilizing LLMs as coordinators that can invoke specialized tools or models for specific tasks. For instance, in radiology, an LLM might analyze a patient's clinical history and symptoms, then call upon a specialized image analysis model to interpret radiographic images. This approach allows for flexibility and modularity, enabling the system to leverage the strengths of different AI models for various diagnostic tasks [36]. By functioning as an orchestrator of specialized models, the LLM can maintain expertise across domains without requiring extensive retraining for each new capability or data type.

Grafting, on the other hand, involves integrating visual processing capabilities directly into language models. This approach has been demonstrated in models like PathChat, which combines a specially trained pathological image visual encoder with a pre-trained large language model [24]. By fine-tuning on diverse image-text instructions, PathChat achieved superior performance in pathological image analysis and diagnostic question-answering tasks compared to other multimodal models. This grafting approach enables more seamless integration of visual and textual information, allowing for interactive diagnostic assistance and multi-round reasoning in complex cases. The direct incorporation of visual processing within the language model architecture creates a more tightly coupled system capable of sophisticated cross-modal reasoning. The unification strategy aims to create a single, cohesive model capable of processing multiple modalities simultaneously. This approach is exemplified by models like the multi-level guided generative adversarial network (MLG-GAN) combined with a multimodal transformer (Mul-T) for incomplete image generation and disease classification [19]. The MLG-GAN generates missing data guided by multi-level information from voxels, features, and tasks, while the Mul-T network models latent interactions and correlations across modalities using a cross-modal attention mechanism. This unified approach has shown promise in improving diagnostic accuracy for conditions such as Alzheimer's disease by leveraging complementary information from different imaging modalities. Rather than treating each data source independently, unification approaches acknowledge and exploit the inherent correlations between different clinical data types.

Each of these design strategies offers unique advantages and challenges. Tool use provides flexibility and modularity but may require more complex orchestration of multiple models. Grafting enables tighter integration of visual and textual processing but may be limited by the capacity of the base language model. Unification offers the potential for more holistic analysis across modalities but can be computationally intensive and challenging to train effectively. The selection of an appropriate strategy depends on the specific clinical context, available computational resources, and the nature of the diagnostic task at hand. The application of multimodal generative AI in diagnostics extends

 doi:10.20944/preprints202508.1425.v1

beyond image analysis. These models are being used to generate synthetic data for research and training purposes, addressing challenges related to data privacy and scarcity [13]. In drug discovery, generative AI models are being employed to generate novel small molecules and predict drug efficacy and safety, potentially accelerating the drug development process [13]. The versatility of these systems allows them to address multiple pain points in the healthcare ecosystem, from improving diagnostic accuracy to expediting therapeutic development.

Despite the promising advancements, several challenges remain in the development and implementation of multimodal generative AI in clinical diagnostics. Data integration and standardization across diverse sources and institutions pose significant hurdles [15]. Ensuring the interpretability and transparency of AI decision-making processes is crucial for building trust among clinicians and patients. Additionally, the need for continuous learning and updating of AI models to keep pace with evolving medical knowledge and diagnostic standards presents ongoing challenges [24]. Regulatory frameworks must also evolve to appropriately evaluate and monitor these complex systems that integrate multiple data types and potentially autonomous decision-making capabilities. As the field progresses, future research directions may include the development of more sophisticated multimodal fusion techniques, improved methods for handling missing or incomplete data across modalities, and the integration of temporal data such as time series information from ECG or EEG alongside imaging modalities [15]. Furthermore, large-scale prospective clinical trials will be essential to validate the efficacy and safety of multimodal generative AI systems in real-world clinical settings. These clinical validations must address not only technical performance metrics but also practical considerations such as workflow integration, user experience for clinicians, and ultimately, impact on patient outcomes.

In conclusion, multimodal generative AI, through various design strategies including tool use, grafting, and unification, shows great promise in enhancing medical diagnostics. By leveraging diverse data types and advanced machine learning techniques, these systems have the potential to improve diagnostic accuracy, streamline clinical workflows, and contribute to more personalized patient care. However, ongoing research and development efforts are needed to address current limitations and ensure the responsible and effective integration of these technologies into clinical practice. The continued evolution of these approaches will likely lead to increasingly sophisticated diagnostic tools that can serve as valuable assistants to healthcare providers, ultimately benefiting patient care across numerous medical specialties.

2.3. Tradeoffs in Model Design

Multimodal generative AI models have emerged as powerful tools in medical diagnostics, offering the potential to enhance diagnostic accuracy, improve workflow efficiency, and advance research capabilities. This review explores the applications, benefits, and challenges of multimodal generative AI in diagnostics, with a focus on model performance and the trade-offs between specialization and generalization. Multimodal AI models integrate diverse data sources such as medical imaging, genomic information, electronic health records, and clinical notes to provide a more comprehensive analysis of patient health [15,37]. This approach aligns with clinical practices where physicians rely on multiple information sources for decision-making. The integration of multimodal data has shown clear advantages in enhancing diagnostic accuracy by combining radiographic images with patient history and clinical notes, reflecting how clinicians naturally synthesize information from various sources to form diagnostic conclusions [15].

Several studies have demonstrated the superiority of multimodal approaches over unimodal methods in various diagnostic tasks. For instance, a multiscale and multimodal deep neural network classifier built with a combination of fluorodeoxyglucose-positron emission tomography (FDG-PET) and structural magnetic resonance imaging (MRI) showed improved performance in diagnosing neurodegenerative disorders [38]. Similarly, in breast cancer detection, multimodal approaches generally outperformed unimodal methods, particularly when integrating imaging data with genomic profiles and clinical records [14]. These findings suggest that the complementary information provided by different modalities can compensate for the limitations of individual data types, leading to more robust

 doi:10.20944/preprints202508.1425.v1

diagnostic outcomes. However, the implementation of multimodal AI in diagnostics presents several challenges. One significant issue is the complexity of integrating heterogeneous data types, which can lead to increased computational demands and potential overfitting [39]. The dimensionality and structural differences between imaging, textual, and genomic data require sophisticated fusion strategies to effectively leverage their combined information content. Additionally, multimodal approaches may not always be necessary or advantageous, depending on the specific clinical context and the quality of individual data modalities [14]. In some cases, the additional complexity of multimodal models may not justify the marginal improvements in performance, particularly when one data modality contains sufficient diagnostic information.

The trade-off between specialization and generalization is a crucial consideration in the development of multimodal AI models for diagnostics. Specialized models trained on specific tasks or disease types may achieve higher accuracy within their narrow domain but may lack generalizability to other conditions or populations. Conversely, more generalized models may offer broader applicability but potentially at the cost of reduced performance in specific scenarios [27]. This balance becomes particularly important in clinical settings where both high accuracy for common conditions and adaptability to rare presentations are desirable characteristics of diagnostic tools. To address these challenges, researchers are exploring various approaches to multimodal integration. Early, intermediate, and late fusion methods, as well as advanced deep multimodal fusion techniques such as encoder-decoder architectures, attention-based mechanisms, and graph neural networks, are being investigated to optimize the integration of diverse data types [14]. These techniques aim to leverage the strengths of each modality while mitigating their respective weaknesses. The choice of fusion strategy significantly impacts how effectively a multimodal model can capture inter-modal relationships and extract complementary information from different data sources.

Recent advancements in multimodal AI include the development of vision-language models that can process both images and clinical reports. These models have shown promise in tasks such as radiology report generation, visual question answering, and cross-modal retrieval [38]. For example, in chest radiograph interpretation, multimodal models have demonstrated the ability to generate preliminary reports and provide diagnostic insights by analyzing both images and associated textual data [37]. Such capabilities represent a significant step toward AI systems that can reason across modalities in a manner similar to human clinicians, potentially improving both the accuracy and efficiency of diagnostic workflows. The potential of multimodal generative AI extends beyond diagnosis to areas such as treatment planning and prognosis. By integrating diverse patient data, these models can potentially provide more personalized and comprehensive assessments of disease progression and treatment efficacy [33]. The ability to synthesize information from clinical histories, laboratory results, imaging studies, and genomic profiles could enable more precise risk stratification and therapy selection. However, realizing this potential requires addressing challenges related to data quality, standardization, and integration of temporal information [15]. Longitudinal data analysis presents particular difficulties due to varying sampling frequencies and missing data points across different modalities. As the field progresses, there is a growing emphasis on the need for explainable AI (XAI) in multimodal diagnostic models. XAI methods such as Grad-CAM, SHAP, and LIME are being explored to enhance the interpretability and transparency of AI-driven diagnostics [14]. This focus on explainability is crucial for building trust among healthcare professionals and ensuring the ethical implementation of AI in clinical practice. Interpretable multimodal models allow clinicians to understand which features from each data source contributed to a particular diagnostic suggestion, facilitating appropriate levels of human oversight and intervention when necessary.

In conclusion, multimodal generative AI holds significant promise for enhancing diagnostic capabilities in medicine. While challenges remain in terms of data integration, model performance, and the balance between specialization and generalization, ongoing research and technological advancements are paving the way for more sophisticated and clinically valuable AI-driven diagnostic tools. Future developments in this field will likely focus on improving data fusion techniques, enhancing

 doi:10.20944/preprints202508.1425.v1

model interpretability, and validating the clinical utility of multimodal AI across diverse healthcare settings and patient populations. The continued evolution of these technologies may ultimately lead to diagnostic systems that not only match but potentially exceed human capabilities in certain aspects of medical diagnosis, while serving as valuable adjuncts to clinical decision-making.

2.4. Specialty Applications Overview

Multimodal generative AI is emerging as a powerful approach for integrating diverse data types to enhance medical diagnostics and decision-making across multiple specialties. Key developments and applications in radiology, pathology, dermatology, and ophthalmology include:

1. Radiology:
   Multimodal AI models that combine imaging data (e.g. X-rays, CT, MRI) with clinical information from electronic health records are showing improved diagnostic accuracy compared to single-modality approaches [15]. For example, models integrating chest X-rays with patient history and clinical notes have demonstrated enhanced performance in disease classification tasks [37]. Advanced techniques like fusion-based methods and representation learning allow these models to effectively combine visual and textual data [24]. There is also growing interest in cross-modality translation, such as automatically generating radiology reports from images [37], which could significantly reduce radiologists' workload while maintaining diagnostic quality.
2. Pathology:
   AI systems that analyze both pathology slide images and molecular/genomic data are being developed to provide more comprehensive tumor characterization and prognostic stratification [22]. The integration of radiomics features from imaging with transcriptomic data has shown superior predictive capability for treatment responses in some cancers compared to single-modality approaches [22]. This comprehensive analysis mirrors the increasing clinical emphasis on integrated diagnostics, where pathologists collaborate with radiologists and other specialists to formulate more precise diagnostic and treatment plans.
3. Dermatology:
   Multimodal models combining clinical images, dermoscopic images, and patient metadata are being explored to improve skin lesion classification and melanoma detection [40]. These integrated approaches aim to mimic the multifaceted diagnostic process of dermatologists, who routinely consider visual features alongside patient history, risk factors, and other clinical information when making diagnostic decisions. By synthesizing these diverse inputs, multimodal systems offer potential improvements in sensitivity and specificity for skin cancer detection.
4. Ophthalmology:
   Generative AI techniques like GANs are being used to create synthetic retinal images to expand training datasets [40]. Multimodal foundational models capable of processing both eye images and clinical text show promise for enhancing diagnostic accuracy, patient education, and clinician training in ophthalmology [40]. These models can potentially detect subtle retinal changes associated with systemic diseases like diabetes and hypertension, facilitating earlier intervention and better management of these conditions.

Across specialties, key advantages of multimodal generative AI include: Multimodal generative AI offers numerous advantages across medical specialties. The integration of diverse data sources significantly improves diagnostic accuracy by leveraging complementary information from different modalities [15,22]. This approach enhances the ability to model complex disease processes and patient-specific factors, allowing for more personalized diagnostic and treatment strategies [22]. Additionally, these systems demonstrate potential to automate time-consuming tasks like report generation, potentially reducing physician workload and administrative burden [37]. They also create opportunities for data augmentation and synthetic data creation, addressing data scarcity issues that have traditionally hindered AI development in specialized medical fields [40]. The combined effect of these advantages is a more comprehensive and nuanced approach to medical diagnostics that more closely

 doi:10.20944/preprints202508.1425.v1

resembles the multifaceted clinical reasoning process employed by experienced clinicians. However, significant challenges remain, including: Despite their promise, multimodal generative AI systems face substantial implementation challenges. There remains a pressing need for large, diverse, high-quality multimodal datasets for training these complex systems [25]. Technical difficulties in data integration and model interpretability present ongoing obstacles, as different data types often require specialized processing approaches to extract meaningful patterns [25]. There are also legitimate concerns about potential biases and errors if models are not carefully validated across diverse populations and clinical scenarios [13]. Furthermore, regulatory and ethical considerations around data privacy and clinical implementation require careful attention to ensure patient protection [13]. These challenges necessitate interdisciplinary collaboration between AI researchers, clinical experts, and regulatory bodies to develop frameworks that balance innovation with patient safety and ethical considerations. As the field advances, careful validation, ethical oversight, and interdisciplinary collaboration will be crucial to realize the potential of multimodal generative AI while ensuring patient safety and equitable care. The technology shows great promise, but should be viewed as a tool to augment rather than replace clinical expertise. Future developments will likely focus on improving model interpretability, establishing rigorous validation protocols across diverse populations, and creating seamless integration pathways into existing clinical workflows. With appropriate development and implementation, multimodal generative AI has the potential to transform medical diagnostics across specialties, potentially improving healthcare outcomes while increasing efficiency and accessibility.

3. Applications Across Medical Specialties

3.1. Radiology Applications

Multimodal generative AI has shown significant potential in transforming diagnostic processes in radiology, particularly in areas such as image-to-text report generation, visual question answering (VQA), image captioning, and the utilization of datasets like ROCO. These advancements are poised to enhance the efficiency and accuracy of radiological diagnoses while potentially reducing the workload on healthcare professionals. Image-to-text report generation has emerged as a crucial application of multimodal generative AI in radiology. This technique involves the automatic conversion of visual data, such as radiographs, into descriptive text reports [14]. The process can significantly streamline the workflow of radiologists by automating the creation of preliminary interpretations for various imaging modalities, including chest radiographs. Recent studies have demonstrated the feasibility and potential clinical value of such systems. For instance, a domain-specific multimodal generative AI model was developed and evaluated for providing preliminary interpretations of chest radiographs, showing promising results in terms of diagnostic accuracy and clinical utility [37]. These systems represent a significant advance in automation of one of the most time-consuming aspects of radiological practice.

Visual Question Answering (VQA) represents another important application of multimodal generative AI in radiology. VQA systems are designed to answer natural language questions about medical images, bridging the gap between visual and textual information [41]. These systems have the potential to assist radiologists in interpreting complex images by providing relevant information in response to specific queries. Recent advancements in VQA for medical imaging include the development of models like Medical Visual Instruction Tuning (Med-VInT), which aims to perform generative-based medical visual question answering [41]. The interactive nature of VQA systems makes them particularly valuable for both education and clinical consultation scenarios where specific diagnostic questions need targeted answers based on image content. Image captioning, closely related to report generation, involves the automatic creation of descriptive captions for medical images. This application can aid in quick summarization of image content and facilitate efficient communication between healthcare professionals. Multimodal learning approaches, which combine information from various data types such as images and text, have shown particular promise in this area [15]. Unlike full report generation, captioning provides concise descriptions that can serve as rapid screening tools or aid in prioritizing worklists for radiologists, potentially addressing workflow bottlenecks in busy clinical environments.

 doi:10.20944/preprints202508.1425.v1

The ROCO (Radiology Objects in COntext) dataset has played a significant role in advancing multimodal AI research in radiology. This large-scale dataset, containing medical images paired with their respective captions, has been instrumental in training and evaluating various multimodal AI models [42]. The availability of such comprehensive datasets is crucial for developing robust and clinically relevant AI systems in radiology. Beyond ROCO, other important datasets such as NIH14 and MIMIC CXR have enabled researchers to train increasingly sophisticated models that demonstrate superior performance across multiple tasks compared to earlier approaches [15]. Multimodal generative AI models in radiology often leverage advanced architectures such as transformers and generative adversarial networks (GANs). These approaches utilize transformer-based text-image pre-training architectures to acquire representations of both modalities mutually [15]. The architectural innovation in this field has been rapid, with models demonstrating increasingly sophisticated capabilities in understanding the complex relationship between visual features in medical images and their corresponding clinical descriptions. These technical advances have led to improvements in classification, retrieval, and image synthesis compared to traditional unimodal transformer models [15], suggesting that the synergy between different data types enhances overall model performance. Despite the promising advancements, several challenges remain in the widespread adoption of multimodal generative AI in clinical radiology practice. These include ensuring the accuracy and reliability of generated reports, addressing potential biases in training data, maintaining patient privacy, and integrating these systems seamlessly into existing clinical workflows [17]. The issue of explainability remains particularly critical in medical applications, as clinicians need to understand how AI systems reach their conclusions to maintain appropriate oversight and responsibility for patient care. Additionally, there is a need for robust validation of AI-generated outputs and clear guidelines for their use in clinical decision-making. Regulatory frameworks are still evolving to address the unique challenges posed by AI systems that generate content rather than simply classify or detect anomalies.

In conclusion, multimodal generative AI is rapidly evolving and shows great promise in enhancing radiological diagnostics through applications like automated report generation, visual question answering, and image captioning. The integration of diverse data modalities and advanced AI architectures is paving the way for more comprehensive and efficient diagnostic tools. The synergistic combination of visual and textual information processing mirrors the cognitive processes radiologists themselves use when interpreting studies and communicating findings. However, continued research, rigorous clinical validation, and careful consideration of ethical and practical implications will be crucial as these technologies move closer to widespread clinical implementation. The goal remains to augment rather than replace human expertise, creating a collaborative workflow where AI handles routine aspects while radiologists focus on complex cases and oversight, ultimately improving patient care through more efficient and accurate diagnostic processes.

3.2. Pathology Applications

Recent advancements in multimodal generative AI have shown promising applications in histopathology image analysis and interpretation. These approaches combine visual information from histology slides with other data modalities to enhance diagnostic capabilities and extract more comprehensive insights. One key area of development is histopathology image question answering (QA) and visual question answering (VQA) for pathology applications. Large-scale datasets like PMC-VQA have been created, containing over 200,000 question-answer pairs across various medical image modalities including histopathology [41]. Models trained on these datasets can generate free-form answers to questions about histology images, providing a more natural interaction paradigm for pathologists. The ability to query image contents using natural language enables fine-grained interpretation of histological features.

Beyond question answering capabilities, multimodal approaches that integrate histopathology images with genomic data have demonstrated improved performance for tasks like cancer subtyping and mutation prediction [27]. Convolutional neural networks trained on both H&E slides and genomic profiles have shown the ability to predict gene mutations and transcriptional subtypes in cancer

 doi:10.20944/preprints202508.1425.v1

samples, allowing extraction of molecular insights directly from histology images and potentially reducing the need for additional molecular testing. Generative models like GANs are also being applied to histopathology for tasks such as stain normalization, virtual staining, and synthetic image generation [43]. These techniques help standardize image appearance across laboratories, translate between staining protocols, and augment training data. The creation of synthetic histopathology images addresses data scarcity issues in rare conditions, though ethical considerations around their use must be carefully evaluated.

The integration of large language models with vision encoders represents a significant advancement, enabling more sophisticated reasoning over histopathology images [41]. These combined architectures can now generate detailed descriptions of pathological findings, answer open-ended questions, and provide explanations for their predictions. This evolution moves beyond simple classification towards AI systems that can engage in nuanced dialogue about histological features, potentially serving as educational tools and decision support systems. Despite these promising developments, several challenges remain in the clinical implementation of multimodal generative AI for histopathology. Ensuring model interpretability and explaining AI decision-making processes is crucial for clinical adoption. Methods like attention visualization and image captioning are being explored to provide more transparency in how models process visual information [14]. The "black box" nature of deep learning models remains a significant concern that must be addressed before widespread clinical implementation can occur.

Data quality, bias mitigation, and privacy protection are other key considerations as larger multimodal datasets are assembled [43]. Careful curation and annotation of histopathology images paired with clinical data is essential to ensure models learn meaningful patterns rather than artifacts. Additionally, techniques to preserve patient privacy when training on sensitive medical data require further development, particularly as multimodal approaches often combine multiple protected health information sources. In conclusion, multimodal generative AI approaches are expanding the capabilities of computational pathology beyond traditional image analysis. By integrating visual, textual, and molecular data, these systems can provide richer and more nuanced interpretation of histopathology samples. However, thoughtful implementation that addresses interpretability, bias, and ethical concerns will be critical as these technologies move towards clinical application. Ongoing research in explainable AI, privacy-preserving learning, and clinical validation will help realize the potential of multimodal generative AI to augment and enhance pathology practice.

3.3. Dermatology and Ophthalmology Applications

Multimodal generative AI is revolutionizing the field of diagnostics, particularly in dermatology and ophthalmology, by enabling conversational support for imaging analysis and interpretation. This integration of visual and language models is enhancing diagnostic accuracy, streamlining clinical workflows, and improving patient care. In dermatology, multimodal AI systems are being developed to analyze skin lesion images in conjunction with clinical information and patient histories. These systems can provide detailed assessments of various skin conditions, assisting dermatologists in their diagnostic process. For instance, a novel AI-empowered methodology has been designed to support dermatologists' workflows in assessing and diagnosing skin conditions. This system employs large language models and transformer-based vision models for image analysis, along with sophisticated machine learning tools for guideline-based segmentation and measurement tasks [44]. By applying sequential logic with agency, the system achieved a weighted accuracy of 87% on the dataset used, demonstrating its reasoning and diagnostic capabilities.

The potential applications of multimodal AI in dermatology extend beyond diagnosis. These systems can be used to generate personalized patient education materials, create synthetic images for training purposes, and even assist in treatment planning. For example, AI-generated images could be used to show patients how their appearance might change after procedures like ptosis surgery or orbital decompression surgery, enhancing the informed consent process [40]. In ophthalmology, multimodal AI is making similar significant strides in diagnostic support and patient care. AI models

 doi:10.20944/preprints202508.1425.v1

analyze various types of ophthalmic imaging data, including fundus photographs, optical coherence tomography (OCT) scans, and external eye photographs. These models integrate imaging findings with clinical information to provide comprehensive diagnostic assessments. One study explored the potential of GPT-4 V, a multimodal large language model, in ophthalmology diagnostics. While the model's performance was not yet suitable for clinical application, it showed promise in simultaneously analyzing and integrating visual and textual data [42]. The study highlighted the potential of multimodal large language models to advance patient care, education, and research in ophthalmology.

Particularly noteworthy is the Eye-AD framework, validated in a multi-center study of 1671 participants, which analyzed OCT angiography images of retinal microvasculature and choriocapillaris to detect Alzheimer's disease and mild cognitive impairment with high accuracy [16]. This approach leverages the retina as a window into central nervous system health, offering a non-invasive biomarker for neurological diseases. Beyond diagnostics, multimodal AI in ophthalmology enhances patient education and clinical communication through AI-generated images and simulations that help patients understand their conditions and potential treatment outcomes. For example, ophthalmologists can create personalized prognostic markers to show patients how their vision may change if their condition worsens or what they could expect from treatment [40]. The integration of multimodal AI in dermatology and ophthalmology imaging analysis offers several advantages. Enhanced diagnostic accuracy results from combining visual data with clinical information and patient histories, providing more comprehensive and accurate assessments of skin and eye conditions. Clinical workflows are streamlined as these systems quickly analyze complex imaging data and provide detailed reports, potentially reducing diagnostic turnaround times. Patient education becomes more personalized through AI-generated images and simulations that improve informed consent and patient engagement. Additionally, these tools provide valuable guidance for primary care physicians and non-specialists in interpreting specialized imaging, potentially improving access to care. For research and training purposes, AI-generated synthetic images and cases can augment training datasets and create educational materials for medical professionals. Despite these promising developments, several challenges need addressing as these technologies evolve. Data bias and inclusivity remain critical concerns, as ensuring AI models are trained on diverse and representative datasets is crucial to avoid perpetuating healthcare disparities [28]. Safety and reliability issues, particularly AI "hallucinations" or incorrect outputs, must be resolved to ensure patient safety and maintain trust in these systems. Privacy and data security require robust measures to protect sensitive patient information and prevent data breaches. The rapid development of AI technologies necessitates appropriate regulatory frameworks to govern their clinical use. Finally, successful implementation depends on developing user-friendly interfaces and clear guidelines for appropriate integration into clinical workflows [40].

In conclusion, multimodal generative AI is poised to significantly enhance conversational support for dermatologic and ophthalmologic imaging analysis. By combining the power of visual and language models, these systems offer the potential to improve diagnostic accuracy, streamline clinical workflows, and enhance patient care. However, careful consideration of ethical, regulatory, and practical challenges is necessary to ensure the responsible and effective integration of these technologies into medical practice.

3.4. Benchmark Overview

The landscape of multimodal generative AI in diagnostics has seen significant advancements in recent years, with researchers developing innovative approaches to integrate diverse data types and improve diagnostic accuracy across medical specialties. This review explores the current state of datasets, benchmarks, and evaluation protocols in this rapidly evolving field. Multimodal datasets have become increasingly prevalent in diagnostic applications, combining imaging data with other clinical information to provide a more comprehensive view of patient health. The MIMIC series integrates chest X-rays, radiology reports, and electronic health records, serving as a cornerstone for multimodal research in radiology [45]. The UK Biobank combines clinical risk factors with fundus photographs for cardiovascular disease prediction, while EchoNet-Dynamic pairs cardiac MRI imaging

 doi:10.20944/preprints202508.1425.v1

with ECG data for predicting cardiac resynchronization therapy response [46]. Other notable examples include CLARO, which integrates CT images with clinical data for lung cancer studies, and LCID, which combines CT images with lung tumor biomarkers [46]. Additionally, datasets for neurological disorders incorporate brain MRI images, electronic health records (EHR), and free-text clinical notes to provide multifaceted views of patient conditions [46].

While these datasets have facilitated significant progress, standardized benchmarks for evaluating multimodal generative AI models in diagnostics remain limited. Researchers have employed various performance metrics and evaluation protocols, but there is a lack of consistency across studies, making direct comparisons challenging. Common performance metrics used in multimodal diagnostic studies include accuracy, precision, recall, and specificity; Area Under the Receiver Operating Characteristic curve (AUC-ROC); F1-score; and Cohen's kappa coefficient. These metrics provide quantitative assessments of model performance but must be considered in the context of clinical relevance and practical utility. Recent studies have demonstrated the potential of multimodal approaches to outperform single-modality models. A study by Zhang et al. showed a 19% increase in AUROC for predicting multiple sclerosis severity by fusing MRI images, EHR data, and clinical notes [35]. In breast cancer detection, the integration of histopathology images with genomic data and clinical records has shown promise in enhancing diagnostic accuracy [14]. Furthermore, multimodal generative AI models in anatomic pathology have demonstrated moderate to high accuracy in tasks such as image classification, segmentation, and text-to-image retrieval [29]. These findings consistently indicate that leveraging complementary information from multiple data sources can lead to more robust and accurate diagnostic systems.

To address the lack of standardized benchmarks, several initiatives and best practices are emerging in the field. Researchers are advocating for multicenter collaborations to validate models on diverse real-world data, addressing the limitation of models trained on data from a few academic institutions [29]. Federated learning approaches are being developed to build consensus models while maintaining data privacy and security, which is particularly important for sensitive medical information [29]. Carefully curated synthetic anatomic pathology data is being explored to address data scarcity issues and improve model generalizability [29]. Additionally, there is a growing emphasis on developing comprehensive evaluation frameworks that assess not only diagnostic accuracy but also model interpretability, fairness, and clinical relevance. Benchmark competitions and challenges have been instrumental in driving progress in multimodal generative AI for diagnostics. These initiatives typically involve standardized datasets and tasks, clear evaluation metrics and protocols, public leaderboards to track progress, and opportunities for researchers to compare approaches and share insights. Such competitions foster innovation and collaboration while establishing common ground for evaluating different methodologies in a controlled and fair manner.

Quantitative performance comparisons across different multimodal approaches have revealed several important insights for diagnostic applications. Studies have explored various fusion techniques, including early, intermediate, and late fusion, as well as more sophisticated techniques like encoder-decoder architectures and attention-based mechanisms [14]. The integration of descriptive text with imaging data has consistently shown significant improvements in diagnostic accuracy. One study reported an overall accuracy increase from 43.2% to 61.4% ($P<0.001$) when combining imaging with descriptive text across multiple models and prompt types [26]. Different multimodal models have shown varying levels of improvement with the addition of multiple data types. For instance, Claude 3.5 Sonnet achieved the highest overall performance (66.2% accuracy) when combining imaging and descriptive text inputs, particularly with AI-generated prompts (71.6% accuracy) [14]. The effectiveness of various prompt engineering strategies has also been evaluated, with AI-generated prompts often yielding superior combined accuracy across models compared to basic prompts, chain-of-thought prompts, reflection prompts, and multiagent approaches [26]. As the field progresses, there is a growing emphasis on explainable AI (XAI) in multimodal diagnostic systems. Methods such as Grad-CAM, SHAP, LIME, and image captioning are being employed to enhance the interpretability of

 doi:10.20944/preprints202508.1425.v1

complex multimodal models [14]. This focus on explainability is crucial for building trust in AI-assisted diagnoses and ensuring that clinicians can understand and validate the model's decision-making process. Explainable models facilitate clinical adoption by providing transparency in how different data modalities contribute to the final diagnostic assessment, potentially highlighting subtle patterns that might be overlooked in conventional analysis.

In conclusion, while significant progress has been made in developing multimodal generative AI models for diagnostics, there remains a pressing need for standardized benchmarks, comprehensive evaluation frameworks, and large-scale validation studies. Future research should focus on establishing consistent evaluation protocols, improving data integration techniques, and enhancing model interpretability to fully realize the potential of multimodal generative AI in clinical diagnostics. Addressing these challenges will require collaborative efforts across disciplines, including computer science, medicine, and biostatistics, as well as engagement with regulatory bodies to ensure that advances in this field translate effectively to improved patient care.

3.5. VQA-RAD and ROCO

Multimodal generative AI models are making significant strides in enhancing radiology image analysis and interpretation, with VQA-RAD and ROCO emerging as important benchmarks for evaluating these systems. These datasets play a crucial role in advancing the field by providing standardized evaluation metrics and fostering innovation in multimodal AI for radiology.

3.5.1. VQA-RAD (Visual Question Answering in Radiology)

VQA-RAD is a specialized dataset designed to evaluate AI models' ability to answer questions about radiology images [41]. This benchmark contains approximately 3,500 question-answer pairs associated with 315 radiology images spanning various modalities, including chest X-rays, CT scans, and MRI images [42]. The dataset incorporates both open-ended and closed-ended questions, comprehensively addressing aspects such as modality recognition, abnormality detection, and anatomical understanding. To ensure clinical relevance and accuracy, radiologists manually created questions and answers for each image in the dataset. Evaluation typically employs accuracy metrics for closed-ended questions and BLEU scores for open-ended questions. Recent advancements in multimodal models have demonstrated promising results, achieving accuracies of up to 86.8% on closed-ended questions and 73.7% on open-ended questions [41]. This performance indicates significant progress in AI systems' ability to understand and interpret radiological imagery in a clinically meaningful context.

3.5.2. ROCO (Radiology Objects in COntext)

ROCO represents a larger-scale dataset focused on radiology image captioning and classification, containing over 81,000 radiology images with associated captions and classifications [42]. The dataset encompasses diverse radiology modalities and anatomical regions, providing broad coverage of clinical scenarios. Each image is paired with a descriptive caption extracted from scientific publications, offering rich contextual information that mirrors how images are described in clinical practice. Unlike VQA-RAD's manual annotation approach, ROCO employs an automated extraction method for captions from figure descriptions in radiology-related research papers, followed by refinement and validation processes. Evaluation of models trained on ROCO commonly utilizes metrics such as BLEU, METEOR, and CIDEr scores for caption generation tasks, along with classification accuracy for image categorization. State-of-the-art models leveraging ROCO have demonstrated substantial improvements in generating clinically accurate and relevant image descriptions [15], highlighting the dataset's value in advancing multimodal AI capabilities in radiology.

3.6. Importance and Impact

These benchmarks are driving significant progress in multimodal AI for radiology through several key mechanisms. By providing consistent metrics for comparing different models and approaches, VQA-RAD and ROCO enable systematic progress tracking in the field, establishing a common ground

 doi:10.20944/preprints202508.1425.v1

for researchers to evaluate competing methodologies. The incorporation of real-world radiology images and expert-curated questions/captions ensures that AI models are evaluated on clinically meaningful tasks rather than artificial scenarios disconnected from medical practice. Furthermore, these datasets encourage the development of AI systems that effectively integrate visual and textual information, mimicking the way radiologists interpret images in context—considering both the visual patterns in an image and the language used to describe findings. The diverse applications supported by these benchmarks, ranging from question answering to image captioning, foster the development of versatile AI assistants capable of supporting various aspects of radiological workflow and interpretation [41,42].

3.7. Limitations and Future Directions

Despite their substantial contributions to advancing multimodal AI in radiology, VQA-RAD and ROCO exhibit certain limitations that warrant consideration. VQA-RAD, in particular, is relatively small compared to other computer vision datasets, containing only 315 images, which may constrain its ability to train and evaluate more complex models requiring extensive data for optimal performance [47]. Additionally, both datasets may not fully represent the diversity of real-world clinical scenarios and rare conditions, potentially limiting the generalizability of models trained exclusively on these resources [48]. Another significant concern is the static nature of these benchmarks in a rapidly evolving field—as radiology techniques and knowledge advance, these datasets may gradually become outdated, failing to reflect contemporary clinical practices and imaging technologies.

Future developments in radiology image analysis benchmarks should address these limitations through several targeted approaches. Regularly refreshing datasets with new images and annotations would help reflect evolving clinical practices and emerging pathologies. Expanding dataset diversity to include a wider range of pathologies, demographics, and imaging modalities would improve model robustness and clinical applicability across varied patient populations. Integration of additional clinical data types, such as patient history and laboratory results, would enable evaluation of more comprehensive diagnostic AI systems that consider the full clinical context when interpreting images [49]. Furthermore, developing standardized metrics to assess AI models' ability to provide interpretable explanations for their predictions would address the critical need for transparency in clinical AI applications.

In conclusion, VQA-RAD and ROCO serve as vital tools for advancing multimodal generative AI in radiology. By providing standardized benchmarks, they enable researchers to quantitatively assess progress in developing AI systems that can interpret and describe medical images with increasing accuracy and clinical relevance. As the field evolves, these datasets and evaluation methods will likely adapt to address current limitations and push the boundaries of AI-assisted radiology, ultimately working toward the goal of creating AI systems that can provide meaningful clinical decision support in radiology practice.

4. Benchmark Datasets and Evaluation

4.1. PathVQA and Other QA Datasets

Multimodal generative AI approaches are transforming diagnostics in pathology by enabling the integration of diverse data types and the generation of synthetic datasets to augment limited real-world data. These technologies have shown particular promise in visual question answering (VQA) applications, where AI systems interpret pathology images and respond to clinical queries in natural language. The development of specialized pathology VQA datasets has been crucial to this progress. PathVQA emerged as one of the first datasets developed for pathology visual question answering, containing 32,795 question-answer pairs from 4,998 pathology images, with both open-ended and close-ended questions [41,50]. Despite its pioneering role, manual audits revealed limitations in the quality of some examples, likely due to the automated nature of its curation, highlighting the need for more rigorous dataset development approaches [50].

 doi:10.20944/preprints202508.1425.v1

To address these limitations, researchers developed more comprehensive benchmarks like PathQABench [50]. This expert-curated dataset includes 105 high-resolution regions of interest from whole slide images covering 11 tissue sites and 54 diagnoses. PathQABench features both open-ended and multiple-choice diagnostic questions, alongside clinical context summaries for each case. The questions span microscopy, diagnosis, clinical aspects, and ancillary testing, providing a more robust foundation for model development and evaluation. Complementing these efforts, researchers are also generating synthetic question-answer pairs about pathology images to expand training data for VQA models, creating more diverse and extensive datasets to improve model performance on real-world diagnostic queries [50]. Beyond dataset development, significant progress has been made in multimodal integration techniques. Advanced models now combine pathology images with other data modalities like genomics, clinical records, and patient histories [22,51]. This integration employs various approaches including early, intermediate, and late fusion methods, as well as attention mechanisms and graph neural networks, to effectively synthesize insights from these diverse data types [14]. Such integration is particularly valuable in pathology diagnostics, where contextual information often proves crucial for accurate interpretation.

Several innovative model architectures have emerged to address pathology VQA tasks. PathChat represents a vision-language AI assistant fine-tuned on over 456,000 diverse visual-language instructions specifically for pathology applications [50]. MedFuseNet offers an attention-based multimodal deep learning model designed for optimal fusion of different input modalities, while MedCLIP employs contrastive learning to scale usable training data by cleverly decoupling images and texts [15]. These architectures demonstrate various approaches to the complex challenge of integrating visual and textual information in the specialized domain of pathology. Performance evaluations have yielded promising results, with studies showing that multimodal generative AI models can achieve state-of-the-art performance on multiple-choice diagnostic questions from diverse tissue origins and disease models [50]. Human expert evaluations have further validated these findings, demonstrating that specialized pathology models often produce more accurate and pathologist-preferable responses to diverse pathology queries compared to general-purpose AI assistants [50]. These assessments provide important evidence for the potential clinical utility of such systems. Despite these advances, significant challenges remain in the development and deployment of multimodal generative AI for pathology diagnostics. Ensuring data quality and addressing biases in both real and synthetic datasets represents an ongoing concern. There is also the complex task of balancing model complexity with interpretability for clinical use, as healthcare applications demand both high performance and transparency. Additionally, ethical concerns and regulatory requirements for AI in healthcare must be carefully addressed, alongside rigorous validation of model performance across diverse patient populations and clinical settings to ensure generalizability and fairness. Looking forward, research continues to focus on developing more comprehensive and clinically relevant pathology VQA datasets that better represent the complexity and diversity of real-world cases. Improvements in model architectures are being pursued to better handle the intricacy of pathology data, particularly at the high resolutions required for accurate diagnosis. Enhancing the explainability and interpretability of AI-generated responses remains a priority for clinical acceptance and regulatory approval. Furthermore, researchers are exploring the integration of temporal data and additional modalities to improve diagnostic accuracy and provide more holistic patient assessments [15].

In conclusion, multimodal generative AI approaches, particularly in visual question answering and synthetic dataset creation, demonstrate significant potential for enhancing pathology diagnostics. These technologies could augment pathologists' capabilities, improve diagnostic accuracy, and potentially increase access to expert-level diagnostics. However, careful validation, ethical considerations, and thoughtful clinical integration strategies remain essential for realizing the full potential of these technologies in practice [14,15,22,25,25,41,50,51].

 doi:10.20944/preprints202508.1425.v1

4.2. MIMIC-CXR and ImageCLEF Challenges

Multimodal generative AI has emerged as a promising approach for enhancing diagnostic capabilities in medical imaging, particularly for large-scale X-ray report datasets and multi-task benchmarks. This field combines multiple data modalities, such as imaging, text reports, and clinical information, to improve the accuracy and comprehensiveness of diagnostic models. The MIMIC-CXR dataset has become a cornerstone for research in this area, containing over 377,000 chest X-rays associated with 227,835 imaging studies from 65,379 patients [16]. This extensive dataset provides paired image-text data crucial for training multimodal AI models. Complementing MIMIC-CXR, the ImageCLEF medical image classification and caption prediction challenges have played a significant role in advancing multimodal approaches. These challenges offer standardized datasets and evaluation frameworks that enable researchers to benchmark their models against state-of-the-art techniques [52].

Recent advances in multimodal transformer architectures have shown promising results in X-ray interpretation tasks. These models can process both image and text inputs, learning joint representations that capture the relationship between visual features and diagnostic language [22]. For instance, the MMBERT model demonstrated strong performance on the MIMIC-CXR dataset for report generation and abnormality classification [52]. This architectural innovation represents a significant step forward in the ability of AI systems to understand and interpret complex medical imagery alongside associated clinical text. Multimodal generative AI models leverage these large-scale datasets to perform various diagnostic tasks. For example, models can be trained to generate detailed radiology reports from chest X-ray images, combining visual analysis with natural language generation [17]. This capability has the potential to assist radiologists by providing initial draft reports and highlighting key findings. Other applications include visual question answering systems that can interpret X-ray images based on natural language queries [41]. These developments demonstrate how multimodal approaches can support clinicians across different aspects of the diagnostic workflow.

The evaluation of these multimodal models employs a combination of automated metrics and human assessment to ensure both technical performance and clinical utility. Automated metrics such as ROUGE-L and BertScore measure the semantic similarity between generated reports and reference texts [19], providing quantitative performance indicators. However, recognizing that these metrics may not fully capture clinical relevance, studies increasingly incorporate evaluations by experienced radiologists who assess the quality, diagnostic accuracy, and clinical relevance of AI-generated reports [53]. This dual evaluation approach helps bridge the gap between computational performance and real-world clinical applicability. Multi-task learning approaches further enhance model capabilities by allowing simultaneous performance of multiple diagnostic tasks, such as abnormality detection, severity grading, and report generation [26]. This integrated approach leverages the complementary nature of different tasks to improve overall performance. Recent comparative studies of multimodal large language models (LLMs) on radiological quiz cases found that models like Claude 3.5 Sonnet achieved the highest overall accuracy (46.3%) in imaging-only inputs, followed closely by GPT-4o (43.5%) and Gemini-1.5-Pro-002 (39.8%) [26]. These findings highlight the current capabilities and limitations of state-of-the-art multimodal models in clinical diagnostic scenarios.

Despite the promising advances, multimodal generative AI faces significant challenges that must be addressed before widespread clinical adoption. Ensuring the reliability and explainability of model outputs remains crucial for clinical integration. Methods for uncertainty quantification and interpretable AI require further development to provide clinicians with appropriate confidence levels in AI-generated diagnoses [14]. Moreover, potential biases in training data can significantly influence model performance, particularly with factors such as case rarity and knowledge cutoff dates [26]. Addressing these limitations requires ongoing refinement of both datasets and modeling approaches. Ethical and regulatory considerations form an essential framework for the development and deployment of these AI systems in healthcare settings. Privacy protection in medical AI systems can be achieved through techniques such as data anonymization, strategic noise injection, and federated learning architectures [14]. Additionally, evaluation of model fairness employs quantitative measures

 doi:10.20944/preprints202508.1425.v1

such as the Gini coefficient and Shannon diversity index to assess output diversity and detect potential biases across different demographic groups [14]. These methodological safeguards are essential to ensure that multimodal AI systems serve all patient populations equitably and respect confidentiality standards. In conclusion, large-scale X-ray datasets and multi-task benchmarks are driving rapid progress in multimodal generative AI for diagnostics. As these techniques continue to mature, they have the potential to significantly enhance radiologists' efficiency and accuracy in interpreting medical images. However, rigorous clinical validation, thoughtful implementation, and ongoing efforts to address ethical and regulatory challenges will be necessary to translate these advances into improved patient care. The integration of multimodal approaches represents not just a technological advance but a fundamental shift in how computational tools can support and enhance clinical decision-making in diagnostic imaging.

4.3. Synthetic Image Generation

The use of generative adversarial networks (GANs) and diffusion models for rare conditions in multimodal generative AI diagnostics represents a promising approach to address the challenges of limited data availability and improve diagnostic accuracy for uncommon diseases. These techniques offer significant potential in synthesizing realistic medical images and augmenting training datasets, particularly for rare conditions where obtaining large, diverse datasets can be difficult.

GANs have shown considerable promise in generating synthetic medical images that can be used to expand limited datasets for rare conditions. For example, in the field of histopathology, the CONCH (CONtrastive learning from Captions for Histopathology) model has demonstrated significant potential in addressing diseases with scarce data [24]. CONCH combines weakly supervised learning with a large dataset of approximately 1.17 million image-text pairs, including both H&E staining and immunohistochemistry (IHC) samples. This approach has shown promising results in zero-shot classification and cross-modal retrieval tasks, indicating its potential utility for rare disease diagnosis. Diffusion models, another class of generative AI techniques, have also gained traction in medical imaging synthesis. These models have shown the ability to generate high-quality, diverse medical images that can be used to augment training datasets for rare conditions. The application of diffusion models in medical imaging can help overcome the limitations of small datasets and improve the performance of diagnostic AI systems for uncommon diseases. The integration of GANs and diffusion models into multimodal AI systems further enhances their potential for rare condition diagnostics. For instance, the development of advanced multimodal generative AI pathology assistants, such as PathChat, demonstrates the power of combining visual and language models for interactive diagnostic assistance [24]. PathChat, which integrates a specially trained pathological image visual encoder with a pre-trained large language model, has shown superior performance in multiple-choice diagnostic tasks and open-ended question answering, particularly for complex cases that may include rare conditions.

These generative models offer numerous advantages in rare condition diagnostics. Through data augmentation, GANs and diffusion models can generate synthetic images to expand limited datasets, enabling more robust training of diagnostic AI models. This expansion contributes to improved generalization capabilities, as the creation of diverse synthetic samples helps AI systems better recognize unseen cases of rare conditions. Additionally, some studies have introduced visualization techniques like attention maps or class activation mapping within generative models, enhancing interpretability and helping pathologists understand the basis of AI decisions for rare conditions [24]. The multimodal integration aspect is particularly valuable, as these generative models can be combined with genomic information and clinical records to provide more comprehensive and accurate diagnoses for rare diseases. Despite their promise, the application of GANs and diffusion models for rare conditions presents several challenges that require careful consideration. Data quality and standardization remain crucial concerns, as ensuring the consistency and quality of synthetic images is essential, especially for rare conditions where real-world examples may be limited [24]. The generation and use of synthetic medical data also raise significant ethical considerations regarding privacy and require careful handling to protect patient confidentiality [54]. Before integration into clinical workflows for rare

 doi:10.20944/preprints202508.1425.v1

condition diagnostics, rigorous validation of synthetic data and generative models is necessary to ensure safety and efficacy. Furthermore, enhancing model interpretability is essential for building trust among clinicians and ensuring appropriate use of these technologies in rare disease diagnosis [19].

In conclusion, the application of GANs and diffusion models in multimodal generative AI for rare condition diagnostics shows great promise. These techniques can help overcome the challenges of limited data availability and improve diagnostic accuracy for uncommon diseases. However, careful consideration of data quality, ethical implications, and clinical validation is necessary to ensure their responsible and effective implementation in healthcare settings. As research in this field progresses, we can expect to see further advancements in the use of generative AI for rare condition diagnostics, ultimately leading to improved patient care and outcomes.

4.4. Synthetic Text and Rare Disease Simulation

Multimodal generative AI models are emerging as powerful tools for enhancing diagnostics and clinical decision support in medicine. By combining and analyzing data from multiple modalities like medical imaging, clinical text, genomics, and other sources, these models can provide more comprehensive and nuanced insights compared to unimodal approaches. The integration of text and imaging data is showing particular promise in clinical applications. Studies have demonstrated that multimodal models combining imaging and clinical text data consistently outperform unimodal approaches in tasks like breast cancer detection [14]. Techniques like early and late fusion are being explored to effectively combine features from different modalities. For instance, TieNet uses a CNN+RNN architecture to jointly analyze chest X-rays and associated reports, while more recent transformer-based models like VisualBERT can process images and text simultaneously, creating a more integrated analysis framework [14]. Large language models (LLMs) are being leveraged to generate synthetic clinical case narratives and pair them with relevant medical images, creating valuable datasets for training and evaluation. This approach helps address challenges related to limited or imbalanced real-world data, especially for rare diseases or uncommon clinical presentations. For example, Onto-CGAN combines disease ontology knowledge with generative adversarial networks to create synthetic data for unseen diseases not present in training sets [55]. Such synthetic data generation techniques can augment existing datasets, enabling the development of more robust diagnostic models that can recognize patterns across a wider spectrum of presentations.

Benchmark datasets incorporating paired image-text data are crucial for advancing this field. The IMAGene project exemplifies efforts to create comprehensive multimodal datasets, integrating clinical, radiomic, genomic, and environmental data for pancreatic cancer risk prediction [18]. Such initiatives facilitate the development and rigorous evaluation of multimodal AI systems by providing standardized data against which different algorithmic approaches can be compared. Recent studies have also investigated the diagnostic performance of commercial multimodal LLMs like GPT-4 and Claude in radiological image interpretation [26]. While these models show promising capabilities, their performance can vary significantly based on factors like case rarity, input modalities, and prompt engineering strategies. The integration of descriptive text alongside images was found to substantially improve diagnostic accuracy across multiple models, highlighting the synergistic value of combining different data types in clinical decision support. As multimodal generative AI continues to evolve, it has the potential to transform cancer diagnostics by enabling more holistic analysis of patient data [56]. These models can integrate diverse information sources to create comprehensive disease profiles, enhancing precision in diagnosis, prognosis, and treatment planning. However, challenges remain, including addressing biases in training data, ensuring interpretability of model outputs for clinician trust, and navigating complex regulatory frameworks for clinical deployment.

In conclusion, multimodal generative AI, particularly in the realm of LLM-generated case narratives and image-text pairing, represents a promising frontier in medical diagnostics. By leveraging the complementary strengths of different data modalities, these approaches can potentially improve diagnostic accuracy, facilitate rare disease detection, and advance personalized medicine. Continued research and development in this area, coupled with careful consideration of ethical and practical

 doi:10.20944/preprints202508.1425.v1

implications, will be crucial for realizing the full potential of multimodal generative AI in clinical practice.

5. Synthetic Data Generation for Rare Diseases

5.1. Privacy and Collaboration via Synthetic Data

Multimodal generative AI is emerging as a powerful tool for enhancing diagnostics and enabling privacy-safe collaboration in clinical research, particularly for rare diseases. This innovative approach combines multiple data modalities and leverages advanced AI techniques to generate synthetic data that closely mimics real patient information while preserving privacy. The generation of synthetic data addresses a critical challenge in rare disease research - the scarcity of patient data due to low disease prevalence. Chang Sun and Michel Dumontier proposed Onto-CGAN, a novel generative framework that integrates knowledge from disease ontologies with Generative Adversarial Networks (GANs) to create synthetic data for unseen diseases not present in the training set [55]. This approach demonstrated the ability to generate unseen disease data with statistical characteristics comparable to real data, offering valuable applications in data augmentation and hypothesis generation for rare diseases. Multimodal AI systems are driving a paradigm shift in modern biomedicine by seamlessly integrating heterogeneous data sources such as medical imaging, genomic information, and electronic health records [22]. This integration enables more comprehensive and precise diagnoses, supporting early disease detection and personalized treatment strategies. For instance, in the realm of biomaterials, AI facilitates the design of patient-specific solutions tailored for tissue engineering, drug delivery, and regenerative therapies. The convergence of multiple data modalities provides a more holistic view of patient conditions, particularly valuable when dealing with complex and rare diseases where single-modality data might be insufficient for accurate diagnosis. The potential of multimodal generative AI extends to various diagnostic applications across medical specialties. In radiology, AI systems can synthesize inputs from imaging, molecular markers, and clinical data to improve diagnostic precision [22]. Similarly, in pathology, advanced multimodal generative AI assistants like PathChat have demonstrated superior performance in analyzing pathological images and answering diagnostic questions conversationally [24]. These systems combine specially trained visual encoders with large language models to achieve truly interactive diagnostic assistance, representing a significant advancement in how clinicians can interact with diagnostic AI tools. One of the key advantages of synthetic data generation is its ability to facilitate cross-border privacy-safe collaboration. Parvin et al. highlighted that synthetic data aims to capture information of diagnostic utility while eliminating the possibility of patient re-identification [57]. This approach allows researchers to share data without violating patient privacy, potentially accelerating collaborative efforts in rare disease research. The privacy-preserving nature of synthetic data is particularly crucial in the context of international research collaborations, where varying regulatory frameworks around patient data protection can create significant barriers.

A recent study by Hagos et al. demonstrated the practical application of synthetic data generation for cross-border collaboration in acute myeloid leukemia (AML) research [58]. The researchers employed two different methodologies of generative AI - CTAB-GAN+ and normalizing flows (NFlow) - to synthesize patient data derived from 1606 AML patients treated within four multicenter clinical trials. Both generative models accurately captured distributions of demographic, laboratory, molecular, and cytogenetic variables, as well as patient outcomes. Importantly, the synthetic data preserved inter-variable relationships and survival curves while safeguarding patient privacy, mitigating the risk of re-identification. This case study provides compelling evidence for how generative AI can overcome traditional barriers to data sharing in clinical research. However, the implementation of multimodal generative AI in clinical practice faces several challenges that must be addressed for widespread adoption. These include ensuring robust data security mechanisms, meeting increasingly stringent regulatory standards across different jurisdictions, and promoting algorithmic transparency to build trust among clinicians and patients [22]. Additionally, there are valid concerns about potential biases in AI models that could perpetuate or exacerbate existing healthcare disparities, particularly if training

 doi:10.20944/preprints202508.1425.v1

data is not sufficiently diverse or representative. Ensuring equitable access to these technologies across different healthcare settings and geographic regions also remains a significant challenge that requires thoughtful policy development. Despite these challenges, the convergence of AI and biotechnology continues to shape a future where healthcare is more predictive, personalized, and responsive. The ability to generate high-quality synthetic data for rare diseases opens new avenues for research collaboration and accelerates the development of diagnostic and therapeutic strategies. As these technologies evolve, it will be crucial to address ethical considerations and establish robust regulatory frameworks to ensure their responsible and effective use in clinical practice [16,27]. Ongoing dialogue between technology developers, healthcare providers, patient advocates, and regulatory bodies will be essential to navigate the complex ethical landscape of synthetic data generation while maximizing its benefits for rare disease research.

In conclusion, multimodal generative AI presents a promising approach to overcoming data scarcity in rare disease research while enabling privacy-safe collaboration across borders. By generating synthetic data that closely mimics real patient information, this technology has the potential to accelerate diagnostic advancements, treatment discoveries, and collaborative efforts in the field of rare diseases. As research in this area progresses, it will be essential to balance the benefits of data sharing with the imperative of patient privacy protection, ensuring that these powerful tools are used ethically and effectively to improve patient care.

5.2. Clinical Trial Design and Validation Needs

Multimodal generative AI has emerged as a promising technology in medical diagnostics, offering the potential to enhance clinical decision-making, improve patient outcomes, and transform healthcare delivery. This review examines the current state and future prospects of multimodal generative AI in diagnostics, with a particular focus on comparative studies with human clinicians and failure analysis. Multimodal AI models integrate diverse data types, including medical imaging, electronic health records, genomic information, and real-time patient data, to provide a more comprehensive view of a patient's condition. This approach mirrors how clinicians synthesize various information sources but with the added capability of detecting patterns and interactions beyond human perception. For instance, in the diagnosis of neurodegenerative disorders like Alzheimer's disease, multimodal AI can simultaneously analyze MRI and PET scans along with clinical data, potentially enhancing diagnostic accuracy and risk prediction [59]. The integration of multiple data modalities allows these systems to develop a more holistic understanding of patient conditions, similar to how clinicians approach diagnosis but with enhanced computational capabilities. Comparative studies between multimodal AI systems and human clinicians have shown promising results across various medical specialties. In pathology, the PathChat system demonstrated superior performance in multiple-choice diagnostic tasks compared to other AI models and human experts. PathChat achieved an accuracy of 78.1% in image-only tasks, which increased to 89.5% when clinical background information was provided [60]. This performance was significantly higher than comparative models and human pathologists, particularly in questions requiring careful examination of histological images. The substantial improvement when clinical information was incorporated underscores the value of multimodal approaches in diagnostic accuracy.

In radiology, multimodal AI models have shown potential in improving diagnostic accuracy and efficiency. For example, a study on chest X-ray interpretation found that AI-assisted radiologists demonstrated improved diagnostic accuracy and reduced reading time compared to radiologists working without AI assistance [61]. Similarly, in dermatology, AI models integrating visual and clinical data have shown comparable or superior performance to dermatologists in diagnosing skin lesions [25]. These findings suggest that multimodal AI systems can serve as valuable diagnostic aids across multiple medical domains. However, it is crucial to note that these comparisons often focus on specific diagnostic tasks and may not fully capture the nuanced decision-making process of experienced clinicians. The integration of AI into clinical practice requires careful consideration of how these tools can augment rather than replace human expertise. Contextual understanding,

 doi:10.20944/preprints202508.1425.v1

ethical considerations, and patient-specific factors that inform clinical judgment remain challenging aspects for AI systems to fully replicate. Failure analysis of multimodal generative AI in diagnostics has revealed several key challenges. The accuracy of AI diagnostic systems is highly dependent on the quality and consistency of input data. Variations in pathological slice preparation, imaging protocols, and diagnostic standards across different institutions can affect model performance [60]. Models trained on data from specific populations or institutions may not perform well when applied to diverse patient groups or in different clinical settings [61]. This limitation highlights the need for diverse, representative datasets in model development and validation.

While some models incorporate visualization techniques like attention maps or class activation mapping, there is still a need for improved methods to help clinicians understand the basis of AI decisions [60]. The "black box" nature of many AI models poses challenges for clinical adoption and regulatory approval. Traditional AI models often struggle with rare diseases or presentations not well-represented in training data. However, innovative approaches like the Onto-CGAN framework have shown promise in generating synthetic data for unseen diseases, potentially addressing this limitation [59]. Many current models focus on static data points, but there is a growing need to incorporate temporal information, such as disease progression over time, into diagnostic models [61]. Additionally, the use of AI in healthcare raises important ethical questions regarding data privacy, informed consent, and the potential for bias in decision-making [62]. Regulatory frameworks are still evolving to address these challenges, creating uncertainty in implementation pathways for novel AI diagnostic tools. To address these challenges and advance the field, several strategies have been proposed. Development of large-scale, diverse multimodal datasets across various anatomical domains and disease types is essential [61]. This includes efforts to create standardized, high-quality datasets that represent diverse patient populations. Implementation of federated learning and edge computing techniques can address data privacy and integration concerns [13]. These approaches allow for model training across multiple institutions without compromising patient data privacy, facilitating the development of more robust and generalizable models.

Continued research into model interpretability and explainability is needed to build trust and facilitate clinical adoption [60]. This includes developing visualization tools and techniques that allow clinicians to understand the reasoning behind AI-generated diagnoses. Prospective, multi-center clinical trials are crucial to validate the real-world performance and impact of multimodal AI systems [60]. Such studies are essential for demonstrating the clinical utility and safety of AI diagnostic tools before widespread implementation. Establishment of quality control and responsibility tracing mechanisms is necessary for long-term, safe implementation of AI in clinical practice [60]. This includes developing clear protocols for AI model maintenance, updating, and monitoring in clinical settings. Integration of AI systems with existing clinical workflows and electronic health record systems will ensure seamless adoption and use by healthcare providers [45], minimizing disruption to established clinical processes while maximizing potential benefits. As the field progresses, there is a growing emphasis on creating AI systems that can engage in interactive, multi-round reasoning with clinicians. This approach, exemplified by systems like PathChat and CareAssist-GPT, allows for clarification of doubts, incorporation of new information, and collaborative decision-making between AI and human experts [45,60]. These interactive systems aim to enhance not only diagnostic accuracy but also patient communication and engagement in the diagnostic process. The evolution toward conversational and collaborative AI systems represents a significant advancement over earlier unidirectional diagnostic tools, potentially addressing many of the limitations identified in failure analyses.

In conclusion, multimodal generative AI holds immense potential to transform medical diagnostics by integrating diverse data types and augmenting clinical decision-making. While comparative studies have shown promising results, ongoing research must address challenges in data quality, interpretability, and real-world validation. The future of AI in diagnostics lies not in replacing human clinicians, but in creating synergistic systems that combine the pattern recognition capabilities of AI with the nuanced judgment and experience of healthcare professionals. As these technologies continue

 doi:10.20944/preprints202508.1425.v1

to evolve, it is crucial to maintain a focus on patient-centered care, ethical considerations, and the responsible development and deployment of AI in healthcare settings.

5.3. Human-AI Collaboration Models

Multimodal generative AI is emerging as a powerful tool in medical diagnostics, with significant potential to enhance clinical decision-making and improve patient outcomes. However, its integration into healthcare systems raises important considerations around trust calibration and the balance between decision support and automation. Multimodal AI systems can integrate diverse data types such as medical imaging, electronic health records, genomic information, and real-time patient data to provide more comprehensive analyses [18,22]. This allows for a more holistic view of patient health and can lead to improved diagnostic accuracy and personalized treatment strategies. For example, in radiology, multimodal approaches combining imaging data with clinical information have shown enhanced performance in disease classification and risk prediction [21]. The ability to synthesize information across different modalities represents a significant advancement over single-modality systems that may miss critical correlations between diverse clinical data sources. A key advantage of multimodal generative AI is its ability to synthesize and interpret complex, heterogeneous data in ways that may surpass human capabilities. Models like PathChat have demonstrated impressive accuracy in pathology diagnosis tasks, outperforming other AI models and approaching expert-level performance [24]. Such systems can potentially augment clinical expertise, especially in areas with limited specialists or for rare diseases. The generative capabilities of these systems also enable them to produce detailed explanations and visualizations that can enhance clinicians' understanding of complex cases and serve as educational tools for medical training. However, the integration of these powerful AI tools into clinical workflows raises critical questions about trust calibration - how much clinicians should rely on AI-generated insights versus their own judgment. Studies have shown that excessive reliance on AI-based diagnostic support tools can lead to automation bias and potentially adverse outcomes [63]. There is a risk of clinicians becoming over-confident in AI systems, especially if the reasoning behind AI decisions is not transparent. Conversely, underutilization of accurate AI recommendations due to distrust could deprive patients of beneficial insights, creating a delicate balance that healthcare organizations must navigate.

To address these trust calibration challenges, several approaches are being explored. These include providing clinicians with the AI system's reasoning process, developing meta-AI systems to indicate when to rely on AI-driven support, and implementing interactive, multi-round dialogue capabilities that allow clinicians to probe the AI's rationale [24,63]. These interaction mechanisms create a collaborative diagnostic environment where clinicians can interrogate the AI's conclusions and understand the underlying evidence and confidence levels. The goal is to create a collaborative human-AI diagnostic model where AI augments rather than replaces clinical expertise, establishing appropriate reliance that leverages the strengths of both human judgment and computational analysis. The balance between using AI for decision support versus automation is another key consideration in the deployment of multimodal generative AI. While AI can potentially automate certain diagnostic tasks, especially in image analysis, the complexity of clinical decision-making and the ethical implications of automated medical decisions suggest that AI should primarily serve in a supportive role [23]. This aligns with the principle of keeping humans "in the loop" for critical healthcare decisions. The appropriate division of labor between AI systems and clinicians requires careful consideration of factors such as task complexity, potential consequences of errors, and the contextual understanding that human experts bring to patient care situations. There are also important ethical and regulatory considerations in deploying multimodal generative AI in healthcare. These include ensuring data privacy and security, addressing potential biases in AI models, maintaining transparency and interpretability of AI decision-making processes, and establishing clear accountability frameworks [17]. As these systems often operate as "black boxes," there is a pressing need for explainable AI approaches in healthcare applications. Regulatory bodies worldwide are working to develop guidelines that ensure multimodal

 doi:10.20944/preprints202508.1425.v1

AI systems meet rigorous standards for safety, effectiveness, and fairness while protecting patient privacy and autonomy in the decision-making process.

Looking ahead, the integration of multimodal generative AI in diagnostics holds immense promise but requires careful implementation. Strategies for effective trust calibration, clear delineation of AI's role in decision support versus automation, and robust ethical and regulatory frameworks are essential. The development of healthcare-specific evaluation metrics and validation methodologies will also be crucial to ensure these systems perform reliably across diverse patient populations and clinical settings. As the field evolves, ongoing research, real-world piloting, and iterative refinement of AI systems in clinical settings will be crucial to realizing the potential of this technology while prioritizing patient safety and care quality. The most successful implementations will likely be those that thoughtfully integrate multimodal AI capabilities into existing clinical workflows, providing seamless decision support while preserving the essential human elements of healthcare delivery.

6. Clinical Validation and Regulatory Perspectives

6.1. Regulatory and FDA Perspectives

Multimodal generative AI is emerging as a powerful tool in medical diagnostics, offering the potential to revolutionize clinical workflows and improve patient outcomes. However, its integration into healthcare systems raises important regulatory and ethical considerations that must be carefully addressed. The U.S. Food and Drug Administration (FDA) has recognized the potential impact of AI on human health and has adopted a framework for regulating AI as Software as a Medical Device (SaMD) [64]. Under this approach, AI products may be reviewed through existing medical device pathways such as 510(k), De Novo, or Premarket Approval (PMA), depending on the level of risk and novelty. To address the unique challenges posed by AI, particularly adaptive algorithms that can change in response to new data, the FDA has proposed several initiatives including the Software Pre-Cert Pilot Program, which focuses on vetting software developers and processes, allowing for streamlined review of some products and requiring ongoing real-world performance monitoring [64]. The FDA has also introduced the AI/ML Action Plan, which implements a total product lifecycle (TPLC) approach for regulating AI/ML-based SaMD [65]. This approach considers the adaptive nature of continuously learning algorithms and recommends new premarket submissions if software modifications introduce new risks or significantly alter functionality. Additionally, the FDA supports the use of Real-World Evidence (RWE) for machine learning validation, which allows developers to demonstrate ongoing accuracy and safety of AI models without requiring full re-approval for minor modifications [65]. For post-market monitoring, the FDA encourages manufacturers to establish transparency in AI model updates through structured risk assessments, labeling modifications, and regulatory submission triggers when significant changes occur. Developers must implement robust post-market monitoring strategies, including bias and fairness assessment tools, automated performance tracking mechanisms, and regulatory submission workflows for major algorithm modifications [65].

In the European Union, a complementary regulatory framework exists in the form of the AI Act, which provides a risk-based approach for AI systems [66]. Under this framework, AI applications in medical diagnostics are considered high-risk and must undergo rigorous third-party evaluations, adhere to strict data management standards, and implement risk mitigation measures. The Act also introduces new obligations for general-purpose AI models, including requirements for technical documentation, disclosure of training data sources, and compliance with EU copyright laws [54]. The integration of multimodal generative AI in diagnostics presents both opportunities and challenges within these regulatory frameworks. These systems have demonstrated promising results in tasks such as medical report generation, visual question answering, and cross-modal retrieval [25]. For example, in radiology, generative AI models could potentially match human expert performance in generating reports across various imaging modalities. However, the complexity of these systems also introduces new sources of potential errors, necessitating thorough error analysis and validation beyond what traditional regulatory approaches might address. Ethical considerations are paramount in the

 doi:10.20944/preprints202508.1425.v1

deployment of generative AI in healthcare and must be considered alongside regulatory compliance. A significant concern is the potential for AI to produce misleading or fabricated information, posing risks of misdiagnosis or inappropriate treatment recommendations [67]. Additionally, the lack of transparency in decision-making processes is problematic, as the closed-source nature of many large language models prevents both patients and healthcare providers from understanding the reasoning behind AI-generated outputs. Further ethical challenges include risks to patient privacy and data security, particularly concerning the use of patient-derived data and AI-generated synthetic data, and the potential depersonalization of care, as AI streamlines administrative tasks and potentially distances providers from patients [67].

To address these challenges and ensure the responsible integration of multimodal generative AI in diagnostics, several key steps are necessary. Regulatory frameworks must be further developed to adapt to the dynamic nature of AI technologies while maintaining high standards for safety and efficacy. This includes implementing rigorous validation processes, such as real-world testing and post-market surveillance, to assess the performance and impact of AI systems across diverse populations and clinical settings. Clear guidelines for transparency and explainability in AI-driven diagnostics must be established, enabling healthcare providers to understand and critically evaluate AI-generated insights [67]. Interdisciplinary research investment is crucial to advance methods for error analysis, bias mitigation, and fairness in multimodal AI systems. This research should inform both regulatory approaches and clinical implementation strategies. Finally, ongoing ethical deliberation involving clinicians, AI developers, policymakers, and patient advocates is essential to navigate the complex implications of AI integration in healthcare [64–66]. As multimodal generative AI continues to evolve, careful consideration of regulatory, clinical, and ethical factors will be essential to harness its potential for improving diagnostic accuracy and patient care while safeguarding against potential risks and unintended consequences. The convergence of robust regulatory oversight, ethical guidelines, and technological innovation will ultimately determine how effectively these powerful AI tools can be integrated into medical diagnostics to benefit patients and healthcare systems [25,67].

6.2. Ethical and Liability Challenges

Multimodal generative AI holds significant potential to transform diagnostics in healthcare, but also raises important ethical, legal, and regulatory challenges related to responsibility, fairness, transparency, and patient autonomy. This review examines key considerations in these areas. Responsibility for AI outputs in diagnostics is a critical concern as generative AI models become more sophisticated and autonomous. There are questions around who bears liability when an AI system makes an incorrect diagnosis or recommendation - the healthcare provider, the AI developer, or some combination [67]. Clear frameworks are needed to delineate responsibilities and establish accountability. Some propose a "human-in-the-loop" approach where AI serves as a decision support tool but final diagnostic decisions remain with human clinicians [68]. However, as AI capabilities grow, maintaining meaningful human oversight may become challenging, especially when clinicians may face cognitive biases leading them to over-rely on seemingly authoritative AI recommendations. Legal frameworks for AI liability in healthcare are still evolving. Existing medical malpractice laws may not adequately address scenarios involving AI-assisted diagnoses [69]. New legislation or case law may be needed to clarify liability in cases where AI recommendations contribute to adverse outcomes. Some experts suggest a "learned intermediary" doctrine, where clinicians maintain primary responsibility for diagnostic decisions but AI developers could be held liable for defects in their systems [68]. This approach attempts to balance accountability while recognizing the unique roles played by human clinicians and AI systems in the diagnostic process.

Fairness is another key issue, as AI models can perpetuate or amplify biases present in training data. Multimodal generative AI models that integrate diverse data types like imaging, genomics, and clinical records may be particularly susceptible to compounded biases [17]. Careful curation of diverse, representative training datasets is crucial for mitigating these risks. Researchers are exploring techniques like fairness-constrained learning and algorithmic auditing to mitigate unfairness

 doi:10.20944/preprints202508.1425.v1

in model outputs [70]. Ongoing monitoring and auditing of deployed models is also important to detect emerging biases that may affect different demographic groups, particularly those historically underrepresented in medical research and clinical trials. Transparency poses significant challenges for complex generative AI models used in diagnostics. The "black box" nature of deep learning models makes it difficult to interpret how they arrive at outputs [67]. This lack of interpretability can erode trust among clinicians and patients, potentially limiting adoption of these technologies despite their potential benefits. Efforts to develop more explainable AI models are ongoing, including visualization techniques to elucidate model decision-making [70]. For multimodal models, explaining how different data modalities are weighted and integrated presents an additional challenge that requires innovative approaches to model interpretability and communication of AI-derived insights. Patient autonomy and informed consent are critical ethical considerations that must be addressed as AI systems become more prevalent in diagnostics. Patients have a right to understand when AI is being used in their care and how it may influence diagnostic or treatment decisions [67]. Clear communication about the role of AI, its potential benefits and limitations, and alternatives is essential for maintaining patient trust and respecting their autonomy. Some experts argue that specific informed consent should be obtained for AI-assisted diagnoses, similar to consent for other medical procedures [69]. This raises practical questions about how to effectively communicate complex technical information about AI systems to patients with varying levels of technical literacy and health knowledge.

Regulatory frameworks are still evolving to keep pace with rapid advancements in generative AI for healthcare. Agencies like the FDA are working to develop appropriate oversight approaches that balance innovation with patient safety [68]. Key areas of focus include validating model performance, ensuring data privacy and security, and establishing standards for clinical deployment. The FDA's "Software as a Medical Device" (SaMD) framework may need to be adapted for AI systems that continuously learn and evolve [68]. The dynamic nature of learning AI systems presents unique challenges for traditional regulatory approaches that assume relatively stable product characteristics over time. Intellectual property concerns also arise with generative AI in healthcare. There are questions around the ownership and patentability of AI-generated diagnostic insights or treatment plans [17]. Clear guidelines are needed to ensure appropriate attribution and protect innovation while promoting knowledge sharing that benefits patients. The collaborative nature of healthcare innovation, involving multiple stakeholders including clinicians, researchers, patients, and technology developers, further complicates intellectual property considerations in this domain. Addressing these ethical, legal, and regulatory challenges will be crucial for realizing the full potential of multimodal generative AI in diagnostics. Ongoing collaboration between AI developers, healthcare providers, ethicists, legal experts, and policymakers is needed to develop responsible approaches that prioritize patient benefit while mitigating risks [71]. This includes establishing governance frameworks, refining liability and intellectual property laws, and creating standards for model development and deployment [22]. Importantly, these frameworks must be flexible enough to adapt to rapidly evolving technological capabilities while maintaining core principles of patient safety, autonomy, and equitable access to healthcare innovations.

With appropriate guardrails in place, generative AI has the potential to significantly enhance diagnostic capabilities and improve patient outcomes. However, maintaining human judgment and preserving the doctor-patient relationship will be essential as these technologies become more widespread. Striking the right balance between leveraging AI's strengths and preserving human medical expertise and compassion should be a key focus as the field continues to evolve. This will require ongoing dialogue among stakeholders, careful evaluation of AI systems in real-world clinical settings, and a commitment to centering patient needs and values in the development and deployment of these powerful new diagnostic tools.

6.3. Key Findings and Path Forward

Multimodal generative AI has made significant progress in clinical diagnostics, offering innovative approaches to enhance diagnostic accuracy, workflow efficiency, and research capabilities.

 doi:10.20944/preprints202508.1425.v1

However, several gaps remain and a future research agenda is needed to fully realize its potential. This review synthesizes key findings and outlines a path forward for multimodal generative AI in clinical diagnostics. Multimodal generative AI has demonstrated promising capabilities in integrating diverse data types to improve diagnostic processes. In the area of enhanced diagnostic accuracy, AI-driven image analysis techniques have shown potential in improving the detection and classification of diseases, particularly in fields like radiology and pathology. For example, multimodal models integrating imaging data with clinical records and genomic information have achieved higher accuracy in breast cancer detection compared to unimodal approaches [14,36,72]. Beyond diagnostics, generative AI models have been successfully applied to workflow optimization by automating routine tasks in healthcare administration and clinical documentation, potentially reducing physician burnout and improving efficiency. Large language models (LLMs) have been implemented to generate summaries of patient data from electronic health records, significantly reducing documentation time for healthcare providers [13]. These advancements represent a critical step toward addressing administrative burdens that often detract from patient care. Personalized medicine has also benefited from multimodal generative AI approaches. By analyzing multiple data modalities including genetic profiles, medical history, and lifestyle factors, these systems can predict individual patient responses to treatments, enabling more tailored therapeutic approaches [13]. This capability marks a paradigm shift from the traditional one-size-fits-all approach to more individualized patient care protocols.

In medical education and training, generative AI has created interactive tools such as virtual patient cases and simulations for cognitive behavioral therapy training [13]. These educational applications provide standardized yet customizable learning experiences that can supplement traditional medical education and help address training disparities across different healthcare settings. The research and development landscape has similarly been transformed, with multimodal generative AI showing promise in drug discovery and clinical trial design. By integrating diverse data sources, these systems help identify potential drug targets and optimize trial protocols, potentially accelerating the development of new therapeutics and improving trial success rates [72].

6.4. Remaining Gaps

Despite these advancements, significant challenges and gaps persist in the implementation of multimodal generative AI for clinical diagnostics. A primary concern is the lack of rigorous clinical validation in real-world settings, particularly for models utilizing multiple modalities [37]. The transition from promising research results to clinically validated tools requires large-scale prospective studies demonstrating efficacy and safety across diverse patient populations, which are currently insufficient in number and scope. Interpretability and explainability present another substantial hurdle. The "black box" nature of many complex multimodal AI models poses challenges for clinical adoption and regulatory approval [16]. Healthcare professionals are understandably reluctant to rely on diagnostic recommendations they cannot verify or understand, particularly when patient outcomes are at stake. Improving the interpretability of these models is therefore crucial for building trust among healthcare professionals and patients. Data quality and standardization issues further complicate implementation efforts. The performance of multimodal AI models is highly dependent on the quality and consistency of input data, yet healthcare data remains notoriously fragmented and heterogeneous. Inconsistencies in data collection, formatting, and labeling across different healthcare systems continue to impede the development of robust, generalizable AI models [19]. Ethical and privacy concerns also demand attention as the field advances. The use of large-scale, multimodal patient data raises important questions regarding data privacy, consent, and potential biases in AI algorithms [67]. These concerns are particularly acute when considering vulnerable populations or sensitive health information, necessitating careful consideration of both technical safeguards and ethical frameworks.

Finally, the integration of multimodal AI tools into existing clinical workflows and healthcare IT systems remains challenging. Even technically superior AI systems may fail to gain traction if they disrupt established workflows or create additional burdens for healthcare providers. Achieving

 doi:10.20944/preprints202508.1425.v1

seamless integration requires careful consideration of user experience, interoperability standards, and the complex sociotechnical environments in which these tools will operate [16].

6.5. Future Research Agenda

To address these gaps and advance the field of multimodal generative AI in clinical diagnostics, a comprehensive research agenda is essential. First and foremost, there is a critical need to conduct large-scale, multicenter clinical trials to validate the performance and clinical impact of multimodal AI diagnostic tools across diverse patient populations and healthcare settings. These studies should assess not only technical performance metrics but also clinically relevant outcomes and potential implementation challenges. Parallel efforts must focus on developing novel techniques to improve the interpretability and explainability of complex multimodal AI models. By enabling clinicians to understand and trust the reasoning behind AI-generated recommendations, these advances would address a major barrier to clinical adoption. Approaches such as attention visualization, feature importance ranking, and natural language explanations warrant further investigation in the context of multimodal systems. Standardized protocols for data collection, preprocessing, and annotation across different modalities are equally important to ensure consistency and improve the generalizability of AI models. Collaborative initiatives involving healthcare institutions, professional societies, and regulatory bodies could establish common data standards and quality assurance practices specifically tailored to multimodal clinical data.

As multimodal AI systems increasingly influence clinical decisions, investigating methods to ensure fairness and mitigate biases becomes imperative. Particular attention should be given to underrepresented patient populations, who may be disadvantaged by algorithms trained predominantly on majority group data. Bias detection tools, diverse training datasets, and fairness-aware learning algorithms represent promising research directions. Privacy concerns can be addressed through research into federated learning and other privacy-preserving techniques that enable collaborative model development while protecting patient privacy and data security. These approaches allow AI models to learn from decentralized data sources without requiring sensitive information to be shared or centralized, potentially resolving the tension between data access and privacy protection. To facilitate clinical adoption, researchers should design and evaluate user-centered interfaces and integration strategies that seamlessly incorporate multimodal AI tools into clinical workflows. This human-centered design approach should maximize utility and adoption by considering the needs, preferences, and constraints of healthcare providers operating in complex clinical environments.

Understanding the long-term impact of multimodal AI adoption requires longitudinal studies investigating effects on clinical outcomes, healthcare costs, and physician workflow. These studies should be alert to potential unintended consequences, such as automation bias or deskilling, which might accompany widespread AI implementation in clinical practice. The regulatory landscape must evolve alongside the technology, necessitating research to develop regulatory frameworks and guidelines specifically tailored to the evaluation and approval of multimodal generative AI tools in healthcare settings. These frameworks should balance the need for innovation with patient safety considerations and provide clear pathways for the validation and approval of increasingly complex AI systems. Emerging diagnostic modalities present fertile ground for multimodal AI applications. Researchers should explore the potential of these approaches in areas such as digital pathology, liquid biopsy analysis, and integrative diagnostics combining molecular and imaging data. These novel applications could expand diagnostic capabilities and enable earlier, more precise disease detection. Finally, the chronic challenge of limited data for model training could be addressed through research into synthetic data generation techniques. These methods could augment limited datasets and improve model performance, particularly for rare diseases or underrepresented patient groups where collecting sufficient real-world data is especially challenging.

In conclusion, while multimodal generative AI has shown great promise in enhancing clinical diagnostics, significant work remains to be done to address existing gaps and realize its full potential. A concerted effort from researchers, clinicians, regulators, and industry partners will be crucial in

 doi:10.20944/preprints202508.1425.v1

advancing this technology and ensuring its safe and effective integration into healthcare systems. By pursuing the research agenda outlined above, the field can move toward more robust, interpretable, and clinically validated multimodal AI systems that genuinely improve patient care.

7. Conclusions

Multimodal generative AI is poised to revolutionize diagnostics and clinical trials, offering unprecedented opportunities for improving patient care while also presenting new challenges that must be carefully addressed. As we look to the future, several key priorities emerge for clinical practice and research. A major priority is developing robust frameworks for seamlessly integrating multiple types of patient data, including imaging, genomics, electronic health records, and real-time physiological monitoring. This multimodal approach allows for a more comprehensive understanding of patient health and disease progression, enabling more accurate diagnostics and personalized treatment plans. Future research should focus on optimizing data fusion techniques and creating standardized protocols for multimodal data collection and analysis to ensure consistency across clinical settings [15].

As AI models become increasingly complex, ensuring their decisions are interpretable and explainable to both clinicians and patients is crucial for clinical implementation. Research efforts should concentrate on developing techniques that provide clear rationales for AI-generated diagnoses and recommendations, such as attention-based visualizations and feature attribution methods. This transparency is essential for building trust among healthcare professionals and facilitating widespread adoption in clinical settings [39]. Advancing personalized medicine requires AI systems that can adapt to individual patient characteristics and evolving clinical knowledge. The development of self-supervised and reinforcement learning approaches that continuously refine diagnostic and treatment recommendations based on patient outcomes and physician feedback represents a significant opportunity for improving clinical trial design and execution. These adaptive systems could dramatically reduce the time and resources needed for patient stratification and treatment optimization [45]. As AI systems play an increasingly significant role in healthcare decision-making, addressing ethical concerns and mitigating potential biases is paramount to ensuring equitable care. Research priorities should include developing robust frameworks for evaluating and mitigating biases in training data and model outputs, particularly when these systems are deployed across diverse patient populations. Establishing clear guidelines for the ethical use of AI in clinical practice will be essential for maintaining patient trust and ensuring regulatory compliance [23]. There is a pressing need for large-scale validation studies and standardized benchmarks to rigorously assess the performance and generalizability of multimodal generative AI models across diverse clinical settings. This includes developing comprehensive evaluation frameworks that consider not only diagnostic accuracy but also clinical relevance, patient outcomes, and potential unintended consequences. Such validation efforts are crucial for transitioning promising research applications into standard clinical care [59]. As multimodal AI systems require access to sensitive patient data, advancing privacy-preserving machine learning techniques is essential for maintaining confidentiality while enabling innovation. Research should focus on developing methods for federated learning, differential privacy, and secure multi-party computation that enable collaborative model training and deployment while safeguarding patient information. These approaches will be particularly important for multi-center clinical trials and international research collaborations [22].

While many current applications focus on specific medical specialties, future research should explore the development of AI systems that can operate across multiple clinical domains. This cross-specialty approach could lead to more holistic patient care and facilitate the discovery of novel inter-disciplinary insights that might otherwise remain undetected. Integration of knowledge across traditionally siloed medical specialties may reveal new biomarkers and treatment approaches [17]. Integrating multimodal generative AI into real-time clinical workflows presents both technical and organizational challenges that must be addressed for successful implementation. Priorities include

 doi:10.20944/preprints202508.1425.v1

developing efficient algorithms for processing and analyzing streaming multimodal data, as well as designing user interfaces that seamlessly incorporate AI-generated insights into clinical decision-making processes without disrupting established workflows or adding to clinician burden [45]. Advancing techniques for generating high-quality synthetic medical data could help address data scarcity issues and enable more robust model training, particularly for rare conditions or underrepresented patient populations. Research should focus on improving the fidelity and clinical relevance of synthetic data across multiple modalities while ensuring patient privacy is maintained. These synthetic datasets could also serve as standardized benchmarks for comparing different AI approaches [19]. By addressing these clinical and research priorities, the medical community can harness the full potential of multimodal generative AI to transform diagnostics and clinical trials, ultimately leading to improved patient outcomes and more efficient healthcare delivery. However, this progress must be balanced with careful consideration of ethical implications, regulatory compliance, and the preservation of the human element in healthcare. The successful integration of these technologies will require close collaboration between AI researchers, clinicians, patients, and regulatory bodies to ensure that advances in multimodal generative AI truly benefit those they are designed to serve.

Author Contributions: Conceptualization, M.M.; methodology, M.M.; software, M.M.; validation, M.M.; formal analysis, M.M.; investigation, M.M.; resources, S.A.G.; data curation, M.M.; writing—original draft preparation, M.M.; writing—review and editing, M.M. and S.A.G.; supervision, M.M.; project administration, M.M. and S.A.G.; All authors have read and agreed to the published version of the manuscript.

Funding: No funding was obtained for this research.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: All the references used in this research review have been obtained from publicly available PubMed research repository.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
LLM	Large Language Model
VQA	Visual Question Answering
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
CT	Computed Tomography
XAI	Explainable Artificial Intelligence
Grad-CAM	Gradient-weighted Class Activation Mapping
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-agnostic Explanations
Med-PaLM	Medical Pathways Language Model
LLaVA-Med	Large Language and Vision Assistant for Medicine
BiomedGPT	Biomedical Generative Pre-trained Transformer
BioGPT-ViT	BioGPT Vision Transformer
MLG-GAN	Multi-Level Guided Generative Adversarial Network
Mul-T	Multimodal Transformer
ECG	Electrocardiogram
EEG	Electroencephalogram
FDG-PET	Fluorodeoxyglucose Positron Emission Tomography
GANs	Generative Adversarial Networks

 doi:10.20944/preprints202508.1425.v1

ROCO	Radiology Objects in COntext
NIH14	National Institutes of Health Chest X-ray Dataset
QA	Question Answering
OCT	Optical Coherence Tomography
MIMIC	Medical Information Mart for Intensive Care
CLARO	CT Imaging of Lung Cancer And Related Outcomes
LCID	Lung Cancer Imaging Database
EHR	Electronic Health Records
AUC-ROC	Area Under the Receiver Operating Characteristic curve
BLEU	Bilingual Evaluation Understudy
MIMIC-CXR	Medical Information Mart for Intensive Care Chest X-Ray
ImageCLEF	Image Cross-Language Evaluation Forum
MMBERT	Multi-Modal Bidirectional Encoder Representations from Transformers
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
CONCH	Contrastive Learning from Captions for Histopathology
IHC	Immunohistochemistry
H&E	Hematoxylin and Eosin
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
Onto-CGAN	Ontology-enhanced Conditional Generative Adversarial Network
AML	Acute Myeloid Leukemia
TPLC	Total Product Life Cycle
PMA	Premarket Approval
SaMD	Software as a Medical Device
FDA	Food and Drug Administration

References

1. Harrer, S.; Shah, P.; Antony, B.; Hu, J. Artificial intelligence for clinical trial design. *Trends in pharmacological sciences* **2019**, *40*, 577–591.
2. Maleki, M.; Ghahari, S. Clinical trials protocol authoring using llms. *arXiv preprint arXiv:2404.05044* **2024**.
3. Askin, S.; Burkhalter, D.; Calado, G.; El Dakrouni, S. Artificial intelligence applied to clinical trials: opportunities and challenges. *Health and technology* **2023**, *13*, 203–213.
4. Maleki, M.; Ghahari, S. Comprehensive clustering analysis and profiling of covid-19 vaccine hesitancy and related factors across us counties: Insights for future pandemic responses. In Proceedings of the Healthcare. MDPI, 2024, Vol. 12, p. 1458.
5. Maleki, M.; Khan, M. Covid-19 health equity & justice dashboard: A step towards countering health disparities among seniors and minority population. *Available at SSRN 4595845* **2023**.
6. Mayorga-Ruiz, I.; Jiménez-Pastor, A.; Fos-Guarinos, B.; López-González, R.; García-Castro, F.; Alberich-Bayarri, Á. The role of AI in clinical trials. *Artificial Intelligence in Medical Imaging: Opportunities, applications and risks* **2019**, pp. 231–243.
7. Maleki, M. Clustering analysis of us covid-19 rates, vaccine participation, and socioeconomic factors. *arXiv preprint arXiv:2404.08186* **2024**.
8. Maleki, M.; Haeri, F. Identification of cardiovascular diseases through ECG classification using wavelet transformation. *arXiv preprint arXiv:2404.09393* **2024**.
9. Woo, M. An AI boost for clinical trials. *Nature* **2019**, *573*, S100–S100.
10. Maleki, M.; Ghahari, S. Impact of Major Health Events on Pharmaceutical Stocks: A Comprehensive Analysis Using Macroeconomic and Market Indicators. *arXiv preprint arXiv:2408.01883* **2024**.
11. Maleki, M.; Bahrami, M.; Menendez, M.; Balsa-Barreiro, J. Social Behavior and COVID-19: Analysis of the Social Factors behind Compliance with Interventions across the United States. *International Journal of Environmental Research and Public Health* **2022**, *19*. https://doi.org/10.3390/ijerph192315716.
12. Angus, D.C. Randomized clinical trials of artificial intelligence. *Jama* **2020**, *323*, 1043–1045.
13. Reddy, S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. *Implementation science : IS* **2024**, *19*, 27. https://doi.org/10.1186/s13012-024-01357-9.

 doi:10.20944/preprints202508.1425.v1

14. Abdullakutty, F.; Akbari, Y.; Al-Maadeed, S.; Bouridane, A. Histopathology in focus: a review on explainable multi-modal approaches for breast cancer diagnosis. *Frontiers in medicine* **2024**, *11*, 1450103. https://doi.org/10.3389/fmed.2024.1450103.
15. Haq, I.U.; Mhamed, M.; Al-Harbi, M.; Osman, H. Advancements in Medical Radiology Through Multimodal Machine Learning: A Comprehensive Overview. *Bioengineering (Basel, Switzerland)* **2025**, *12*. https://doi.org/10.3390/bioengineering12050477.
16. Rouzrokh, P.; Khosravi, B.; Faghani, S.; Moassefi, M. A Current Review of Generative AI in Medicine: Core Concepts, Applications, and Current Limitations. *Current reviews in musculoskeletal medicine* **2025**, *18*, 246–266. https://doi.org/10.1007/s12178-025-09961-y.
17. Rashidi, H.H.; Pantanowitz, J.; Chamanzar, A.; Fennell, B. Generative Artificial Intelligence in Pathology and Medicine: A Deeper Dive. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* **2025**, *38*, 100687. https://doi.org/10.1016/j.modpat.2024.100687.
18. Araújo, C.C.; Frias, J.; Mendes, F.; Martins, M. Unlocking the Potential of AI in EUS and ERCP: A Narrative Review for Pancreaticobiliary Disease. *Cancers* **2025**, *17*. https://doi.org/10.3390/cancers17071132.
19. Gao, X.; Shi, F.; Shen, D.; Liu, M. Multimodal transformer network for incomplete image generation and diagnosis of Alzheimer's disease. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* **2023**, *110*, 102303. https://doi.org/10.1016/j.compmedimag.2023.102303.
20. Kunze, K.N. Generative Artificial Intelligence and Musculoskeletal Health Care. *HSS journal : the musculoskeletal journal of Hospital for Special Surgery* **2025**, *21*, 15563316251335334. https://doi.org/10.1177/15563316251335334.
21. Oettl, F.C.; Zsidai, B.; Oeding, J.F.; Hirschmann, M.T. Beyond traditional orthopaedic data analysis: AI, multimodal models and continuous monitoring. *Knee surgery, sports traumatology, arthroscopy : official journal of the ESSKA* **2025**, *33*, 2269–2275. https://doi.org/10.1002/ksa.12657.
22. Parvin, N.; Joo, S.W.; Jung, J.H.; Mandal, T.K. Multimodal AI in Biomedicine: Pioneering the Future of Biomaterials, Diagnostics, and Personalized Healthcare. *Nanomaterials (Basel, Switzerland)* **2025**, *15*. https://doi.org/10.3390/nano15120895.
23. Tortora, L. Beyond Discrimination: Generative AI Applications and Ethical Challenges in Forensic Psychiatry. *Frontiers in psychiatry* **2024**, *15*, 1346059. https://doi.org/10.3389/fpsyt.2024.1346059.
24. Gao, Y.; Wen, P.; Liu, Y.; Sun, Y. Application of artificial intelligence in the diagnosis of malignant digestive tract tumors: focusing on opportunities and challenges in endoscopy and pathology. *Journal of translational medicine* **2025**, *23*, 412. https://doi.org/10.1186/s12967-025-06428-z.
25. Rao, V.M.; Hla, M.; Moor, M.; Adithan, S. Multimodal generative AI for medical image interpretation. *Nature* **2025**, *639*, 888–896. https://doi.org/10.1038/s41586-025-08675-y.
26. Han, T.; Jeong, W.K.; Shin, J. Diagnostic performance of multimodal large language models in radiological quiz cases: the effects of prompt engineering and input conditions. *Ultrasonography (Seoul, Korea)* **2025**, *44*, 220–231. https://doi.org/10.14366/usg.25012.
27. Shao, J.; Ma, J.; Zhang, Q.; Li, W. Predicting gene mutation status via artificial intelligence technologies based on multimodal integration (MMI) to advance precision oncology. *Seminars in cancer biology* **2023**, *91*, 1–15. https://doi.org/10.1016/j.semcancer.2023.02.006.
28. Pfob, A.; Sidey-Gibbons, C.; Barr, R.G.; Duda, V. The importance of multi-modal imaging and clinical information for humans and AI-based algorithms to classify breast masses (INSPiRED 003): an international, multicenter analysis. *European radiology* **2022**, *32*, 4101–4115. https://doi.org/10.1007/s00330-021-08519-z.
29. Ullah, E.; Baig, M.M.; Waqas, A.; Rasool, G. Multimodal Generative AI for Anatomic Pathology-A Review of Current Applications to Envisage the Future Direction. *Advances in anatomic pathology* **2025**. https://doi.org/10.1097/PAP.0000000000000498.
30. Jain, S.S.; Elias, P.; Poterucha, T.; Randazzo, M. Artificial Intelligence in Cardiovascular Care-Part 2: Applications: JACC Review Topic of the Week. *Journal of the American College of Cardiology* **2024**, *83*, 2487–2496. https://doi.org/10.1016/j.jacc.2024.03.401.
31. Hagos, D.H.; Aryal, S.K.; Ymele-Leki, P.; Burge, L.L. AI-driven multimodal colorimetric analytics for biomedical and behavioral health diagnostics. *Computational and structural biotechnology journal* **2025**, *27*, 2219–2232. https://doi.org/10.1016/j.csbj.2025.05.015.
32. Javan, R.; Kim, T.; Mostaghni, N. GPT-4 Vision: Multi-Modal Evolution of ChatGPT and Potential Role in Radiology. *Cureus* **2024**, *16*, e68298. https://doi.org/10.7759/cureus.68298.

 doi:10.20944/preprints202508.1425.v1

33. Geersing, G.J.; de Wit, N.J.; Thompson, M. Generative artificial intelligence for general practice; new potential ahead, but are we ready? *The European journal of general practice* **2025**, *31*, 2511645. https://doi.org/10.1080/13814788.2025.2511645.
34. Maleki, M. Advancing Healthcare Accessibility through a Neighborhood Search Recommendation Tool. *Available at SSRN 4825773* **2024**.
35. Maleki, M. Evaluating the Reproducibility of ICU Patient Readmission using RNN and ODE models. *Available at SSRN 4825763* **2024**.
36. Brodsky, V.; Ullah, E.; Bychkov, A.; Song, A.H. Generative Artificial Intelligence in Anatomic Pathology. *Archives of pathology & laboratory medicine* **2025**, *149*, 298–318. https://doi.org/10.5858/arpa.2024-0215-RA.
37. Hong, E.K.; Ham, J.; Roh, B.; Gu, J. Diagnostic Accuracy and Clinical Value of a Domain-specific Multimodal Generative AI Model for Chest Radiograph Report Generation. *Radiology* **2025**, *314*, e241476. https://doi.org/10.1148/radiol.241476.
38. Hong, G.S.; Jang, M.; Kyung, S.; Cho, K. Overcoming the Challenges in the Development and Implementation of Artificial Intelligence in Radiology: A Comprehensive Review of Solutions Beyond Supervised Learning. *Korean journal of radiology* **2023**, *24*, 1061–1080. https://doi.org/10.3348/kjr.2023.0393.
39. Chang, C.; Shi, W.; Wang, Y.; Zhang, Z. The path from task-specific to general purpose artificial intelligence for medical diagnostics: A bibliometric analysis. *Computers in biology and medicine* **2024**, *172*, 108258. https://doi.org/10.1016/j.compbiomed.2024.108258.
40. Sonmez, S.C.; Sevgi, M.; Antaki, F.; Huemer, J. Generative artificial intelligence in ophthalmology: current innovations, future applications and challenges. *The British journal of ophthalmology* **2024**, *108*, 1335–1340. https://doi.org/10.1136/bjo-2024-325458.
41. Zhang, X.; Wu, C.; Zhao, Z.; Lin, W. Development of a large-scale medical visual question-answering dataset. *Communications medicine* **2024**, *4*, 277. https://doi.org/10.1038/s43856-024-00709-2.
42. Sorin, V.; Barash, Y.; Konen, E.; Klang, E. Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) - A Systematic Review. *Academic radiology* **2020**, *27*, 1175–1185. https://doi.org/10.1016/j.acra.2019.12.024.
43. Alajaji, S.A.; Khoury, Z.H.; Elgharib, M.; Saeed, M. Generative Adversarial Networks in Digital Histopathology: Current Applications, Limitations, Ethical Considerations, and Future Directions. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* **2024**, *37*, 100369. https://doi.org/10.1016/j.modpat.2023.100369.
44. Panagoulias, D.P.; Tsoureli-Nikita, E.; Virvou, M.; Tsihrintzis, G.A. Dermacen analytica: A novel methodology integrating multi-modal large language models with machine learning in dermatology. *International journal of medical informatics* **2025**, *199*, 105898. https://doi.org/10.1016/j.ijmedinf.2025.105898.
45. Algarni, A. CareAssist GPT improves patient user experience with a patient centered approach to computer aided diagnosis. *Scientific reports* **2025**, *15*, 22727. https://doi.org/10.1038/s41598-025-01518-w.
46. Teoh, J.R.; Dong, J.; Zuo, X.; Lai, K.W. Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications. *PeerJ. Computer science* **2024**, *10*, e2298. https://doi.org/10.7717/peerj-cs.2298.
47. Brin, D.; Sorin, V.; Barash, Y.; Konen, E. Assessing GPT-4 multimodal performance in radiological image analysis. *European radiology* **2025**, *35*, 1959–1965. https://doi.org/10.1007/s00330-024-11035-5.
48. Sosna, J.; Joskowicz, L.; Saban, M. Navigating the AI Landscape in Medical Imaging: A Critical Analysis of Technologies, Implementation, and Implications. *Radiology* **2025**, *315*, e240982. https://doi.org/10.1148/radiol.240982.
49. Hacking, S. Foundation models in pathology: bridging AI innovation and clinical practice. *Journal of clinical pathology* **2025**, *78*, 433–435. https://doi.org/10.1136/jcp-2024-209910.
50. Lu, M.Y.; Chen, B.; Williamson, D.F.K.; Chen, R.J. A multimodal generative AI copilot for human pathology. *Nature* **2024**, *634*, 466–473. https://doi.org/10.1038/s41586-024-07618-3.
51. Ferber, D.; El Nahhas, O.S.M.; Wölflein, G.; Wiest, I.C. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature cancer* **2025**. https://doi.org/10.1038/s43018-025-00991-6.
52. Lee, S.; Youn, J.; Kim, H.; Kim, M. CXR-LLaVA: a multimodal large language model for interpreting chest X-ray images. *European radiology* **2025**, *35*, 4374–4386. https://doi.org/10.1007/s00330-024-11339-6.
53. Gupta, A.; Rajamohan, N.; Bansal, B.; Chaudhri, S. Applications of artificial intelligence in abdominal imaging. *Abdominal radiology (New York)* **2025**. https://doi.org/10.1007/s00261-025-04990-0.

 doi:10.20944/preprints202508.1425.v1

54. Van Booven, D.J.; Chen, C.B.; Malpani, S.; Mirzabeigi, Y. Synthetic Genitourinary Image Synthesis via Generative Adversarial Networks: Enhancing Artificial Intelligence Diagnostic Precision. *Journal of personalized medicine* **2024**, *14*. https://doi.org/10.3390/jpm14070703.
55. Sun, C.; Dumontier, M. Generating unseen diseases patient data using ontology enhanced generative adversarial networks. *NPJ digital medicine* **2025**, *8*, 4. https://doi.org/10.1038/s41746-024-01421-0.
56. Yang, Y.; Shen, H.; Chen, K.; Li, X. From pixels to patients: the evolution and future of deep learning in cancer diagnostics. *Trends in molecular medicine* **2025**, *31*, 548–558. https://doi.org/10.1016/j.molmed.2024.11.009.
57. Segal, B.; Rubin, D.M.; Rubin, G.; Pantanowitz, A. Evaluating the Clinical Realism of Synthetic Chest X-Rays Generated Using Progressively Growing GANs. *SN computer science* **2021**, *2*, 321. https://doi.org/10.1007/s42979-021-00720-7.
58. Eckardt, J.N.; Hahn, W.; Röllig, C.; Stasik, S. Mimicking clinical trials with synthetic acute myeloid leukemia patients using generative artificial intelligence. *NPJ digital medicine* **2024**, *7*, 76. https://doi.org/10.1038/s41746-024-01076-x.
59. Ibrahim, M.; Khalil, Y.A.; Amirrajab, S.; Sun, C. Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. *Computers in biology and medicine* **2025**, *189*, 109834. https://doi.org/10.1016/j.compbiomed.2025.109834.
60. Liu, F.; Zhou, H.; Wang, K.; Yu, Y. MetaGP: A generative foundation model integrating electronic health records and multimodal imaging for addressing unmet clinical needs. *Cell reports. Medicine* **2025**, *6*, 102056. https://doi.org/10.1016/j.xcrm.2025.102056.
61. Hirosawa, T.; Harada, Y.; Tokumasu, K.; Ito, T. Evaluating ChatGPT-4's Diagnostic Accuracy: Impact of Visual Data Integration. *JMIR medical informatics* **2024**, *12*, e55627. https://doi.org/10.2196/55627.
62. Hosny, A.; Bitterman, D.S.; Guthier, C.V.; Qian, J.M. Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study. *The Lancet. Digital health* **2022**, *4*, e657–e666. https://doi.org/10.1016/S2589-7500(22)00129-7.
63. Sakamoto, T.; Harada, Y.; Shimizu, T. Facilitating Trust Calibration in Artificial Intelligence-Driven Diagnostic Decision Support Systems for Determining Physicians' Diagnostic Accuracy: Quasi-Experimental Study. *JMIR formative research* **2024**, *8*, e58666. https://doi.org/10.2196/58666.
64. Potnis, K.C.; Ross, J.S.; Aneja, S.; Gross, C.P. Artificial Intelligence in Breast Cancer Screening: Evaluation of FDA Device Regulation and Future Recommendations. *JAMA internal medicine* **2022**, *182*, 1306–1312. https://doi.org/10.1001/jamainternmed.2022.4969.
65. Han, G.R.; Goncharov, A.; Eryilmaz, M.; Ye, S. Machine learning in point-of-care testing: innovations, challenges, and opportunities. *Nature communications* **2025**, *16*, 3165. https://doi.org/10.1038/s41467-025-58527-6.
66. Ratkevičiūtė, K.; Aliukonis, V. Exploring Opportunities and Challenges of AI in Primary Healthcare: A Qualitative Study with Family Doctors in Lithuania. *Healthcare (Basel, Switzerland)* **2025**, *13*. https://doi.org/10.3390/healthcare13121429.
67. Hasan, S.S.; Fury, M.S.; Woo, J.J.; Kunze, K.N. Ethical Application of Generative Artificial Intelligence in Medicine. *Arthroscopy : the journal of arthroscopic & related surgery : official publication of the Arthroscopy Association of North America and the International Arthroscopy Association* **2025**, *41*, 874–885. https://doi.org/10.1016/j.arthro.2024.12.011.
68. Kumar, R.; Waisberg, E.; Ong, J.; Paladugu, P. Artificial Intelligence-Based Methodologies for Early Diagnostic Precision and Personalized Therapeutic Strategies in Neuro-Ophthalmic and Neurodegenerative Pathologies. *Brain sciences* **2024**, *14*. https://doi.org/10.3390/brainsci14121266.
69. Sablone, S.; Bellino, M.; Cardinale, A.N.; Esposito, M. Artificial intelligence in healthcare: an Italian perspective on ethical and medico-legal implications. *Frontiers in medicine* **2024**, *11*, 1343456. https://doi.org/10.3389/fmed.2024.1343456.
70. Jha, D.; Durak, G.; Das, A.; Sanjotra, J. Ethical framework for responsible foundational models in medical imaging. *Frontiers in medicine* **2025**, *12*, 1544501. https://doi.org/10.3389/fmed.2025.1544501.
71. Huang, J.; Wittbrodt, M.T.; Teague, C.N.; Karl, E. Efficiency and Quality of Generative AI-Assisted Radiograph Reporting. *JAMA network open* **2025**, *8*, e2513921. https://doi.org/10.1001/jamanetworkopen.2025.13921.
72. Lipkova, J.; Chen, R.J.; Chen, B.; Lu, M.Y. Artificial intelligence for multimodal data integration in oncology. *Cancer cell* **2022**, *40*, 1095–1110. https://doi.org/10.1016/j.ccell.2022.09.012.

 doi:10.20944/preprints202508.1425.v1