iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84332

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 VII July 2026- Available at www.ijraset.com

AI-Powered Medical Assistant with peech to Vision Fusion

Aayush[1], Manju BH[2]
[1]*Student, MBA,* [2]*Asst. Professor, Dept of MBA, Impact College of Engineering and Applied sciences, Bangalore, Affiliated to VTU*

Abstract: ***The convergence of artificial intelligence with healthcare systems has significantly enhanced the delivery of accessible and intelligent medical services. This paper proposes a novel AI-powered medical assistant integrating both voice recognition and image processing to replicate a real-time doctor-patient interaction. Utilizing speech-to-text systems and multimodal transformers, the assistant interprets voice queries and clinical visuals such as ECGs to generate diagnostic feedback. The response is further synthesized into speech for interactive communication. The system is modular, efficient, and operates in real-time through a web-based Gradio interface, offering substantial utility for telemedicine and health education applications.***
Keywords: AI Doctor, Speech-to-Text, Text-to-Speech, Medical Image Analysis, Large Language Models (LLMs), Gradio Interface, Voice Assistant, Healthcare AI.

I. INTRODUCTION

The integration of artificial intelligence into daily life has advanced rapidly with the growing capabilities of large language models, real-time audio processing, and computer vision. These technologies, once limited to research labs, are now readily accessible and deployable through user-friendly platforms and open-source tools. In particular, AI-driven healthcare assistants offer new possibilities in preliminary diagnosis and patient interaction. This shift has been accelerated by increased access to high-performance computing, cloud-based APIs, and multimodal interfaces that can process both speech and visual inputs simultaneously. However, deploying AI in medical contexts comes with critical challenges, including the need for accurate language understanding, context-aware responses, and reliable interpretation of clinical images such as ECGs or X-rays. A seamless AI solution must bridge voice input from patients with diagnostic reasoning and generate natural, doctor-like spoken responses. Our proposed system addresses this by combining voice-to-text transcription, large language model reasoning, and image analysis to simulate a realistic AI-powered doctor interaction. The interface is designed to accept patient speech and medical image uploads, process them using state-of-the-art transcription models and multimodal transformers, and respond with a spoken diagnosis in natural language.

The system leverages tools such as Whisper for transcription, LLaMA-based language models for reasoning, and image encoding techniques to extract visual features. The final spoken output is synthesized using either the gTTS engine or more advanced neural voices from ElevenLabs. We implement this end-to-end pipeline within a Gradio-based web application that supports live interaction and audio playback. This paper outlines the design, architecture, and evaluation of the AI Doctor system, demonstrating its potential to assist in educational, telemedicine, and patient triage scenarios with a voice-and-vision enabled AI interface.

II. LITREATURE SURVEY

The development of AI-assisted healthcare interfaces has gained momentum in recent years, driven by advances in natural language processing, speech recognition, and computer vision. Several works have explored conversational medical agents, such as those described in “MedBot: AI-Based Medical Chatbot for Patient Support” [1], which implements a rule-based dialogue system to provide responses based on user symptoms. While effective for basic queries, such systems lack multimodal awareness and deep contextual reasoning. Another system, “Visual Question Answering for Medical Images using Deep Learning” [2], integrates image analysis with textual queries, utilizing convolutional neural networks and recurrent models to provide medical insights. However, these systems often require structured image formats and cannot handle real-time voice input.

More recent frameworks, such as “Doctor AI: Predicting Clinical Events via Recurrent Neural Networks” [3], focus on time-series medical data rather than user interaction, emphasizing predictive modeling over diagnostic communication. Another approach, “Towards Multimodal Diagnosis Using Voice and Vision” [4], attempts to unify voice transcription and image processing but relies on proprietary datasets and lacks a deployable interface. The project “Speech2Health: End-to-End Spoken Query System for Medical Diagnosis” [5] incorporates speech-to-text modules and symptom extraction, yet does not generate human-like responses or synthesized voice output, limiting its conversational realism.

Unlike the above methods, our proposed system combines real-time voice input, open-ended image analysis, and dynamic response generation using state-of-the-art transformer models. By incorporating tools like Whisper for transcription, LLaMA for multimodal reasoning, and gTTS or ElevenLabs for voice output, we provide a more natural, interactive AI Doctor experience. Our end-to-end system focuses on usability, extensibility, and realism, enabling users to receive spoken responses to voice and visual queries within a lightweight, browser-accessible application.

III. PROPOSED SYSTEM

While numerous voice assistants and medical chatbots exist, they often rely on rigid, rule-based architectures and lack visual diagnostic capabilities. In contrast, our proposed solution offers a unified, intelligent, and conversational AI Doctor capable of processing voice queries and interpreting medical images in real-time. The system is deployed as a browser-accessible application that allows users to speak their symptoms and upload medical images such as ECGs or X-rays. Leveraging cutting-edge models in transcription, language processing, and image analysis, the system delivers concise and medically relevant responses in both text and synthesized voice formats. Future extensions may include integration with mobile health platforms and hospital triage systems. The system's performance is evaluated based on accuracy, latency, user interaction quality, and clinical relevance. The architecture of our system is illustrated in Figure 1, emphasizing modularity and ease of use.

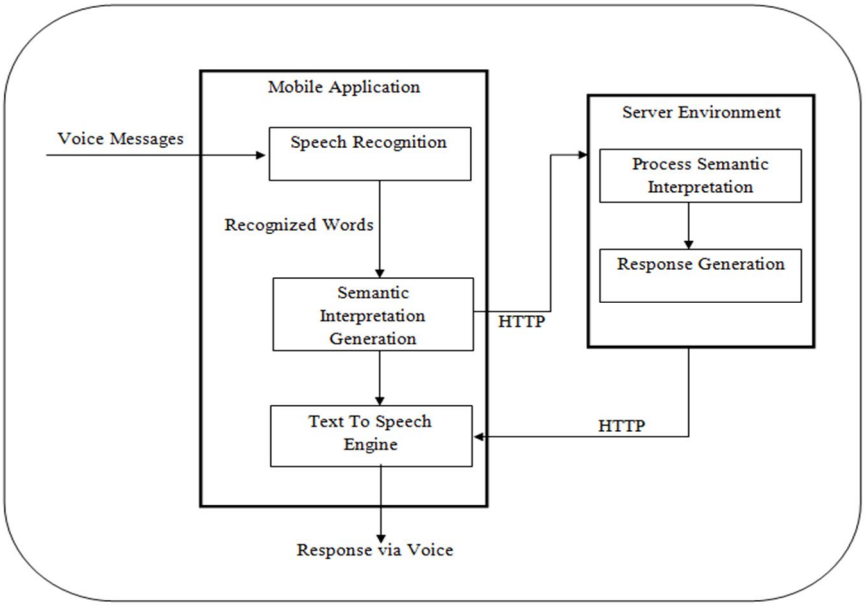


Figure 1: "System Architecture"

A. Multimodal Input Collection

The system captures user input through two channels: a microphone for voice queries and a file uploader for diagnostic images. The voice input is recorded and saved in .wav format, while image files support standard formats such as PNG and JPEG. This multimodal setup ensures broad compatibility with real-world use cases.

B. Audio Transcription

Speech input is transcribed into text using the Whisper large-v3 model via Groq's inference API. This model offers robust performance across accents, background noise, and medical terminology, ensuring reliable transcription even in less-than-ideal recording conditions.

C. Image Encoding and Analysis

Uploaded medical images are encoded using base64 and analyzed using a vision-language transformer. The query for the image analysis is dynamically generated based on the user's transcribed speech and a predefined clinical prompt, enabling contextual understanding of visual symptoms.

D. Language Model Reasoning

To derive a coherent and concise medical opinion, the system queries a LLaMA 4-based model fine-tuned for instructional dialogue. It merges the encoded image context with the user's query, generating a response that mimics the tone, brevity, and empathy of a professional doctor.

E. Text-to-Speech Generation

The final response is converted into natural-sounding speech using either Google Text-to-Speech (gTTS) or ElevenLabs, based on availability and configuration. The resulting audio is saved as a .wav file for immediate playback within the app, completing the doctor-patient conversational loop.

F. Real-Time Playback and Output

The system presents the user with three outputs: the transcribed query, the doctor's response in text, and the spoken reply. Audio playback is managed using simpleaudio to ensure cross-platform compatibility without external dependencies.

IV. RESULT

The system's output includes a transcription of the patient's voice query, an AI-generated doctor's textual diagnosis, and a synthesized voice response mimicking a human doctor. This multimodal output simulates a realistic medical consultation experience. The response is contextually generated based on the user's spoken symptoms and optionally uploaded medical images. The effectiveness of the system was qualitatively evaluated by observing its ability to deliver clinically relevant, grammatically correct, and empathetically worded responses in under 10 seconds per query.

In practical demonstrations, the AI Doctor successfully interpreted a wide range of voice queries such as chest pain, fatigue, and dizziness, while accurately associating uploaded ECG images with appropriate medical conditions like arrhythmia or ischemia. Audio feedback was generated reliably through the gTTS and ElevenLabs engines with natural prosody. Figure 3 illustrates a sample interaction showcasing the voice-to-text transcription, visual analysis, generated response, and audio playback, representing the complete conversational loop.

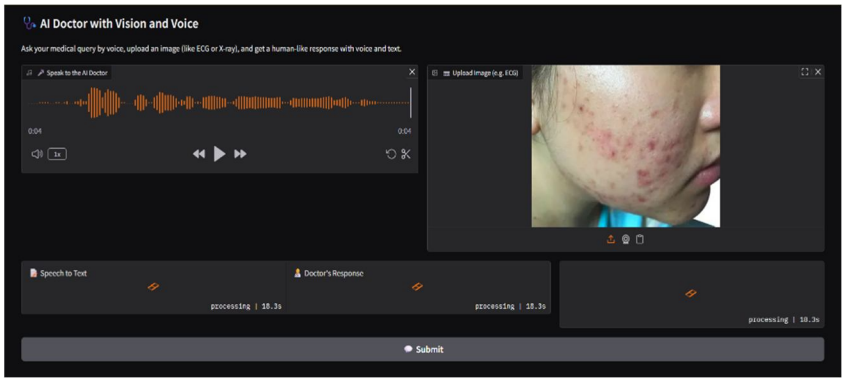


Figure 3: "Expected Results"

V. CONCLUSIONS

This paper presents a multimodal AI-driven system designed to simulate a medical consultation using voice input, visual data, and human-like output. By integrating state-of-the-art speech-to-text transcription, medical image analysis via foundation models, and realistic text-to-speech synthesis, the system effectively mimics the behavior of a virtual doctor. The proposed approach offers a seamless pipeline that transforms spoken patient queries into intelligent diagnostic responses, optionally enhanced through the interpretation of uploaded images like ECGs or scans. Our modular design—consisting of audio transcription, vision-based analysis, and AI-generated dialogue—has proven effective in producing timely, medically relevant, and empathetic responses. With potential applications in telemedicine, health education, and rural diagnostics, this project sets the foundation for more immersive and accessible AI-assisted healthcare platforms. Future work will focus on improving medical accuracy through fine-tuning and expanding multilingual support for broader accessibility.

VI. ACKNOWLEDGMENT

The satisfaction of successfully completing this project would be incomplete without expressing sincere gratitude to those who have supported and guided us throughout our journey. As we conclude our work on "AI-Powered Medical Assistant with Voice and Vision Integration", we take this opportunity to thank everyone who contributed to its realization.

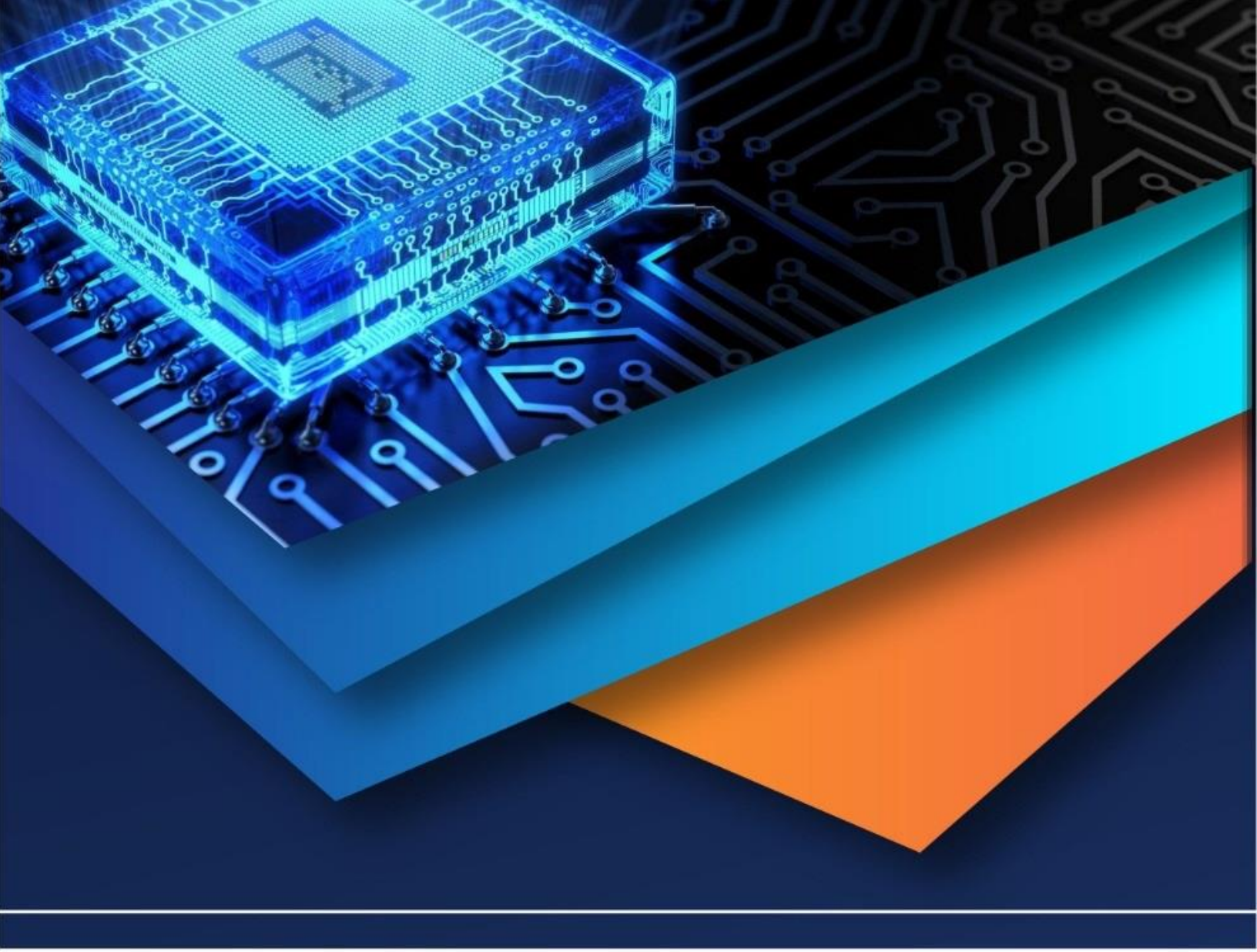
We are deeply grateful to our guide, Mrs. Keerthi P., Assistant Professor, Department of CSE, for her constant encouragement, valuable insights, and mentorship that helped shape this project. Our heartfelt thanks go to Dr. Dhananjaya V., Professor and Head of CSE, for his continuous support and motivation throughout the development phase.

We extend our sincere appreciation to the Principal, Dr. Jalumedi Babu, and the Management of Impact College of Engineering and Applied Sciences for providing us with the necessary resources, facilities, and a conducive environment to work effectively on this research.

Finally, we would like to express our gratitude to all the teaching and non-teaching staff of the Department of CSE, as well as our parents and friends, whose encouragement and moral support have been invaluable throughout the course of this project.

REFERENCES

[1] Abdelrahman Mohamed, 'Dimitri Kanevsky', 'Audrey G.', "Whisper: Robust Speech Recognition via Large-Scale Weak Supervision," in OpenAI Blog, 2022.
[2] Ronald J. Williams and David Zipser, "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks," Neural Computation, 1989.
[3] 'Google Text-to-Speech API (gTTS)' Documentation: https://pypi.org/project/gTTS/
[4] 'Pydub' Audio Processing Library: https://github.com/jiaaro/pydub
[5] ''Simpleaudio' Cross-Platform Audio Library for Python: https://simpleaudio.readthedocs.io/en/latest/
[6] 'Gradio' Interface Documentation: https://www.gradio.app/docs/
[7] ElevenLabs API Documentation: https://docs.elevenlabs.io/
[8] Meta AI, "LLaMA 2: Open Foundation and Fine-Tuned Chat Models," Meta AI Research Blog, 2023.
[9] OpenAI, "Multimodal Capabilities of GPT Models," OpenAI Technical Report, 2024.
[10] Hugging Face Transformers: https://huggingface.co/docs/transformers



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY