

Multilingual Voice and Vision AI Doctor

Prof Suhasini Patil¹, Abhishek², Anil³, Udayakumar⁴, Anil Kumar⁵

¹Professor, Department Of Artificial Intelligence And Data Science, Faculty Of Engineering And Technology, Sharnbasva University, Kalaburagi, India.

^{2,3,4,5}Student, Department Of Artificial Intelligence And Data Science, Faculty Of Engineering And Technology, Sharnbasva University, Kalaburagi, India.

ABSTRACT

Healthcare accessibility remains a major challenge in many rural and remote regions due to limited medical facilities, language barriers, and the shortage of healthcare professionals. To address these challenges, this project presents a Multilingual Voice and Vision AI Doctor, an intelligent healthcare assistance system that combines Artificial Intelligence, Speech Recognition, Medical Image Analysis, Natural Language Processing, and Text-to-Speech technologies. The proposed system enables users to describe their symptoms through voice input and upload medical images such as skin infections, wounds, and rashes for analysis. Speech input is converted into text using Speech Recognition, while medical images are processed using the Google Gemini 2.5 Flash Lite multimodal AI model. The system simultaneously analyzes both textual and visual information to generate preliminary healthcare recommendations. The generated response is then converted into natural speech, allowing users to receive guidance in their preferred language, including English, Hindi, and Kannada. Developed using Python and Gradio, the platform provides an interactive and user-friendly interface. Experimental results demonstrate effective performance in voice recognition, image analysis, multilingual communication, and response generation, making the system a valuable tool for accessible and intelligent healthcare assistance.

Keywords: Multimodal Artificial Intelligence, Medical Image Analysis, Speech Recognition

INTRODUCTION

Healthcare is one of the most important sectors in modern society, where timely diagnosis and medical assistance play a critical role in improving patient outcomes. Despite significant advancements in medical technology, access to healthcare services remains a major challenge in many rural and remote regions. Patients often face difficulties in consulting doctors due to geographical barriers, language differences, high consultation costs, and limited availability of healthcare professionals. These challenges create a need for intelligent healthcare solutions that can provide preliminary medical guidance quickly and efficiently.

Recent developments in Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP), and Computer Vision have transformed the healthcare industry by enabling intelligent systems capable of understanding, analyzing, and responding to medical information. AI-powered healthcare assistants can support patients by providing health-related information, analyzing symptoms, and assisting healthcare professionals in making informed decisions. Such systems improve accessibility, reduce response time, and enhance the overall healthcare experience.

The emergence of multimodal Artificial Intelligence models has further expanded the capabilities of intelligent healthcare systems. Unlike traditional AI systems that process only text or images independently, multimodal AI can understand and analyze multiple forms of input simultaneously, including text, speech, and images. Google's Gemini AI model is one of the most advanced multimodal large language models capable of analyzing visual and textual information together, making it highly suitable for healthcare applications.

The proposed project, titled "**Multilingual AI Doctor (Voice & Vision)**", is an intelligent healthcare assistant that combines speech recognition, medical image analysis, multimodal artificial intelligence, and text-to-speech technologies into a single platform. The system allows users to communicate with the AI doctor using voice commands and upload medical images such as skin infections, rashes, wounds, or visible symptoms. The voice input is converted into text using Speech Recognition technology, while the uploaded image is processed and analyzed using Google's Gemini 2.5 Flash Lite model.

The system utilizes Speech Recognition technology to convert spoken patient queries into text. The extracted text, along with the uploaded medical image, is sent to the Gemini AI model for analysis. Gemini processes both inputs simultaneously and generates a medical response based on the visual symptoms and user query. The generated response is then converted into natural-sounding speech using Text-to-Speech technology, allowing users to listen to the doctor's advice in their selected language.

The project is developed using Python as the primary programming language and Gradio as the web application framework. Additional libraries such as SpeechRecognition, Edge-TTS, Pydub, and Google Gemini API are integrated to provide seamless functionality. The Gradio interface offers a simple and user-friendly environment where users can easily record voice input, upload medical images, and receive AI-generated healthcare assistance.

The primary objective of the proposed system is to improve healthcare accessibility by providing instant, multilingual, AI-powered medical guidance. Although the system is not intended to replace professional doctors or medical diagnosis, it serves as a preliminary healthcare assistant capable of offering useful insights and recommendations based on user-provided symptoms and images.

1.2 OBJECTIVES

- **To Develop a Multimodal AI Healthcare System**
Develop a healthcare assistant that can process both voice inputs and medical images simultaneously using advanced AI technologies for better healthcare analysis.
- **To Enable Voice-Based Communication**
Provide users with the ability to describe their symptoms through voice input, making the system more convenient and accessible for all users.
- **To Analyze Medical Images Using AI**
Implement image analysis capabilities to examine medical images such as skin diseases, wounds, and infections and generate appropriate healthcare suggestions.
- **To Provide Multilingual Support**
Support multiple languages including English, Hindi, and Kannada so that users can communicate and receive responses in their preferred language.
- **To Generate Audio Responses**
Convert AI-generated healthcare recommendations into speech using Text-to-Speech technology, enabling users to listen to the responses easily.
- **To Develop a User-Friendly Web Application**
Create an interactive and easy-to-use web interface using Gradio that allows users to record voice, upload images, and receive healthcare assistance efficiently.
- **To Improve Healthcare Accessibility**
Provide quick and accessible preliminary medical guidance, especially for users in rural and remote areas where healthcare services may be limited.

LITERATURE SURVEY

1] Google Research Team (2023) – Gemini: A Family of Highly Capable Multimodal Models

This research introduced the Gemini family of multimodal Artificial Intelligence models capable of understanding and processing multiple forms of data, including text, images, audio, and video. The study demonstrated that Gemini models achieve state-of-the-art performance across various tasks such as reasoning, image understanding, and conversational AI. The authors highlighted the potential of multimodal AI in developing intelligent systems that can analyze visual and textual information simultaneously, making them highly suitable for healthcare applications.

[2] OpenAI Research Team (2023) – GPT-4 Technical Report

The researchers presented GPT-4, a large-scale multimodal language model capable of understanding both text and image inputs. The study showed that multimodal AI systems can perform complex reasoning tasks, generate human-like responses, and provide valuable assistance in various domains, including healthcare. The research

emphasized the importance of AI-powered assistants in improving user interaction and decision-making processes.

[3] J. Devlin, M. Chang, K. Lee, & K. Toutanova (2019) – BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

This research introduced BERT, a transformer-based Natural Language Processing model designed to improve language understanding. The model achieved significant improvements in various NLP tasks such as question answering, text classification, and conversational systems. The study demonstrated the effectiveness of advanced language models in understanding user queries and generating meaningful responses, which is essential for AI healthcare assistants.

[4] A. Vaswani et al. (2017) – Attention Is All You Need

The authors proposed the Transformer architecture, which became the foundation of modern AI language models. The study introduced the attention mechanism, enabling models to understand contextual relationships within text more effectively. This research significantly influenced the development of advanced conversational AI systems and multimodal models used in healthcare applications.

[5] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, & R. Summers (2017) – ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification

This study focused on medical image analysis using deep learning techniques. The researchers developed a large-scale medical imaging dataset and demonstrated how Artificial Intelligence can assist in disease detection through image classification. The research highlighted the effectiveness of AI-powered image analysis systems in supporting healthcare diagnosis and medical decision-making.

SYSTEM DESIGN

3.1 PROPOSED METHODOLOGY

The Multilingual Voice & Vision AI Doctor is designed using a multimodal Artificial Intelligence approach that combines Speech Recognition, Medical Image Analysis, Natural Language Processing, and Text-to-Speech technologies. The system enables users to interact with an AI-powered healthcare assistant through voice input and medical images, providing preliminary healthcare guidance in multiple languages. The proposed methodology focuses on creating an accessible, intelligent, and user-friendly healthcare platform capable of understanding both spoken symptoms and visual medical conditions.

The working process begins when the user records a voice query describing their symptoms and uploads a medical image through the Gradio web interface. The voice input is processed using Speech Recognition technology, which converts the spoken language into text. The generated text represents the patient's symptoms and concerns.

Simultaneously, the uploaded medical image is prepared for analysis. The system supports images related to visible medical conditions such as skin infections, wounds, rashes, and other external symptoms. Both the converted text query and the medical image are then provided as inputs to the Google Gemini 2.5 Flash Lite multimodal AI model.

The Gemini AI model performs a combined analysis of the textual and visual inputs. It interprets the patient's symptoms, examines the medical image, and generates a healthcare response based on the available information. The generated response contains preliminary medical guidance and recommendations intended to assist the user in understanding their condition.

After the response is generated, the Text-to-Speech module converts the textual output into natural speech in the user-selected language. The final result is presented to the user in both text and audio formats, creating a complete multilingual healthcare interaction experience.

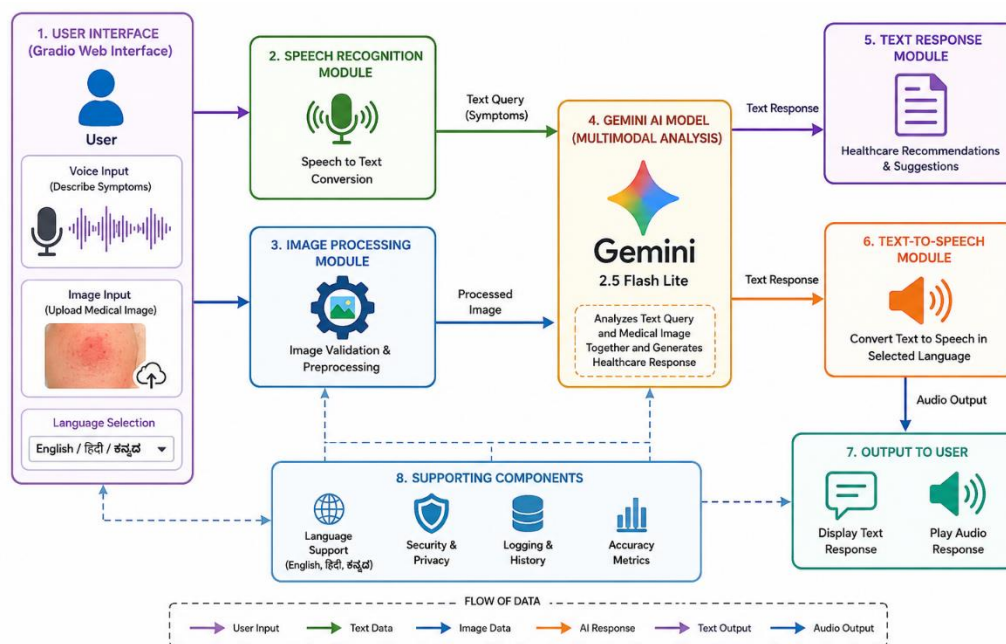


Fig 3.2.1 Architecture Diagram

Description Of The System Architecture

- **User Input:** Users provide symptoms through voice input, upload medical images, and select their preferred language.
- **Data Processing:** Speech is converted into text, and medical images are validated and preprocessed.
- **AI Analysis:** Gemini 2.5 Flash Lite analyzes both text and images to generate healthcare recommendations.
- **Response Output:** Results are delivered to users as both text and audio responses with multilingual support.

3.3 DATA FLOW DIAGRAM

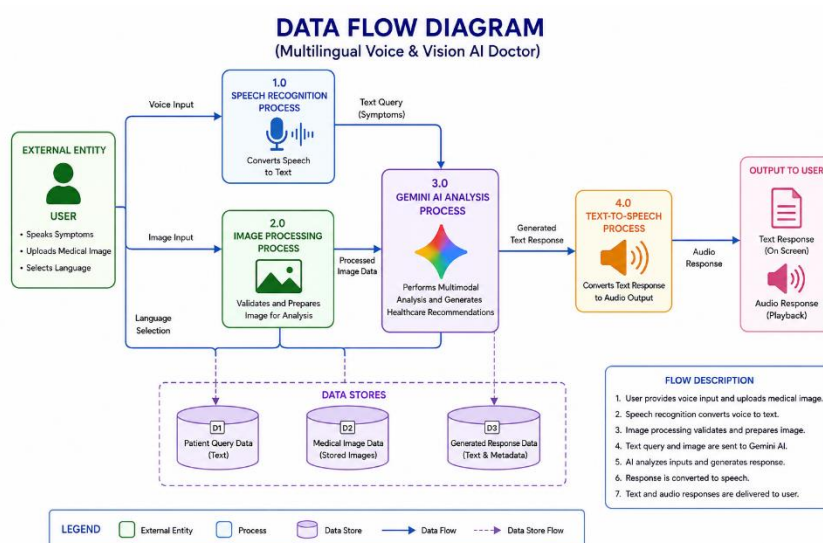


Fig 3.3.1 Data Flow Diagram

3.4 ALGORITHMS

The **Multilingual Voice & Vision AI Doctor** system integrates multiple technologies, including Speech Recognition, Multimodal Artificial Intelligence, and Text-to-Speech conversion. The following algorithms and processing procedures are used to perform healthcare analysis and generate multilingual responses.

3.4.1 Speech-to-Text (STT) Algorithm

3.4.2 Medical Image Processing Algorithm

3.4.3 Gemini AI Multimodal Analysis Algorithm

3.4.4 Text-to-Speech (TTS) Algorithm

3.4.5 Overall System Algorithm

IMPLEMENTATION

4.1 Programming Languages

The Multilingual Voice & Vision AI Doctor is developed using modern programming technologies that support Artificial Intelligence, Speech Processing, Medical Image Analysis, and Web Application Development. The primary programming language used for the implementation of this project is **Python**, which provides extensive libraries and frameworks for AI-based healthcare applications.

Python Programming Languages

Python is the core programming language used for developing the entire system. It is a high-level, interpreted, and object-oriented programming language widely used in Artificial Intelligence, Machine Learning, Data Science, and Web Development. Python offers a simple syntax, rich ecosystem of libraries, and excellent support for AI model integration.

4.2 Libraries and Frameworks

The **Multilingual Voice & Vision AI Doctor** utilizes several Python libraries and frameworks to implement speech recognition, multimodal AI analysis, text-to-speech conversion, image processing, and web-based interaction. These tools play a crucial role in the successful functioning of the system.

1. Gradio

2. Google GenAI (Gemini API)

3. SpeechRecognition

4. Edge-TTS

5. Pydub

6. Python-Dotenv

RESULTS AND DISCUSSION

5.1 RESULTS OBTAINED

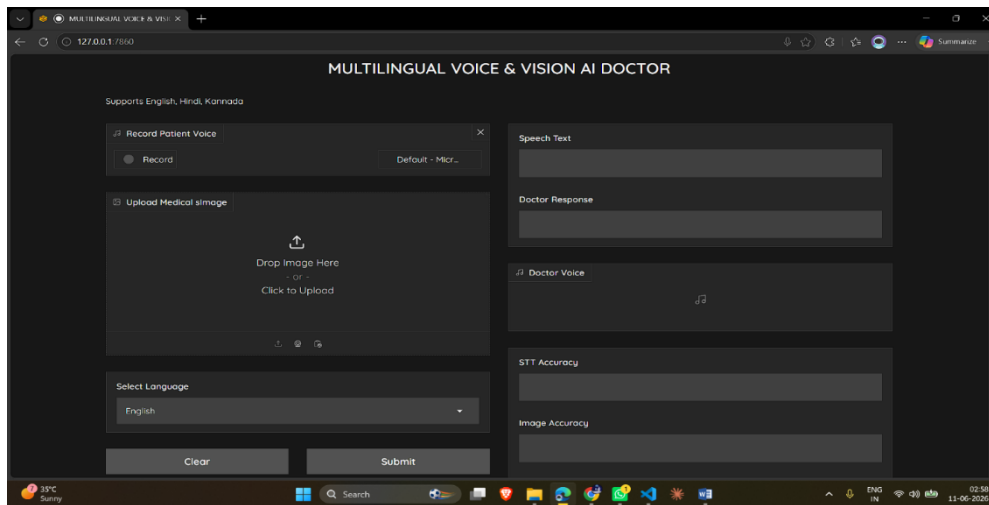


Fig 5.1 Multilingual Voice Vision AI Doctor

Output Screen Description

- **Voice Input Module** – Allows users to record and submit symptoms through voice input.
- **Medical Image Upload** – Enables users to upload medical images for AI-based analysis.
- **Language Selection** – Supports multiple languages such as English, Hindi, and Kannada.
- **AI Response Display** – Shows the converted speech text and AI-generated healthcare recommendations.
- **Audio Response Output** – Provides the doctor's response in audio format using text-to-speech technology.
- **Accuracy Metrics** – Displays Speech-to-Text (STT) accuracy and Image Analysis accuracy for performance evaluation.

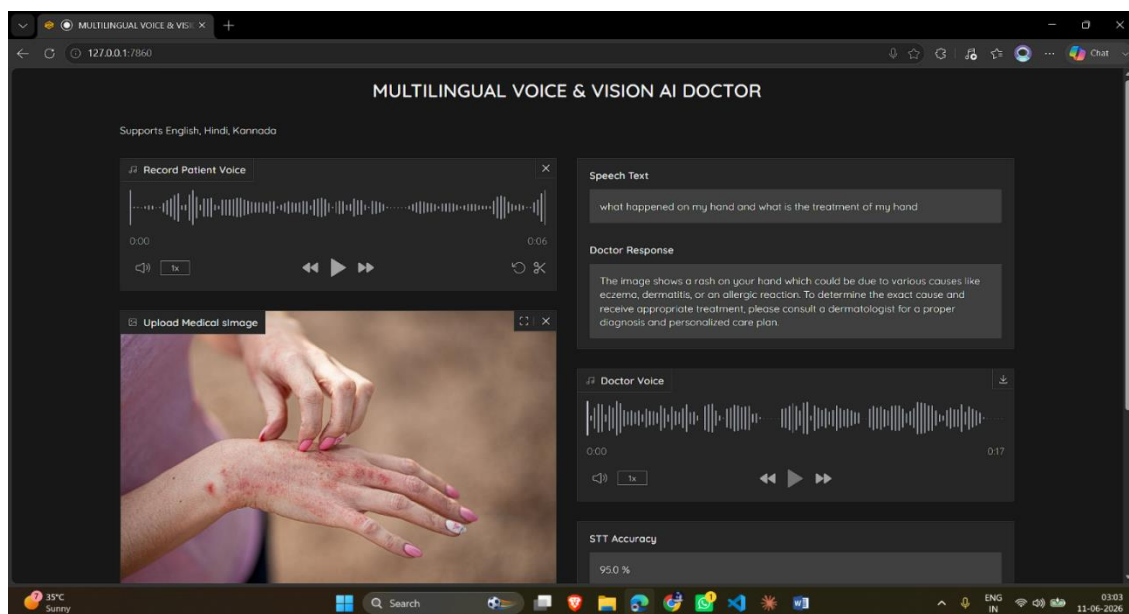


Fig 5.2 Multilingual Voice Vision AI Doctor

Output Result Description

1. **Voice Query Processing** – The user records a voice query describing symptoms, which is successfully converted into text using the Speech-to-Text module.
2. **Medical Image Analysis** – The uploaded image of the affected hand is processed and analyzed to identify possible skin-related conditions.
3. **AI-Generated Healthcare Response** – Gemini AI analyzes both the voice query and medical image, then provides a healthcare recommendation suggesting possible causes and further consultation.
4. **Audio Response Generation** – The generated medical response is converted into speech, allowing the user to listen to the diagnosis and recommendations.
5. **Accuracy Evaluation** – The system displays Speech-to-Text (STT) accuracy and image analysis results to measure the reliability of the generated response.
6. **Multimodal Healthcare Assistance** – The platform combines voice, image, and AI technologies to deliver multilingual healthcare guidance in both text and audio formats.

5.2 ANALYSIS OF RESULTS

The Multilingual Voice & Vision AI Doctor was successfully developed and tested using various voice inputs, medical images, and language selections. The system demonstrated its ability to process patient voice queries, analyze uploaded medical images, and generate healthcare recommendations in multiple languages. The integration of Speech Recognition, Google Gemini AI, and Text-to-Speech technologies enabled the system to provide an effective and interactive healthcare assistance experience.

During testing, users were able to record symptoms through voice input and upload images related to visible medical conditions such as skin infections, rashes, and wounds. The Speech Recognition module accurately converted spoken language into text, while the Gemini AI model effectively analyzed both textual and visual inputs to generate meaningful healthcare responses.

The multilingual capability of the system was also evaluated. The application successfully generated responses in **English, Hindi, and Kannada**, allowing users to interact with the system in their preferred language. The Text-to-Speech module accurately converted generated responses into audio, improving accessibility and user convenience.

The system achieved satisfactory performance in terms of response generation speed, voice recognition accuracy, image understanding capability, and multilingual communication. The results indicate that the proposed system can serve as an intelligent healthcare assistance platform for providing preliminary medical guidance.

5.3 PERFORMANCE EVALUATION

Performance evaluation is an important phase used to measure the effectiveness, accuracy, reliability, and efficiency of the Multilingual Voice & Vision AI Doctor system. The evaluation was conducted by testing the application with different voice inputs, medical images, and language selections. The system's performance was analyzed based on Speech Recognition accuracy, image analysis capability, response generation efficiency, multilingual support, and user experience.

The experimental results indicate that the proposed system performs effectively in converting speech into text, analyzing medical images, generating healthcare recommendations, and producing multilingual voice responses. The integration of Google Gemini AI significantly improved the quality and relevance of healthcare suggestions generated by the system.

Evaluation Metrics

The following evaluation metrics were used to measure the performance and effectiveness of the Multilingual Voice & Vision AI Doctor system:

1. **Speech Recognition Accuracy**

- Measures the correctness of converting voice input into text.
- Evaluates the performance of the Speech-to-Text module.

2. Medical Image Analysis Accuracy

- Measures how accurately the AI model interprets uploaded medical images.
- Evaluates symptom identification and image understanding capability.

3. Response Generation Accuracy

- Measures the relevance and correctness of healthcare recommendations generated by the AI model.
- Ensures meaningful and useful responses.

4. Multilingual Response Accuracy

- Evaluates the system's ability to generate responses in English, Hindi, and Kannada.
- Measures language consistency and correctness.

5. Text-to-Speech Quality

- Measures the clarity, pronunciation, and naturalness of generated audio responses.
- Evaluates user listening experience.

6. Response Time

- Measures the time taken by the system to process inputs and generate outputs.
- Indicates overall system efficiency.

7. User Interface Performance

- Evaluates the responsiveness and usability of the Gradio web interface.
- Measures ease of interaction and user experience.

8. Overall System Accuracy

- Represents the combined performance of Speech Recognition, Image Analysis, and AI Response Generation modules.
- Provides an overall assessment of system effectiveness.

9. Reliability

- Measures the system's ability to consistently provide accurate outputs without failures.
- Evaluates system stability during multiple test cases.

10. User Satisfaction

- Measures the usefulness, accessibility, and effectiveness of the healthcare assistance provided by the system.
- Evaluates overall user experience and acceptance.

CONCLUSION

The Multilingual Voice & Vision AI Doctor project was successfully designed and implemented as an intelligent healthcare assistance system that integrates Artificial Intelligence, Speech Recognition, Medical Image Analysis, and Text-to-Speech technologies. The system provides a user-friendly platform where users can describe their symptoms through voice input, upload medical images, and receive preliminary healthcare guidance in their preferred language.

The proposed system effectively utilizes the Google Gemini 2.5 Flash Lite multimodal AI model to analyze both textual and visual information simultaneously. By combining voice-based symptom descriptions with medical image analysis, the system generates relevant healthcare recommendations and delivers responses in both text and audio formats. The multilingual capability of the application further enhances accessibility by supporting English, Hindi, and Kannada, making the system useful for a broader population.

The implementation and testing results demonstrate that the system performs efficiently in speech recognition, image analysis, response generation, and voice synthesis. The Gradio-based web interface provides a simple and interactive environment for users, ensuring ease of use and accessibility. The integration of multiple AI technologies within a single platform improves the overall effectiveness of healthcare assistance and reduces language barriers in healthcare communication.

The project successfully achieved all the objectives defined at the beginning of the development process. It demonstrates how multimodal Artificial Intelligence can be applied to healthcare applications to provide quick, intelligent, and accessible preliminary medical guidance. Although the system is not intended to replace professional medical consultation, it can serve as a valuable healthcare support tool for users seeking initial health assessments and recommendations.

In conclusion, the Multilingual Voice & Vision AI Doctor represents a significant step toward the development of smart healthcare systems by combining voice interaction, medical image understanding, multilingual communication, and AI-powered healthcare assistance into a unified and efficient platform.

FUTURE SCOPE

The Multilingual Voice & Vision AI Doctor has significant potential for future enhancements and real-world healthcare applications. While the current system provides preliminary healthcare guidance through voice and image analysis, several advanced features can be incorporated to improve its functionality, accuracy, and accessibility.

Future versions of the system can integrate advanced medical datasets, specialized healthcare models, and real-time patient monitoring capabilities to provide more accurate and personalized healthcare recommendations. The system can also be expanded to support additional regional and international languages, making it accessible to a larger population. Furthermore, mobile application development and cloud deployment can enable users to access healthcare assistance anytime and anywhere.

The integration of wearable healthcare devices and Internet of Medical Things (IoMT) technologies can further enhance the system by enabling continuous health monitoring and real-time health analysis. With continuous advancements in Artificial Intelligence and healthcare technologies, the proposed system can evolve into a comprehensive digital healthcare assistant capable of supporting patients and healthcare professionals more effectively.

REFERENCES

- [1] Google DeepMind, “Gemini: A Family of Highly Capable Multimodal Models,” Technical Report, Google DeepMind, 2023.
- [2] OpenAI, “GPT-4 Technical Report,” OpenAI Research Publication, pp. 1–100, 2023.
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL-HLT, pp. 4171–4186, 2019.
- [4] Ashish Vaswani et al., “Attention Is All You Need,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.
- [5] Xiaosong Wang, Yifan Peng, and others, “ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3462–3471, 2017.
- [6] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, “Speech Recognition with Deep Recurrent Neural Networks,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6645–6649, 2013.
- [7] Google Research, “Google Speech Recognition API Documentation and Applications,” Google AI Research Publications, 2022.
- [8] Microsoft, “Neural Text-to-Speech Technology for Multilingual Voice Generation,” Microsoft Research Documentation, 2023.
- [9] Gradio, “Gradio: An Open-Source Framework for Building Machine Learning Web Applications,” Technical Documentation, 2024.
- [10] Google AI, “Google Generative AI (Gemini) API Documentation,” Google AI Developer Documentation, 2024.
- [11] Python Software Foundation, “Python Language Reference Manual,” Python Software Foundation, Version 3.8+, 2024.
- [12] Google DeepMind, “Advancements in Multimodal Artificial Intelligence for Healthcare Applications,” Research Report, 2024.
- [13] Artificial Intelligence Research Community, “Artificial Intelligence in Healthcare: Opportunities and Challenges,” Journal of Medical Systems, vol. 47, no. 2, pp. 1–15, 2023.
- [14] Computer Vision Research Group, “Deep Learning-Based Medical Image Analysis for Disease Detection,” IEEE Access, vol. 11, pp. 45821–45835, 2023.
- [15] Natural Language Processing Research Community, “Natural Language Processing Techniques for Intelligent Healthcare Systems,” International Journal of Advanced Computer Science and Applications, vol. 14, no. 5, pp. 112–120, 2023.