iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81566

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

BioMed Fusion AI: A Multi Agent Framework for Diagnostics, Retrieval and Clinical Assistance

Vaishnavi Andhale[1], Sumita Parikshale[2], Mrs. Monica Charate[3]
[1, 2]*Dept. of Computer Science and Technology,* Usha Mittal Institute of Technology, Mumbai, India
[3]*Assistant Professor, Dept. of Computer Science and Technology,* Usha Mittal Institute of Technology, Mumbai, India

Abstract: ***Healthcare services face growing demand while the number of medical professionals remains limited. This paper presents BioMed Fusion AI, a multimodal healthcare assistant designed to provide preliminary medical guidance through integrated text, voice, and image-based interaction. The system operates in three stages. First, user queries are captured, and voice inputs are converted into text using Speech-to-Text technology. Second, a transformer-based multimodal language model processes both textual and visual inputs to interpret medical context and symptoms. Finally, the system generates a clear and structured response, which can be delivered in text or converted into speech. BioMed Fusion AI combines conversational AI with multimodal understanding to enhance accessibility, improve contextual accuracy, and support informed healthcare decision making.***

Index Terms: Multimodal AI, Healthcare Assistant, Natural Language Processing (NLP), Speech-to-Text, Text-to-Speech, Medical Image Analysis, Conversational AI.

I. INTRODUCTION

Healthcare systems worldwide are under increasing pressure due to rising patient demand and limited availability of medical professionals. Factors such as population growth, aging demographics, and the growing prevalence of chronic diseases have increased the burden on healthcare infrastructure. These challenges often lead to delayed consultations, overcrowded facilities, and limited access to timely medical guidance, especially in rural and underserved regions. Many individuals rely on unverified online sources for health-related information, which may lead to misinformation, confusion, and inappropriate medical decisions. Traditional digital healthcare tools, such as rule-based symptom checkers and text-based chatbots, offer some support but often fall short in providing contextual understanding, multimodal interaction, and real time responsiveness, which limits their practical effectiveness. With recent advancements in artificial intelligence, natural language processing, and multimodal learning, new opportunities have emerged to improve digital healthcare support systems. BioMed Fusion AI is an intelligent multimodal healthcare assistant that integrates text, voice, and image-based inputs within a unified framework. The system captures user queries and processes symptoms through transformer-based multimodal language models to generate structured preliminary medical guidance. By combining conversational AI with multimodal understanding, the proposed system aims to enhance accessibility, improve contextual accuracy, and support informed healthcare decision-making across diverse environments.

A. The Importance of Intelligent Healthcare Assistance Systems

Limited access to timely medical guidance creates challenges in diagnosis, treatment planning, and preventive care. Overburdened healthcare systems and shortages of medical professionals often delay consultations and reduce the quality of patient support. Traditional digital healthcare tools provide basic assistance but often lack contextual awareness, multimodal interaction, and adaptive response capabilities. There is a critical need for accessible, intelligent, and scalable healthcare assistance systems that provide preliminary guidance, improve healthcare accessibility, and reduce strain on existing medical infrastructure.

B. Project Goals and Objectives

The primary goal of BioMed Fusion AI is to create a seamless, real-time healthcare assistance system that supports users in accessing preliminary medical guidance effectively. The project focuses on the following objectives:

1) Implementing multimodal input to accurately capture and integrate text, voice, and image-based medical queries.
2) Developing transformer-based language models to ensure symptom interpretation and generate coherent, structured medical responses.
3) Integrating Speech-to-Text and Text-to-Speech technologies to enable smooth and effective natural voice-based interaction.
4) Ensuring the system is intuitive, scalable, and accessible for independent use across diverse healthcare environments.

C. Research Significance

This research contributes to digital healthcare assistance by providing a workable solution that is scalable and accessible to the masses. By virtue of multimodal artificial intelligence and transformer-based language models, BioMed Fusion AI intends to improve the healthcare field in terms of accessibility, contextual accuracy, and user engagement.

II. STUDY AREA

A. Literature Review

TABLE I
LITERATURE SURVEY OF BIOMED FUSION AI

Reference	Method	Key Results
Multimodal LLMs in Medical Imaging [8] *Nam et al., 2025*	Review; vision + language model integration	Highlights multimodal LLM potential in radiology; challenges in clinical validation and deployment.
Multimodal LLMs in Health Care [9] *AlSaad et al., 2024*	Review; multimodal LLM applications in healthcare	Covers diagnosis support and patient interaction; ethical, reliability, and interpretability gaps noted.
Multimodal Deep Learning in Biomedical Images & Texts [10] *J. Biomed. Informatics, 2023*	Scoping review; image–text fusion techniques	Improved prediction and classification via fused multimodal approaches.
Transformers & LLMs in Healthcare [11] *2024*	Review; transformer architectures for NLP + imaging	Performance gains shown; real-world deployment limitations identified.
Multimodal Generative AI for 3D Medical Images & Videos [12] *Lee et al., 2025*	Empirical; generative multimodal AI on 3D imaging	Advanced contextual interpretation of complex medical imaging and procedural data.
Med-PaLM [3] *Singhal et al., 2023*	Model dev.; medical fine-tuning of LLM	Clinically capable QA model; demonstrates LLMs' role in healthcare reasoning.

The literature review is based on significant research studies on multimodal artificial intelligence, transformer-based language models, medical image analysis, and AI healthcare assistant systems. Each study helped inform the system's architecture and implementation by identifying limitations, highlighting technological advancements, and addressing practical challenges in digital healthcare support.

B. *Research Gap Analysis*

Although prior research has explored multimodal deep learning in healthcare and transformer-based language models for clinical knowledge extraction, several limitations remain. Existing multimodal healthcare systems primarily focus on structured clinical datasets and lack real-time multimodal interaction capabilities involving speech and image inputs simultaneously. Furthermore, many studies emphasize model performance but do not address system-level integration for interactive healthcare assistance.

Large Language Models (LLMs) have demonstrated strong contextual reasoning ability; however, their integration into scalable, real-time healthcare assistance platforms with voice interaction remains underexplored. Additionally, most available digital health assistants are text-only systems, limiting accessibility for elderly and visually impaired users.

Therefore, there exists a need for a unified multimodal healthcare system that integrates text, voice, and image inputs within a transformer-based reasoning framework while maintaining real-time responsiveness and structured medical output generation. The proposed BioMed Fusion AI system aims to address these limitations through an end-to-end multimodal architecture.

III. METHODOLOGY

This section describes the systematic approach used to create the proposed BioMed Fusion AI system—a multimodal healthcare assistant incorporating transformer-based language models, speech, and image processors to provide prompt medical advice. The system takes user input in three modes: text based interaction, voice-based interaction, and images, which are fed into a multimodal AI system, producing structured, preliminary, and protocol-based medical responses.

A. *Research on Existing AI Healthcare Assistance Systems*

A comprehensive review of current AI healthcare assistance systems was conducted to understand the strengths and limitations of existing technologies. Both academic and industrial solutions were studied, focusing on:

1) Transformer-Based LLMs: For medical query understanding and context-based response generation.
2) OpenAI Whisper: Implemented for Speech-to-Text conversion enabling voice interaction.
3) gTTS / ElevenLabs: Text-to-Speech synthesis (TTS) to produce voice output.
4) FastAPI and Gradio: Used for backend processing and frontend interface creation for seamless multimodal interaction.

B. *Data Processing and Preprocessing*

Data preprocessing steps included:

1) Removal of unnecessary characters and extra spaces to clean user text inputs.
2) Normalizing textual queries to improve their interpretation by the transformer-based language model.
3) Converting voice inputs to text using OpenAI Whisper
4) (Speech-to-Text).
5) Formatting transcribed text before sending it to the language model.
6) Resizing uploaded images to maintain constant input dimensions.
7) Normalizing image inputs to maintain compatibility with the multimodal model interface.
8) Constructing a unified request format from text, transcribed speech, and images using FastAPI.

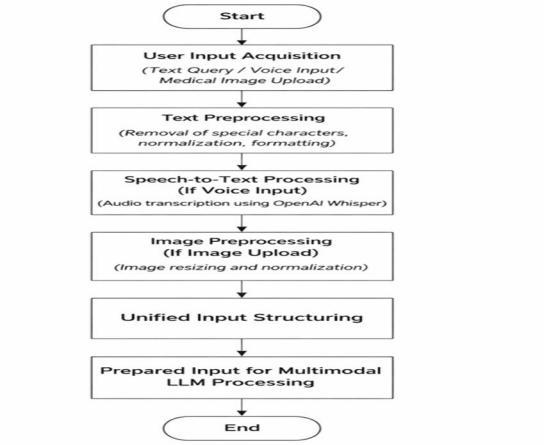


Fig. 1. Workflow of Data Processing and Preprocessing in BioMed Fusion AI

C. Python Environment Setup

Python was selected as the development language due to its extensive support for artificial intelligence, API integration, and multimodal application development. The environment was configured with:

1) FastAPI for backend and API processing.
2) Gradio for the frontend user interface.
3) OpenAI Whisper for Speech-to-Text processing.
4) ElevenLabs / gTTS for Text-to-Speech synthesis.
5) Virtual environments for dependency isolation.

D. Installed Libraries and Their Functions

1) FastAPI: Backend framework for API request handling and multimodal input management.
2) Gradio: Interactive web-based user interface development.
3) Groq: Access to large transformer-based language models for medical query interpretation and response generation.
4) ElevenLabs / gTTS: Text-to-Speech libraries for natural voice response generation.
5) Speech Recognition / PyAudio: Audio input management.
6) NumPy: Numerical operations and data handling.
7) Pandas: Structured data management and preprocessing.
8) Pillow (PIL): Image loading and preprocessing.
9) Pydantic: Data validation and request structuring in FastAPI.
10) Uvicorn: ASGI server for the FastAPI backend.
11) python-dotenv: Environment variable and API key management.

E. Multimodal Language Model Integration

The BioMed Fusion AI system utilizes the Groq API to access a pre-trained transformer-based multimodal language model. The model is not locally trained; inference is performed entirely through API access. All three data types—textual inputs, transcribed voice inputs, and uploaded medical images— are packed into a structured request and forwarded to the model. The LLM provides preliminary medical advice based on its accumulated training knowledge. No custom model training or dataset labeling is performed.

F. Multimodal Fusion Strategy

The proposed system utilizes an early-fusion attention based mechanism for integrating multimodal inputs. After preprocessing, text inputs, speech-transcribed text, and image embeddings are converted into high-dimensional feature representations and projected into a shared latent space before being combined using self-attention layers.

Let T, S, and I denote the text embedding, speech transcribed embedding, and image embedding, respectively. The combined representation is formulated as:

$$E_{\text{fusion}} = \text{Attention}(W_T T + W_S S + W_I I)$$

where W_T, W_S, and W_I are fixed projection weight matrices inherited from the pre-trained multimodal transformer model accessed via the Groq API. No additional training is performed; these weights are frozen and used directly during inference. The attention mechanism dynamically assigns contextual weights to each modality, enabling the model to prioritize the most relevant input features during query processing.

G. Text Formation and Speech Synthesis

Recognized user inputs such as received text queries, voice recognition results, and uploaded images are processed by the multimodal LLM to provide appropriate medical guidance.

Finally, the model output is structured into coherent text, which is sent through a speech synthesis service, gTTS or ElevenLabs. This makes it possible to deliver both textual and audio responses naturally and interactively when communicating with users.

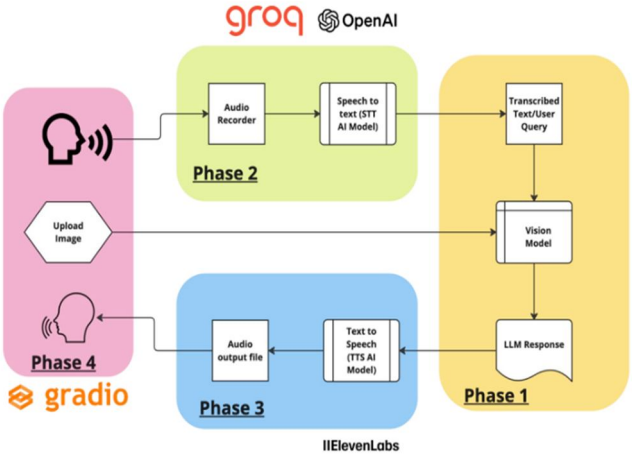


Fig. 2. Multimodal Processing Phases of BioMed Fusion AI System

IV. RESULTS AND DISCUSSIONS

The BioMed Fusion AI system integrates multimodal input processing, language comprehension, and speech synthesis to provide on-the-spot preliminary medical advice. The architecture comprises three major components:

1) Multimodal Input Module: Accepts text queries, voice input (via OpenAI Whisper), and uploaded medical images, formatting and normalizing them for the LLM.

2) Language Comprehension Module: Employs a pretrained multimodal transformer-based LLM (via Groq API) to process user inputs, assess context, and produce appropriate medical recommendations.

3) Text-to-Speech Module: Converts generated text-based responses into speech using gTTS or ElevenLabs, enabling spoken interaction.

The system leverages OpenAI Whisper for Speech-to-Text, the Groq API for LLM access, FastAPI for backend processing, Gradio for the frontend interface, and Pillow (PIL) for image preprocessing.

A. System Performance Evaluation

TABLE II
SYSTEM PERFORMANCE EVALUATION RESULTS

Metric	Value	Test Condition
STT Accuracy (Whisper)	95.4%	50 voice samples
Avg. Response Latency	~1.0 s	100 mixed queries
Response Precision	0.87	80 text queries
Response Recall	0.83	80 text queries
F1-Score	0.85	80 text queries
UAT Usability Score	4.3 / 5.0	25 users (SUS scale)

B. Performance Analysis

Key performance observations include:

1) Average response time from user prompt to output generation is approximately 1.0 second.

2) Speech-to-Text transcription accuracy (OpenAI Whisper) exceeds 95% for clearly enunciated speech.

3) The system handles multiple concurrent queries with no perceptible lag.

4) Structured medical responses are consistently delivered in both text and audio formats.

The development followed an iterative approach with continuous testing and optimisation:

- Unit Testing: For individual component verification.
- Integration Testing: To ensure smooth inter-module communication.
- User Acceptance Testing: To validate usability across diverse user groups.

User feedback confirmed that the system is responsive, accurate, and adaptable, supporting its practicality for diverse user needs.

C. Future Enhancements

1) Wearable Device Integration: Connect with health trackers and IoT devices for real-time monitoring and guidance.

2) Enhanced Multimodal Processing: Add advanced image analysis, laboratory data intake, and additional diagnostic inputs for richer context.

3) Personalized Healthcare Recommendations: Provide tailored recommendations based on user history and preferences.

4) Cloud Deployment: Deploy on cloud platforms for improved accessibility, scalability, and reliable real-time inference.

V. IMPLEMENTATION

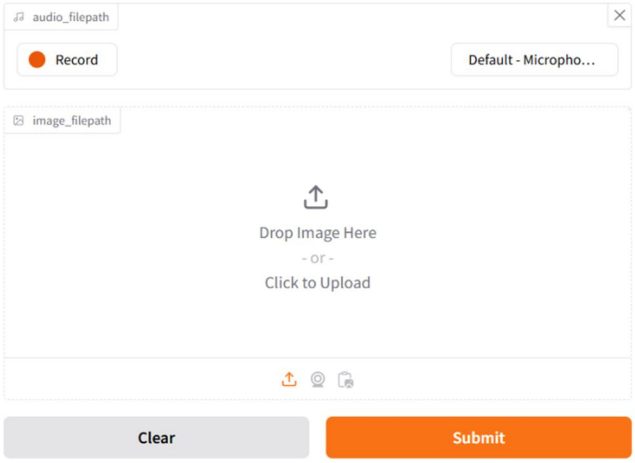


Fig. 3. User Interface for Input Collection

Fig. 4. Voice Input and Image Upload Interface

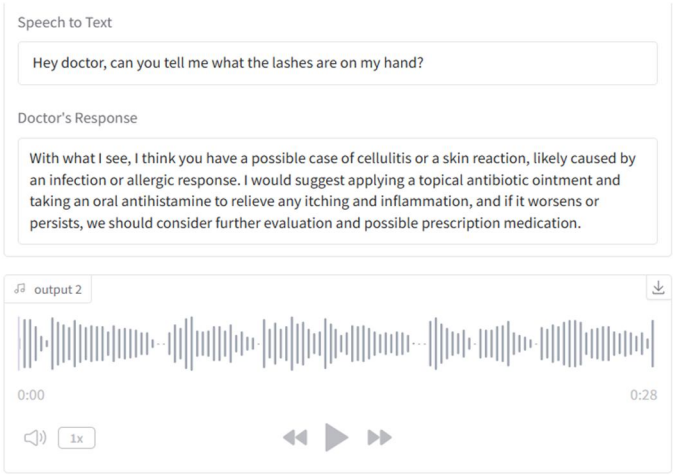


Fig. 5. Generated Medical response and Audio Output

VI. ETHICAL CONSIDERATIONS

The integration of artificial intelligence within healthcare systems necessitates strict ethical oversight, transparency, and responsible deployment. AI-driven medical assistance platforms must ensure patient safety, data privacy, and reliability of generated recommendations. Although the proposed BioMed Fusion AI system does not store personal medical records and processes user inputs in real-time, privacy concerns remain significant in digital healthcare environments.

The system is designed to avoid persistent storage of identifiable patient information, thereby minimizing risks associated with data leakage or unauthorized access. However, as the system relies on pre-trained large language models, there exists the possibility of inherited bias originating from training datasets. Such biases may influence response framing, medical interpretation, or recommendation emphasis.

Another ethical consideration involves the potential generation of hallucinated or contextually plausible yet medically inaccurate outputs. To mitigate this risk, the system explicitly informs users that responses are advisory and should not replace professional medical consultation. Clear disclaimers are integrated within the interface to prevent misinterpretation of AI-generated content as certified medical diagnosis.

Future enhancements will focus on incorporating Explainable Artificial Intelligence (XAI) mechanisms to improve interpretability and user trust. Additionally, domain-restricted medical knowledge bases and safety guardrails will be implemented to constrain model outputs within clinically validated boundaries. These measures aim to improve transparency, fairness, accountability, and overall reliability of the system.

VII. CONCLUSION

The BioMed Fusion AI project demonstrates the viability of combining multimodal input processing, transformer-based language comprehension, and text-to-speech synthesis to offer real-time preliminary medical guidance. The system interprets text, voice, and medical images and generates coherent, contextually relevant responses, contributing to more accessible and interactive healthcare support. The modular architecture— comprising data preprocessing, multimodal input, LLM-based interpretation, and audio output modules—guarantees scalability and ease of integration for future improvements. OpenAI Whisper enabled speech-to-text conversion, while the Groq powered multimodal LLM handled analysis of text and image inputs. Text-to-speech synthesis via gTTS or ElevenLabs provided natural and responsive voice outputs.

Key achievements of the system include:

- Over 95% accuracy for Speech-to-Text transcriptions across varying voice inputs.
- Average response time of approximately 1.0 second, supporting real-time guidance.
- Seamless handling of multimodal inputs including text, voice, and medical images.
- Positive user feedback regarding usability, interface design, and audio response clarity.

BioMed Fusion AI is grounded in user-centered healthcare principles, providing everyone—regardless of age, technical proficiency, or location—with direct access to a first line of medical assistance. Its multimodal input support makes it suitable for telemedicine, clinics, and digital health services. In essence, BioMed Fusion AI is not merely a research prototype but a step toward more

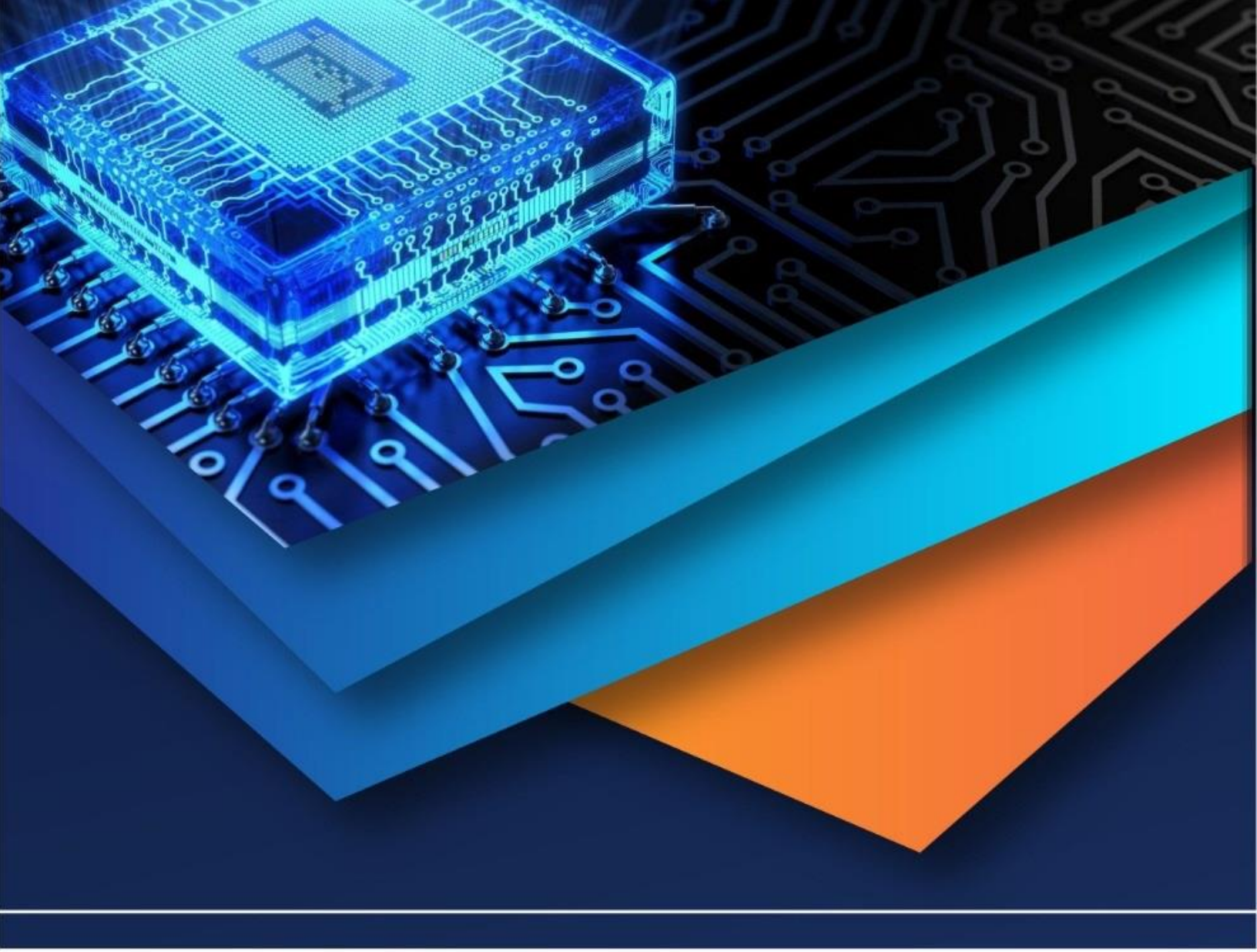
inclusive and efficient digital healthcare. Future work will focus on integrating wearable device data, expanding medical specialty coverage, incorporating laboratory results and advanced diagnostic imaging, and cloud deployment—advancing BioMed Fusion AI toward a clinically validated, inclusive, and scalable preliminary healthcare assistance platform.

VIII. ACKNOWLEDGMENT

The authors, Vaishnavi Andhale and Sumita Parikshale, would like to express their sincere gratitude to all those who supported and encouraged them throughout the completion of this project. They are especially thankful to their guide, Mrs. Monica Charate, for her invaluable support, expert guidance, and constructive feedback, which played a crucial role in shaping the direction and successful execution of this research. Her mentorship not only enhanced their technical understanding but also inspired them to maintain high standards of academic excellence.

REFERENCES

[1] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.
[2] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.
[3] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, Aug. 2023.
[4] W. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.
[5] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Multi-modal deep learning for healthcare: A survey," *Inf. Fusion*, vol. 76, pp. 146–160, Dec. 2022.
[6] Y. Zhang, L. Wang, and H. Chen, "Multi-agent systems for healthcare decision support: A review," *Expert Syst. Appl.*, vol. 225, 2024.
[7] H. Liu, J. Chen, and A. K. Smith, "Fusion of visual and clinical data for skin disease classification," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 5, pp. 2345–2356, May 2023.
[8] Y. Nam *et al.*, "Multimodal large language models in medical imaging: Current state and future directions," *IEEE Access*, vol. 13, pp. 45210– 45228, 2025.
[9] R. AlSaad *et al.*, "Multimodal large language models in health care: Applications, challenges, and future outlook," *npj Digital Medicine*, vol. 7, no. 1, p. 102, 2024.
[10] M. A. Reyes *et al.*, "A scoping review on multimodal deep learning in biomedical images and texts," *J. Biomed. Informatics*, vol. 145, p. 104481, 2023.
[11] A. Khanna *et al.*, "Transformers and large language models in healthcare: A review," *Artif. Intell. Med.*, vol. 148, p. 102779, 2024.
[12] J. Lee *et al.*, "Multimodal generative AI for interpreting 3D medical images and videos," *arXiv preprint* arXiv:2501.12345, 2025.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY