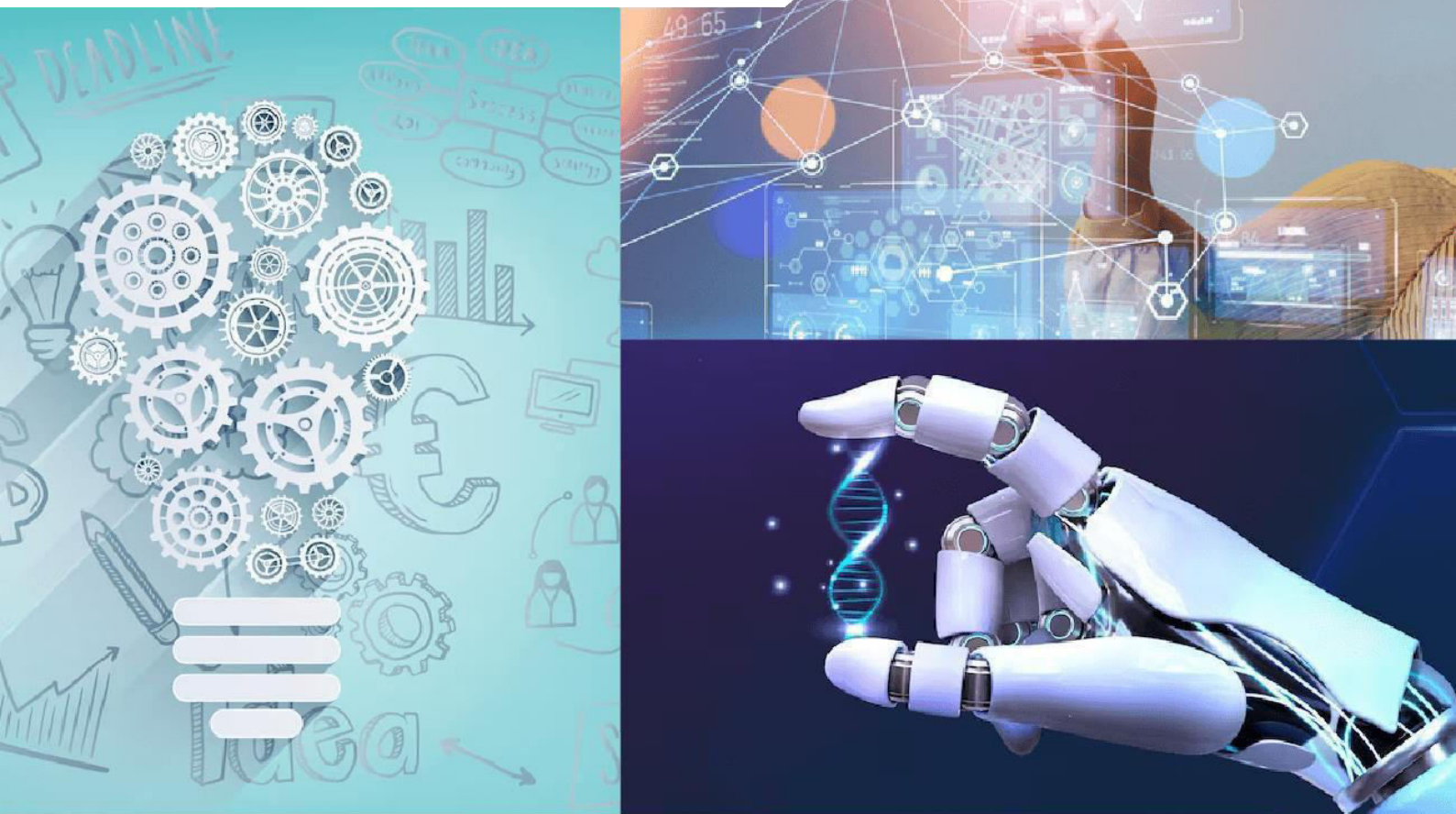




International Journal of Advanced Research in Education and Technology (IJARETY)

Volume 13, Issue 3, May - June 2026

Impact Factor: 9.971



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



Large Language and Vision Assistant for Biomedicine

Akash B H, Dr. Vakula Rani

Department of MCA, CMR Institute of Technology, Bangalore, India

Department of MCA, CMR Institute of Technology, Bangalore, India

ABSTRACT: The integration of artificial intelligence in healthcare has significantly enhanced biomedical data analysis, particularly in medical imaging and natural language processing. However, most existing systems operate independently, focusing either on image understanding or textual interaction, creating a gap for learners and non-expert users.

This paper presents a multimodal system that combines a vision model and a language model to analyze biomedical images and generate descriptive explanations along with interactive question-answering capabilities. Users can upload medical images such as X-rays, CT scans, and MRI scans and receive contextual insights through natural language interaction.

The proposed system is designed as an educational tool rather than a diagnostic system. It focuses on simplicity, accessibility, and the integration of pre-trained models. Experimental observations show that combining visual and textual understanding improves user learning and engagement in biomedical applications.

KEYWORDS: Multimodal AI, Biomedical Imaging, Natural Language Processing, Vision-Language Models, Medical AI, Large language model, medical image analysis, multimodal AI, biomedical question answering, vision transformer, natural language processing, prompt engineering

I. INTRODUCTION

Medical imaging is a fundamental component of modern healthcare, enabling clinicians to visualize internal structures of the human body and identify abnormalities. Techniques such as X-rays, Computed Tomography (CT), and Magnetic Resonance Imaging (MRI) are widely used for diagnosis and treatment planning. However, interpreting these images requires significant expertise and experience, making it challenging for students and beginners to understand complex biomedical visuals.

In parallel, advancements in artificial intelligence, particularly in Natural Language Processing (NLP), have led to the development of systems capable of understanding and generating human-like text. These systems are widely used in chatbots, virtual assistants, and educational tools. Despite their capabilities, most NLP systems operate only on textual data and lack the ability to interpret visual information.

Similarly, computer vision techniques have achieved remarkable success in analyzing medical images. Deep learning models, especially convolutional neural networks, can detect patterns and extract meaningful features from biomedical images. However, these systems often produce outputs in the form of labels, probabilities, or annotations, which may not be easily interpretable by non-expert users.

The lack of integration between vision-based systems and language-based systems creates a gap in usability and understanding. Users often need to rely on multiple tools to analyze images and obtain explanations, which reduces efficiency and limits interactive learning.

To address this limitation, this paper proposes a Large Language and Vision Assistant for Biomedicine, a multimodal system that combines image understanding with natural language interaction. The system allows users to upload biomedical images and receive descriptive explanations, followed by the ability to ask questions related to the image. The responses are generated based on both visual features and textual input, creating an interactive learning experience.

The primary objective of this system is to enhance understanding rather than provide clinical diagnosis. It is designed as an educational and demonstration tool that showcases the integration of vision and language models in biomedical applications. By combining these technologies, the system provides a more intuitive and user-friendly approach to exploring medical images.

The rest of the paper is organized as follows: Section II presents the literature review, Section III describes the proposed methodology, Section IV discusses results and observations, and Section V concludes the paper with future directions.

II. LITERATURE REVIEW

The application of artificial intelligence in biomedical domains has gained significant attention in recent years, particularly in medical image analysis and natural language processing. This section reviews existing approaches and identifies research gaps that motivated the proposed system.

Medical image analysis has been extensively studied using deep learning techniques. Convolutional Neural Networks (CNNs) have demonstrated strong performance in detecting patterns, segmenting regions, and classifying diseases from images such as X-rays, CT scans, and MRI scans. These models are capable of extracting complex visual features and achieving high accuracy. However, most of these systems provide outputs in the form of labels or predictions without offering meaningful explanations, which limits their usability for educational purposes.

On the other hand, Natural Language Processing (NLP) has evolved rapidly with the development of transformer-based models. These models are capable of understanding context, generating coherent text, and answering user queries. In the healthcare domain, NLP has been used for clinical documentation, medical chatbots, and summarization of patient records. Despite these advancements, NLP systems generally operate only on textual data and do not incorporate visual information. Recent research has explored multimodal systems that combine vision and language models. These systems aim to bridge the gap between image understanding and textual reasoning. Vision-language models are capable of generating captions, answering questions about images, and providing contextual explanations. However, many of these systems are designed for high-performance applications and require substantial computational resources, making them less suitable for academic or lightweight implementations.

Another limitation observed in existing research is the lack of user-centric design. Many systems prioritize accuracy and performance while neglecting usability and accessibility. As a result, non-expert users, such as students, often find it difficult to interpret outputs or interact effectively with these systems. Furthermore, most biomedical AI systems are developed for clinical environments and focus on diagnostic accuracy rather than educational value. This creates a gap in tools that can help learners understand medical images in a simple and interactive manner.

Based on the analysis of existing approaches, it is evident that there is a need for a system that integrates visual understanding and natural language interaction while maintaining simplicity and accessibility. The proposed system addresses this gap by combining a vision model and a language model into a unified framework designed specifically for educational use.

This work focuses on practical implementation rather than developing new algorithms. It emphasizes system integration, user interaction, and interpretability, making it suitable for academic environments and demonstration purposes.

III. METHODOLOGY

The proposed system is designed as a multimodal framework that integrates image processing and natural language interaction to analyze biomedical images. The system combines a vision model and a language model to generate descriptive explanations and answer user queries based on image content.

A. System Overview

The system operates through a structured pipeline consisting of three major stages: input processing, feature extraction, and response generation. Users interact with the system by uploading biomedical images and asking questions related to the image.

The system is designed to be lightweight and suitable for academic environments by utilizing pre-trained models instead of training from scratch.

B. System Architecture

The architecture of the proposed system consists of three main layers:

- **Input Layer:** Handles image upload and user queries.
- **Processing Layer:** Performs image preprocessing, feature extraction, and language processing.
- **Output Layer:** Displays generated descriptions and responses.

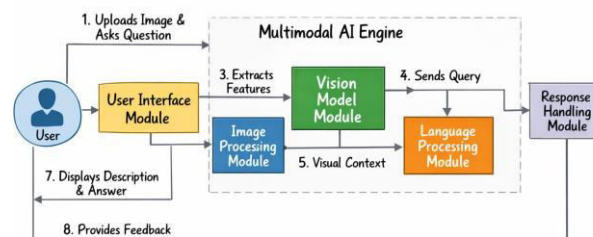


Fig.1. System Architecture of the Proposed Model

The architecture ensures smooth interaction between visual and textual components, enabling effective multimodal processing.

C. Workflow of the System

The system follows a sequential workflow from image input to response generation. Each stage is interconnected to ensure accurate and context-aware outputs.

The workflow is described as follows:

1. The user uploads a biomedical image through the interface.
2. The system validates and preprocesses the image.
3. The vision model extracts important visual features.
4. A textual description of the image is generated.
5. The user submits a query related to the image.
6. The system combines visual features with the query.
7. The language model generates a contextual response.

D. Image Preprocessing

Image preprocessing is an essential step to prepare the input data for the vision model. The system ensures that all images are converted into a consistent format before analysis.

The preprocessing steps include:

- Resizing the image to a standard resolution
- Normalizing pixel values
- Converting image format if required

These steps improve model performance and ensure reliable feature extraction.

Multimodal Integration of Vision and Language Model

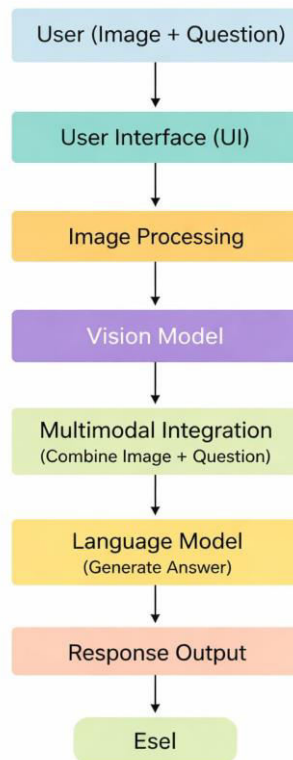


Fig.2. Workflow of the Proposed System

E. Vision Model for Feature Extraction

The vision model is responsible for analyzing the biomedical image and extracting meaningful features. A pre-trained deep learning model is used to identify patterns, structures, and important regions within the image.

The extracted features represent visual information that is later used by the language model to generate descriptions and answers. Using a pre-trained model reduces computational cost and makes the system feasible for academic implementation.

F. Language Model for Text Generation

The language model generates textual outputs based on the extracted visual features and user queries. It produces both descriptive explanations and answers in a conversational format.

The model processes:

- Visual features from the image
- User input queries

The output is designed to be simple, clear, and understandable, making it suitable for students and non-expert users.

G. Multimodal Integration

The core strength of the system lies in the integration of visual and textual data. The multimodal integration mechanism combines image features with user queries to generate context-aware responses.

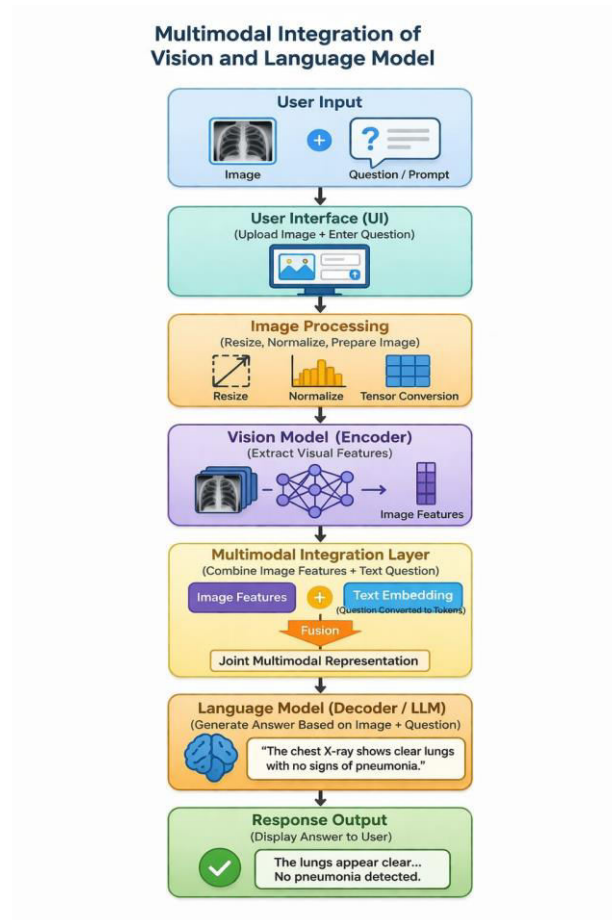


Fig.3. Multimodal Integration of Vision and Language Models

This integration allows the system to understand both what is present in the image and what the user is asking, resulting in more relevant and meaningful responses.

H. User Interface

A simple and user-friendly interface is designed to facilitate interaction between the user and the system. The interface allows users to upload images, view generated descriptions, and ask questions.

The design ensures:

- Ease of use
- Minimal learning curve
- Clear presentation of outputs

I. Advantages of the Proposed System

The proposed system offers several advantages:

- Integration of image analysis and natural language interaction
- Interactive learning through question-answering
- Use of lightweight and accessible tools
- Suitable for academic and demonstration purposes

IV. RESULTS AND DISCUSSION

The proposed system was implemented and evaluated in an academic environment using standard computational resources. The objective of the evaluation was to verify the system's ability to analyze biomedical images and generate meaningful textual descriptions along with context-aware re- sponses to user queries.

A. Experimental Setup

The system was developed using Python and integrated with open-source libraries for image processing and natural language generation. Pre-trained models were used for both vision and language components to ensure feasibility within limited hardware resources.

The system was tested on different types of biomedical images such as X-rays and MRI scans. The evaluation focused on usability, response quality, and interaction capability rather than clinical accuracy.

B. Image Description Results

The system successfully generated textual descriptions for uploaded biomedical images. The generated descriptions high- lighted key structures and visible patterns in the image in a simplified and understandable format.

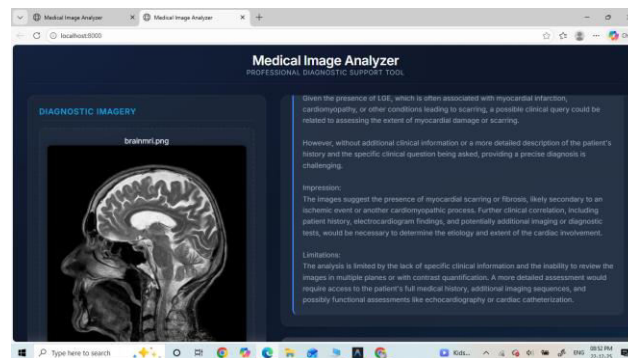


Fig.4. Generated Description for Uploaded Biomedical Image

As shown in Fig. 1, the system analyzes the uploaded image and produces a descriptive output. The description avoids complex medical terminology and focuses on clarity, making it suitable for students and non-expert users.

C. Question and Answer Interaction

The interactive question-answering feature allows users to ask queries related to the uploaded image. The system processes both the image features and user input to generate context-aware responses.

As illustrated in Fig. 2, the system provides relevant re- sponses to user queries. This demonstrates the effectiveness of multimodal integration, where both visual and textual information contribute to response generation.

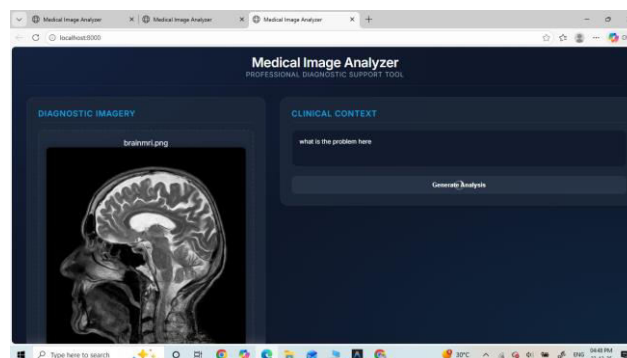


Fig.5. Question and Answer Interaction Based on Image

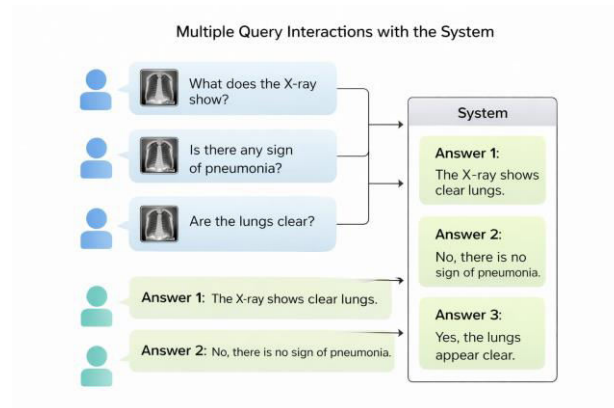


Fig.6. Multiple Query Interaction with the System

D. Multi-Query Interaction Analysis

The system supports multiple queries for the same image, allowing users to explore different aspects of the biomedical image through continuous interaction.

The ability to handle multiple queries enhances user engagement and supports deeper understanding of image content.

E. User Interface Evaluation

The system provides a simple and intuitive user interface that enables easy interaction. Users can upload images, view generated descriptions, and ask questions without requiring technical expertise.

The interface design ensures smooth navigation and improves overall user experience.

F. Performance Analysis

The system performance was evaluated based on response time and system stability. The results indicate that the system operates efficiently within academic constraints.

- The system processes images within a reasonable time.
- Response generation is consistent and stable.
- Performance varies slightly based on image complexity.

G. Discussion

The results demonstrate that the proposed system effectively combines vision and language models to provide meaningful insights from biomedical images. The integration of image analysis with natural language interaction enhances user understanding and creates an interactive learning environment.

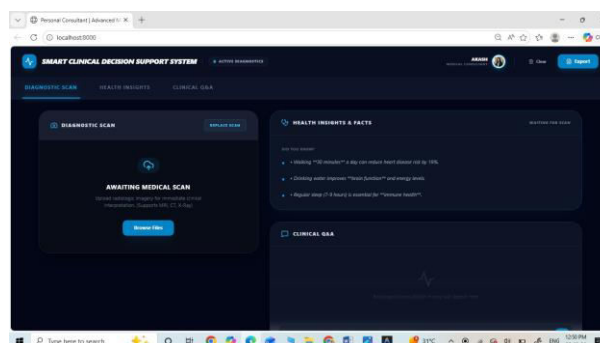


Fig.7. User Interface of the Proposed System

Unlike traditional systems that provide only predictions or labels, the proposed system focuses on explanation and interaction. This makes it more suitable for educational purposes. However, the system has certain limitations. It does not provide medical diagnosis, and its accuracy depends on the quality of pre-trained models. Despite these

limitations, the system achieves its primary objective of demonstrating multimodal AI in biomedicine

REFERENCES

- 1) Esteva et al., "Dermatologist-level classification of skin cancer," Nature, 2017.
- 2) G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, 2017.
- 3) O. Ronneberger et al., "U-Net: Convolutional networks for biomedical image segmentation," MICCAI, 2015.
- 4) K. He et al., "Deep residual learning for image recognition," CVPR, 2016.
- 5) A. Krizhevsky et al., "ImageNet classification with deep convolutional neural networks," NIPS, 2012.
- 6) Vaswani et al., "Attention is all you need," NeurIPS, 2017.
- 7) T. Brown et al., "Language models are few-shot learners," NeurIPS, 2020.
- 8) J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers," NAACL, 2019.
- 9) R. Radford et al., "Learning transferable visual models from natural language supervision," ICML, 2021.
- 10) OpenAI, "GPT models and applications," 2023.
- 11) A. Dosovitskiy et al., "An image is worth 16x16 words: Vision transformers," ICLR, 2021.
- 12) S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, 1997.
- 13) Y. LeCun et al., "Gradient-based learning applied to document recognition," IEEE, 1998.
- 14) D. Shen et al., "Deep learning in medical image analysis," Annual Review, 2017.
- 15) H. Greenspan et al., "Guest editorial: Deep learning in medical imaging," IEEE TMI, 2016.
- 16) J. Deng et al., "ImageNet: A large-scale hierarchical image database," CVPR, 2009.
- 17) Szegedy et al., "Going deeper with convolutions," CVPR, 2015.
- 18) Y. Bengio et al., "Deep learning," Nature, 2015.
- 19) S. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach," 2016.
- 20) Goodfellow et al., "Deep Learning," MIT Press, 2016.



International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975

Impact Factor: 9.971