© October 2025 | IJIRT | Volume 12 Issue 5 | ISSN: 2349-6002

A Multimodal Medical AI Model for Clinical Imaging and Question Answering (Med Vision AI)

Siddharth Padwal[1], Vivek Sonone[2], Omkar Bangar[3], Dr. Puja Gholap[4], Dr.S.S. Khatal[5]
[1,2,3]Student, Department of Computer Engineering, Sharadchandra Pawar College of Engineering and Technology, Junnar, Pune, Maharashtra, India
[4,5]Internal Guide, Department of Computer Engineering, Sharadchandra Pawar College of Engineering and Technology, Junnar, Pune, Maharashtra, India

***Abstract*—Artificial Intelligence (AI) is fundamentally transforming healthcare, particularly in the field of medical diagnostics that involves clinical imaging and natural language processing. A primary challenge with existing AI models is their unimodal focus, concentrating on either image analysis or text comprehension. This specialization limits their utility in clinical settings that require a comprehensive, context-aware understanding of patient data. This paper introduces "A Multimodal Medical AI Model for Clinical Imaging and Question Answering (QA)," a project that confronts this limitation by developing an integrated AI framework. This framework is engineered to simultaneously interpret and reason over both visual and textual medical data. Our methodology employs sophisticated transformer-based architectures to process medical imagery and associated clinical reports. These distinct data streams are integrated using a multimodal fusion technique that harmonizes the features from each source. The resulting unified representation allows the system to handle intricate medical questions formulated in natural language, thereby emulating the diagnostic thought process of a clinical expert. To validate its real-world applicability, the model undergoes training and validation using reputable public datasets, including MIMIC-CXR and MedQA. A user-friendly interface is also developed to allow healthcare professionals to upload medical images and submit diagnostic inquiries. By leveraging the synergy between medical imaging, NLP, and deep learning, this initiative aims to elevate diagnostic accuracy, lessen the workload on medical staff, and broaden access to advanced medical knowledge, especially in underserved regions. This holistic approach signifies a major leap forward in the creation of more intelligent and effective healthcare technologies.**

***Index Terms*—Clinical Decision Support, Deep Learning, Medical Imaging, Multimodal AI, Natural Language Processing, Transformer Models, Visual Question Answering (VQA).**

I. INTRODUCTION

The intersection of artificial intelligence and healthcare has catalyzed a new wave of medical innovation, particularly within diagnostics and clinical imaging. Contemporary medical data is intrinsically multimodal, consisting of visual assets like X-rays, CT scans, and MRIs, in addition to unstructured textual information from physicians' notes and lab results. Legacy AI systems generally handle these data modalities separately, a practice that restricts their capacity to deliver holistic insights within a clinical framework. This research puts forth a consolidated AI model designed to operate on both visual and textual inputs for high-stakes clinical tasks such as providing diagnostic support and answering questions. The urgency for this technology is amplified by the growing prevalence of various diseases and the intensified pressures on healthcare practitioners. The model is especially valuable in remote or developing areas where access to specialized medical knowledge is constrained. Its scope of application includes radiology, general clinical diagnostics, medical education, and patient care management systems.

II. PROBLEM STATEMENT

In the clinical environment, healthcare professionals face immense pressure to analyze large volumes of heterogeneous data with high speed and accuracy. A persistent challenge in patient care is the rate of diagnostic mistakes that originate from the flawed interpretation of imaging or textual reports, which can negatively affect patient health outcomes. Current AI solutions do not adequately capitalize on the synergistic potential between visual medical data and its corresponding textual context, which curtails their usefulness in complex, real-world medical scenarios. The principal difficulty is in engineering an AI system that can simultaneously process and understand both images and text in a contextually meaningful way to provide precise and useful feedback. Furthermore, for any technology deployed in a critical sector like medicine, the transparency and dependability of its conclusions are of utmost importance.

III. RELATED WORK

A. Literature Survey

This project leverages recent progress in the domain of medical vision-language models. BioGPT-VL (2024) expands upon the BioGPT framework by integrating a visual component, using a vision transformer to process medical imagery and achieving notable results in tasks like visual question answering and image-text retrieval[1]. Med-PaLM (2023) is a prominent large language model tailored for medical QA datasets[2]. It has reached expert-level proficiency on key benchmarks such as MedQA, though its capabilities are limited to textual data without image analysis[3]. MedGemma (2024) functions as an adaptable vision-language model, trained on vast collections of biomedical images and text[4]. It adeptly manages diverse tasks, from generating image descriptions to answering questions and identifying diseases, showing powerful zero-shot learning performance[5].

B. Existing Systems

Contemporary systems are often specialized for a single data type[6]:

- CheXNet is a well-known deep learning model utilizing the DenseNet-121 architecture to detect pneumonia from chest radiographs, but its functionality is confined to image classification[7].
- BioGPT is a transformer-based model focused on text, having been pre-trained on an extensive collection of biomedical publications[8]. It shows excellence in textual analysis but has no capacity to process visual information[9].
- Med-PaLM demonstrates strong capabilities on medical text benchmarks but is unable to interpret radiological scans or other forms of visual data[10].
- Generalist models like BLIP and Flamingo can produce captions for medical images, but they often do not achieve the level of domain-specific precision necessary for clinical use[11].

IV. PROPOSED SYSTEM

Our proposal outlines a multimodal AI system that merges visual feature extraction from medical images with the semantic interpretation of clinical texts to address queries provided in natural language[12]. This system is designed with adaptability, interpretability, and robustness in mind for a wide array of diagnostic applications, achieved by enabling combined multimodal reasoning and interactive QA[13].

A. Objectives

The main objectives of this undertaking are:

- To construct an AI model capable of interpreting and analyzing both medical images and their corresponding textual data[14].
- To establish a question-answering function that draws upon integrated visual and textual clinical information[15].
- To provide assistance to medical practitioners in the identification of anomalies in medical imaging[16].
- To decrease the time required for diagnosis while increasing the accuracy of clinical evaluations[17].
- To unify advanced NLP and computer vision techniques within a singular multimodal system[18].
- To bolster the transparency and interpretability of the AI's operational decisions[19].
- To develop an intuitive user interface that simplifies interaction with the AI[20].

B. System Architecture

The system is built on a modular design where user engagement starts at an interface for uploading an image and submitting text[21]. These inputs are processed by parallel encoders (Image and Text)[22]. A Fusion Layer then synthesizes the features from both encoders, and the final answer is generated by a QA Output module[23].

- Frontend: A user-centric interface designed for image and query submission[24].
- Backend: Responsible for AI model serving, all data preprocessing tasks, and API management[25].
- Database: Used for storing medical questions, clinical reports, and the outputs generated by the model[26].

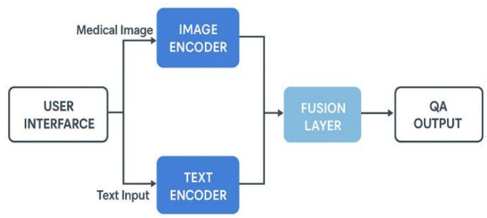

Fig. 1 System Diagram [27]

C. Module Descriptions

- Image Preprocessing Module: This module standardizes incoming medical images by managing format conversions, normalization, and their preparation into tensors[28].
- Text Preprocessing Module: It prepares all clinical text and user-submitted questions for model ingestion through cleaning, tokenization, and encoding procedures[29].
- Vision Encoder Module: This component uses a pretrained transformer to extract meaningful feature representations from the provided image[30].
- Text Encoder Module: It converts clinical documentation and user questions into high-dimensional embedding vectors via a specialized transformer model[31].
- Multimodal Fusion and QA Module: This core module amalgamates the visual and textual embeddings to execute logical reasoning and formulate an answer[32].
- User Interface Module: This module delivers the front-end platform that allows for seamless user interaction with the entire system[33].

V. METHODOLOGY

Our development life cycle is managed using an Agile methodology[34]. The model's training is founded on transformer-based architectures for both its natural language processing and computer vision functionalities[35].

A. Technology Stack

- Programming Languages: Python is the primary language, with JavaScript available as an option for frontend development[36].
- Frameworks: Key frameworks include PyTorch or TensorFlow, HuggingFace Transformers, and either Flask or FastAPI for the backend[37].
- Databases: Data storage is handled by MongoDB or PostgreSQL[38].
- Tools/APIs: Essential tools include OpenCV and either NLTK or spaCy[39].
- Platforms: The project utilizes cloud infrastructure from Google Cloud or AWS, Jupyter Notebooks for rapid prototyping, and Docker for containerized deployment[40].

B. Implementation and Training

The implementation is broken down into these essential phases[41]:

1. Model Selection: We employ a Vision Transformer (ViT) for the image encoding task and a language model predicated on T5 or BioBERT for text processing[42].
2. Fusion Mechanism: Cross-attention mechanisms are employed to effectively align the different data modalities[43].
3. Data Acquisition: The model is trained on datasets that are publicly available, such as MIMIC-CXR, MedQA, and CheXpert[44]. The initial step involves acquiring and curating this data[45].
4. Training: The training protocol starts with applying transfer learning to the vision and language encoders (pretraining)[46]. It then moves to training the cross-attention layers to capture cross-modal relationships (fusion training), and finally, it involves fine-tuning on specific downstream tasks to achieve domain adaptation[47].

VI. EVALUATION AND EXPECTED OUTCOMES

The system's effectiveness will be gauged using a blend of automated evaluation techniques and reviews by domain experts[48].

- Automated Metrics: Performance will be quantified using Accuracy, F1 score, and AUC for classification, in addition to BLEU, ROUGE, and METEOR for assessing QA capabilities[49].

- Expert Review: Medical specialists will conduct manual validations of the model's answers to verify their clinical accuracy and relevance[50].
- Ablation Studies: These experiments will be performed to isolate and measure the contribution of the multimodal fusion component to the overall performance[51].

The anticipated deliverables include:

- A complete and functional multimodal AI system adept at answering clinical queries based on both imaging and textual data[52].
- An operational prototype that includes an interface for uploading images and posing questions in natural language[53].
- A set of quantitative performance metrics that validate the system's diagnostic accuracy and clinical utility[54].
- Features that support explainability, for instance, visual attention heatmaps like Grad-CAM[55].

VII. CONCLUSION AND FUTURE ENHANCEMENTS

This research is centered on the creation of a cohesive multimodal AI model that aims to overcome the constraints of unimodal systems by integrating visual and textual data for clinical question answering[56]. This methodology has the potential to refine diagnostic precision, alleviate the operational burden on healthcare professionals, and democratize access to medical expertise[57].

Future development paths may include[58]:

- Feature Integration: Incorporating a voice-based interface for question answering, deploying the service for real-time use on a cloud infrastructure, and enabling integration with Electronic Health Records (EHR)[59].
- Scalability: Customizing the model for use in specialized medical fields like cardiology and oncology, and potentially developing a Software as a Service (SaaS) platform for clinical institutions[60].

REFERENCES

[1] Chen Wei, Sandeep Gupta., "A Finetuned Multimodal Deep Learning Model for Xray Image Diagnosis", 2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE) | 979-8-3503-5317-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/CISCE62493.2024.10653104

[2] Kishan Patel, Honghao Wang., "Evaluating Machine Learning Based Automatic Quality Assurance of Magnetic Resonance Images in Healthcare System", 2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT) | 979-8-3315-4121-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICAICCT64383.202 4.10912290

[3] Arvind Kumar, Anupam Agrawal., "Towards Precision Healthcare: Leveraging Pre-trained Models and Attention Mechanisms in Medical Visual Question Answering for Clinical Decision Support", 2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON) | 979-8-3503-7289-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/NMITCON62075.2024.10699307

[4] Tariq Mahmood., "Recent advancement and future prospect in active deep learning for medical image segmentation and classification," in Digital Object Identifier 10.1109/ACCESS.2023.3313977, September 2023.

[5] J. Singhal, et al., "Large language models encode clinical knowledge," npj Digital Medicine, vol. 6, no. 1, pp. 1-9, 2023.

[6] K. Narayanaswamy, et al., "Med-palm 2: Towards expert-level medical question answering with a large language model," arXiv preprint arXiv:2305.09617, 2023.

[7] J. Johnson, et al., "BioViL-T: Self-supervised vision-and-language pretraining of biovil using clinical and biomedical data," IEEE Transactions on Medical Imaging, vol. 42, no. 6, pp. 1445-1458, June 2023.

[8] T. Wang, et al., "Unified medical vision-language pretraining for bimodal image-text understanding and generation," in Proceedings of the IEEE/CVF CVPR 2023, pp. 18121-18131, June 2023.