International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

AI DOCTOR – VOICE & VISION

Mr. Subrat Kumar Behera
Student, Dept. of CSE,
GIFT Autonomous, Bhubaneswar

Mr. Smruti Ranjan Satapathy
Student, Dept. of CSE,
GIFT Autonomous, Bhubaneswar

Prof. Dr. Jagannath Ray
Project Guide, Dept. of CSE,
GIFT Autonomous, Bhubaneswar

Abstract— **The rapid growth of AI-driven digital health platforms and increasing complexity of healthcare accessibility challenges have created significant demand for intelligent, multimodal healthcare assistants capable of delivering real-time medical guidance. Traditional healthcare systems often fail to bridge the gap between patients and timely medical support, relying on manual processes that are slow and limited in scope. This paper presents the design and implementation of an AI-Powered Multimodal Healthcare Assistant, AI Doctor – Voice & Vision, which leverages Multi-Agent AI architecture and Retrieval-Augmented Generation (RAG) to automate end-to-end preliminary healthcare guidance. The system integrates a Voice Processing Agent, NLP Agent, Computer Vision Agent, and Disease Prediction Agent, orchestrated through a sequential pipeline. The NLP Agent processes symptoms extracted from voice or text input. The Computer Vision Agent analyses uploaded medical images using CNN models for disease detection. The Disease Prediction Agent synthesizes findings and generates healthcare suggestions via RAG-enhanced knowledge retrieval. A Gradio-based frontend provides voice input, image upload, diagnosis results, and text-to-speech responses. Results demonstrate effective multimodal interaction with approximately 4-second end-to-end pipeline execution and 89% average RAG retrieval relevance.**

Keywords— **Multi-Agent AI, Retrieval-Augmented Generation, Speech Recognition, Computer Vision, NLP, Deep Learning, Gradio, Healthcare Assistant, Disease Prediction, TensorFlow.**

I. INTRODUCTION

A. Background

In the modern digital era, Artificial Intelligence has become one of the most rapidly growing technologies across sectors including education, banking, transportation, and healthcare. AI-based healthcare systems are enabling faster diagnosis, medical prediction, virtual assistance, and continuous health monitoring by combining Machine Learning, Deep Learning, Natural Language Processing (NLP), and Computer Vision [1].

The increasing penetration of smartphones and internet connectivity has made it feasible to deploy intelligent healthcare services accessible from any location. Despite major advances in medical science, significant populations — particularly in rural and remote areas — continue to face barriers including shortage of doctors, high consultation costs, long waiting times, and lack of healthcare awareness. Delayed medical guidance frequently results in severe health complications that early intervention could have avoided [2].

Traditional healthcare systems depend primarily on in-person interaction, and existing digital health applications mostly support only text-based communication without leveraging voice interaction, image analysis, or intelligent disease prediction. To overcome these limitations, AI Doctor – Voice & Vision develops an integrated healthcare assistant combining Speech Recognition, NLP, Computer Vision, and Multi-Agent AI to provide preliminary healthcare guidance through voice, text, and image input.

B. Problem Statement

Healthcare accessibility continues to be a major challenge in many regions, especially in rural and underdeveloped areas. Many people do not receive timely medical support due to shortage of doctors, high consultation costs, long waiting times, and lack of nearby healthcare facilities. Existing healthcare applications provide only basic text-based support and do not effectively utilize modern AI technologies such as voice interaction, image analysis, and intelligent disease prediction [3]. Most systems lack the capability to understand symptoms through natural conversation or analyse visible medical conditions through images. There is a clear need for an intelligent system providing instant preliminary medical guidance, supporting voice, text, and image-based communication, improving accessibility for rural and remote users while enabling early disease detection.

C. Motivation

The motivation behind this project stems from the increasing need for accessible, intelligent, and efficient healthcare solutions. The lack of healthcare facilities in rural areas, the shortage of qualified doctors, and the importance of early disease detection create a strong imperative for AI-powered healthcare assistants available anytime and anywhere. Advances in AI technologies — specifically Machine Learning, NLP, Speech Recognition, and Computer Vision — have made it feasible to create systems capable of understanding human speech, analysing images, and generating intelligent healthcare responses. The growing availability of open-source frameworks and widespread smartphone use further inspire the creation of digital healthcare platforms [4].

D. Objectives

The primary objectives of the AI Doctor – Voice & Vision platform are as follows:

- To design and develop an AI-based multimodal healthcare assistant supporting voice, text, and image input.
- To implement Speech Recognition for voice-based symptom analysis and NLP for understanding user queries.
- To integrate Computer Vision and Deep Learning for image-based disease detection.
- To deploy a Multi-Agent pipeline for symptom analysis, disease prediction, and recommendation generation.
- To implement Retrieval-Augmented Generation (RAG) for accurate, data-grounded healthcare responses.
- To develop a Gradio-based interactive frontend with text-to-speech voice output functionality.

Received: 20-04-2026 | Accepted: 28-05-2026 | Published: 04-06-2026

International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

- To demonstrate the practical application of AI technologies in preliminary healthcare assistance.

E. Scope of the Project

The scope of this project includes the design and implementation of a full-stack AI-based healthcare assistant capable of processing voice, text, and image inputs and delivering structured healthcare guidance through a multi-agent pipeline and interactive dashboard. The system focuses on preliminary symptom analysis, disease prediction, and healthcare awareness, leveraging RAG and LLM-based reasoning while keeping the system practical and deployable without heavy infrastructure requirements.

II. LITERATURE REVIEW

A. Existing Healthcare Analytics Systems

Traditional healthcare analytics platforms such as EHR systems, symptom checkers, and telemedicine tools have become standard in healthcare delivery. These platforms provide useful functionalities but rely heavily on manual input and human interpretation. Text-based chatbots and voice assistants have gained popularity but typically focus on a single interaction modality and lack integrated multimodal capabilities combining voice, image analysis, and intelligent disease prediction simultaneously [5].

B. AI in Healthcare Analytics

Artificial Intelligence has emerged as a transformative technology in healthcare analytics, enabling systems to analyse vast patient data and detect disease patterns that are difficult to identify manually. Machine Learning improves disease prediction accuracy while Deep Learning models detect abnormalities in medical images — X-rays, MRI scans, and skin disease photographs — with clinically significant accuracy. NLP enables healthcare chatbots to understand and respond to medical queries naturally. Research consistently demonstrates that AI-driven healthcare systems improve insight quality and enhance early disease detection [6].

C. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation, introduced by Lewis et al. (2020), enhances LLM outputs by grounding them in retrieved context from an external knowledge base. Unlike fine-tuned models, RAG systems retrieve relevant documents at inference time, significantly reducing hallucination and improving factual accuracy. In healthcare applications, RAG systems retrieve disease-related data, treatment suggestions, and medical guidelines from curated knowledge bases, ensuring AI-generated healthcare recommendations are anchored to verified medical information rather than generic model outputs [7].

D. Multi-Agent AI Frameworks

Multi-agent systems involve multiple AI agents collaborating to perform distinct tasks. Research consistently demonstrates that agent specialization produces higher-quality outputs than single-agent approaches on complex tasks. In AI healthcare systems, multi-agent frameworks allow separate agents to handle voice processing, NLP symptom analysis, image analysis, disease prediction, and recommendation generation independently while communicating via a structured pipeline. This modular approach improves scalability, flexibility, and overall performance [8].

E. Research Gap and Proposed Contribution

From the reviewed literature, there is a clear gap in systems that combine RAG-based medical knowledge retrieval with multi-agent analytical pipelines, real-time multimodal interaction supporting voice, text, and image input simultaneously, voice-enabled accessible healthcare interaction for elderly and rural users, and a deployable full-stack architecture with minimal infrastructure requirements. To address this, this project proposes AI Doctor – Voice & Vision — an integrated platform combining Speech Recognition, NLP, Computer Vision, Deep Learning, RAG-based knowledge retrieval, and a multi-agent pipeline deployed via a Gradio-based interactive frontend.

III. SYSTEM OVERVIEW

A. Proposed System

AI Doctor – Voice & Vision is a web-based healthcare assistant that allows users to record voice symptoms, type healthcare queries, or upload images of visible medical conditions. Users receive NLP-based symptom analysis, AI-powered disease predictions, and actionable healthcare suggestions with precautionary recommendations. They can also listen to voice responses via text-to-speech output. The system provides a Gradio dashboard with a voice recording interface, image upload panel, speech-to-text output display, doctor response panel, and voice response playback.

B. System Architecture

The system follows a six-layer pipeline architecture: (1) Input Layer — voice, text, and image inputs captured through the Gradio frontend; (2) Processing Layer — Speech Recognition converts voice to text; (3) NLP Layer — symptom extraction and intent recognition applied to text input; (4) Vision Layer — Computer Vision and CNN models analyse uploaded medical images; (5) Agent Layer — multi-agent pipeline executes Disease Prediction and Recommendation Agents; (6) Presentation Layer — Gradio dashboard displays results with text-to-speech voice response output.

C. Technology Stack

Layer	Technology	Purpose
Input	Gradio	Voice, text & image UI
Voice	SpeechRecognition / Whisper	Speech-to-text conversion
NLP	spaCy / NLTK	Symptom extraction & NLP
Computer Vision	OpenCV / CNN (TensorFlow)	Image disease detection
Prediction	Scikit-learn / PyTorch	Disease prediction models
RAG	LangChain / ChromaDB	Medical knowledge retrieval
Voice Output	gTTS / pyttsx3	Text-to-speech responses
Database	MongoDB / SQLite	Healthcare data storage

Table I: Technology Stack of AI Doctor – Voice & Vision

D. Key Functional Modules

Module	Description	Technology
Voice Agent	Converts speech to text	SpeechRecognition, Whisper

International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

NLP Engine	Symptom & intent extraction	spaCy, NLTK
Computer Vision	Medical image analysis	OpenCV, CNN, TensorFlow
Disease Prediction	Predicts disease from symptoms	Scikit-learn, PyTorch
RAG Module	Retrieves medical knowledge	ChromaDB, LangChain
Recommendation	Generates healthcare guidance	LLM + RAG
TTS Module	Converts response to voice	gTTS, pyttsx3
Dashboard	Interactive user interface	Gradio, HTML/CSS

Table II: Key Functional Modules

E. Advantages of Proposed System

The AI Doctor – Voice & Vision platform eliminates the need for physical hospital visits for basic preliminary symptom checking through full multimodal pipeline automation. It provides data-grounded healthcare suggestions via RAG, reducing hallucination and increasing medical response reliability. The modular multi-agent design ensures quality at each analytical stage. The platform is accessible to elderly and non-technical users through voice interaction. The system is deployable on standard hardware without heavy GPU infrastructure requirements.

IV. METHODOLOGY

A. Overall Workflow

The system operates through structured steps. First, the user provides input through voice recording, text entry, or medical image upload via the Gradio interface. Second, Speech Recognition converts audio to text. Third, the NLP Engine preprocesses text, extracts symptoms, and identifies healthcare intent. Fourth, Computer Vision analyses uploaded medical images using CNN models for disease pattern detection. Fifth, the Disease Prediction Agent synthesizes symptom and image analysis data and generates disease predictions. Sixth, the RAG module retrieves relevant medical knowledge from ChromaDB. Seventh, the Recommendation Agent generates healthcare suggestions and precautions. Results are displayed in the chat panel and delivered as voice response via text-to-speech.

B. Data Collection and Handling

The system utilises healthcare datasets for training and testing Machine Learning models. Disease symptoms datasets provide structured symptom-disease mappings for classification model training. Medical image datasets include labelled skin disease images, wound classifications, and visible infection patterns used to train CNN models. Healthcare knowledge documents are preprocessed into text chunks, embedded using sentence-transformer models, and stored in ChromaDB for RAG retrieval. The data handling process includes input validation, missing value handling, data normalisation, image resizing and enhancement, and serialization of records for embedding and model training.

C. RAG Implementation

The RAG system forms the medical knowledge retrieval backbone. During indexing, medical knowledge documents are split into text chunks, embedded using the all-MiniLM-L6-v2 sentence-transformer model, and stored in ChromaDB with source metadata. During retrieval, the user query and extracted symptoms are embedded, cosine similarity search is performed against ChromaDB collections, and the top-k most semantically relevant medical records are returned as structured context for the Recommendation Agent prompt.

Score Range	Similarity	Action
> 0.85	High relevance	Primary context
0.70 – 0.85	Medium relevance	Secondary context
< 0.70	Low relevance	Exclude from prompt

Table III: RAG Retrieval Scoring Criteria

D. Multi-Agent Pipeline

The multi-agent pipeline consists of four specialised agents operating sequentially. The Voice Processing Agent receives audio input, applies noise reduction, and outputs transcribed text. The NLP Agent preprocesses transcribed or typed text, applies named entity recognition for symptom identification, and outputs a structured symptom JSON. The Computer Vision Agent receives the uploaded medical image, applies preprocessing, and runs the trained CNN classification model to identify disease patterns with confidence scores. The Disease Prediction Agent receives the combined symptom JSON and image classification output, queries the RAG module for relevant medical context, and generates disease predictions with healthcare suggestions and precautionary recommendations.

E. Backend API and Frontend

The backend handles system logic, AI model communication, and disease prediction processing, implemented in Python using Flask or FastAPI. Backend APIs receive requests from the Gradio frontend, process healthcare data, and communicate with AI modules. The Gradio frontend uses a modular layout with voice recording interface, image upload panel, speech-to-text display, doctor response panel, and voice response playback. Average token consumption per recommendation pipeline execution is approximately 1,800 tokens with response time under 2 seconds per agent call.

V. SYSTEM DESIGN

A. Use Case Diagram

The Use Case Diagram illustrates interaction between users and the AI Doctor – Voice & Vision platform. Both General Users and Elderly Users can record voice symptoms, enter text symptoms, upload medical images, view disease predictions, receive healthcare recommendations, and play voice responses. The Administrator role has additional privileges for managing medical datasets, retraining AI models, and system configuration. The system connects to an external Medical Knowledge Base for RAG-based information retrieval during recommendation generation.

B. ER Diagram

The ER Diagram represents data entities and relationships. The User entity (user_id, name, age) initiates sessions in the Interaction entity (interaction_id, user_id FK, timestamp, input_type). Each interaction includes a VoiceInput record (voice_id, audio_data, transcribed_text) and/or MedicalImage record (image_id, image_path). Symptom records are extracted from voice and text input. The DiseasePrediction

International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

entity (prediction_id, disease_name, confidence_score) links to Recommendation records (recommendation_id, precautions, treatment_suggestion). The MedicalKnowledge entity (knowledge_id, content, embedding) is referenced during RAG retrieval.

C. Sequence Diagram

The Sequence Diagram illustrates step-by-step interaction during a user session. The user provides voice or text input and optionally uploads a medical image via the Gradio UI. The Voice Processing Module converts audio to text and forwards it to the NLP Engine. The NLP Engine extracts symptoms and passes them to the Disease Prediction Agent. If an image is uploaded, the Computer Vision Module processes it concurrently. The Prediction Agent queries the RAG Module, which retrieves relevant records from ChromaDB. Results are displayed in the Gradio UI and converted to voice output via text-to-speech.

D. System Architecture Diagram

The system architecture follows a layered pipeline structure. The browser-based Gradio UI captures voice, text, and image input. The Python Backend invokes the Voice Processing Module and NLP Engine for text-based input, while the Computer Vision Module handles image analysis concurrently. Processed outputs feed into the Multi-Agent Orchestrator managing sequential execution of the Disease Prediction and Recommendation Agents. The RAG Module connects to ChromaDB for medical knowledge retrieval. Responses flow back through the pipeline to produce diagnosis output and voice responses via text-to-speech.

VI. IMPLEMENTATION

A. Voice Processing Implementation

The voice processing module is implemented using Python's SpeechRecognition library and OpenAI Whisper for high-accuracy transcription. It reads audio recorded via the Gradio microphone interface, applies noise reduction using the noisereduce library, and transcribes speech to text for downstream NLP processing. Whisper is used for improved accuracy in noisy environments. The pipeline runs in under 1 second for audio clips up to 10 seconds in length.

B. NLP Implementation

The NLP module is implemented using spaCy and NLTK. It receives transcribed or typed text, applies tokenisation, stopword removal, and lemmatisation. Named Entity Recognition with a custom healthcare entity model identifies symptoms, body parts, and medical terms. Intent classification determines whether the user is describing symptoms, asking for information, or requesting image analysis. Extracted symptom entities are serialised as a structured JSON for the Disease Prediction Agent. Processing time is under 200 milliseconds per query.

C. Computer Vision Implementation

The Computer Vision module is implemented using OpenCV for image preprocessing and a CNN model trained using TensorFlow/Keras for image classification. Input images are resized to 224×224 pixels, normalised, and augmented during training. The CNN architecture follows a ResNet-inspired design trained on dermatological and wound classification datasets. The module outputs a disease class label with a confidence score. Image preprocessing and classification execute in under 1.5 seconds per image.

D. Multi-Agent Pipeline Implementation

Agent	Input	Output
Voice Processing	Audio recording	Transcribed text
NLP Agent	Text / transcription	Symptom JSON
Computer Vision	Medical image	Disease class + score
Disease Prediction	Symptom JSON + image	Predictions + advice

Table IV: Multi-Agent Pipeline Implementation

Each agent executes independently and communicates results via a shared state dictionary managed by the backend orchestrator. Error handling includes fallback responses for API failures and confidence thresholding to avoid low-quality predictions being presented to users.

E. Challenges and Solutions

Challenge	Solution
Background noise in voice input	noisereduce preprocessing + Whisper ASR
Low image quality affecting CNN	Image enhancement & contrast normalisation
RAG hallucination in medical context	Confidence thresholding + score filtering
Session state not persisting	Backend state dictionary management
Large image embedding time	Batch embedding with progress indicator

Table V: Implementation Challenges and Solutions

VII. RESULTS AND ANALYSIS

A. Functional Results

Module	Test Case	Status
Voice Processing	Speech-to-text conversion	Pass
NLP Engine	Symptom extraction from text	Pass
Computer Vision	Skin disease classification	Pass
Disease Prediction	Multi-symptom prediction	Pass
RAG Retrieval	Medical knowledge retrieval	Pass
Recommendation	Precaution & advice generation	Pass
TTS Output	Text-to-speech voice response	Pass
Dashboard	Image upload & voice recording	Pass

Table VI: Functional Test Results

B. RAG Retrieval Accuracy

Query	Score	Accuracy
Fever, cough, fatigue	0.91	High
Skin rash with redness	0.87	High

Received: 20-04-2026 | Accepted: 28-05-2026 | Published: 04-06-2026

International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

Chest pain & breathlessness	0.83	High
Joint pain in elderly	0.76	Med-High

Table VII: RAG Retrieval Accuracy

C. Agent Pipeline Performance

Agent	Avg. Time	Quality
Voice Processing	< 1.0 sec	High transcription accuracy
NLP Agent	< 0.2 sec	Accurate symptom extraction
Computer Vision	< 1.5 sec	Reliable image classification
Prediction Agent	~1.2 sec	Actionable health guidance
Full Pipeline	~4.0 sec	Healthcare-ready output

Table VIII: Agent Pipeline Performance

D. System Performance

Parameter	Observation
End-to-End Pipeline Time	~4.0 seconds average
ChromaDB Query Time	< 400 ms
Image Classification Time	< 1.5 seconds
Voice Transcription Time	< 1.0 second
Dashboard Load Time	< 3 seconds
TTS Response Generation	< 0.8 seconds

Table IX: System Performance

E. User Experience Analysis

The Gradio dashboard provided a smooth and intuitive user experience across all test scenarios. The voice recording interface was accessible and responsive for both technical and non-technical users including elderly participants. The image upload functionality performed reliably across different image formats and sizes. The text-to-speech voice response added significant accessibility value. The dashboard load time of under 3 seconds and sub-4-second end-to-end pipeline execution met performance expectations for a real-time healthcare assistant platform.

VIII. DISCUSSIONS

A. Advantages of the Proposed System

The AI Doctor – Voice & Vision platform offers several key advantages over traditional healthcare applications. It eliminates the need for physical visits for basic symptom checking — once a user records voice symptoms or uploads an image, the entire pipeline executes automatically without further intervention. By grounding all healthcare recommendations in semantically retrieved data from ChromaDB, the system ensures suggestions are based on verified medical knowledge rather than generic AI-generated content, significantly reducing hallucination and increasing patient trust [9].

The multi-agent architecture ensures each cognitive stage is handled by an agent specifically designed for that task, producing higher-quality outputs than a single-agent approach. The Gradio voice interface allows elderly patients without technical expertise to interact with complex AI healthcare analysis through natural voice communication. The system can run on a standard laptop without requiring GPU infrastructure or complex cloud deployment.

B. Limitations of the System

Several limitations were identified during development and testing. The system provides preliminary guidance only and is not intended to replace professional clinical diagnosis. Prediction accuracy depends on the quality and diversity of training datasets. Voice recognition accuracy decreases in high-noise environments despite preprocessing. Image classification performance depends on image quality and lighting conditions [11].

Limitation	Impact
Preliminary guidance only	Cannot replace professional diagnosis
Dataset coverage limits	May miss rare or complex conditions
Voice recognition in noise	Reduced accuracy in noisy environments
No user authentication	No multi-user access control implemented

Table X: System Limitations

C. Real-World Applicability

Rural and remote users can access the platform for basic healthcare guidance without travelling to hospitals. Elderly users can interact naturally through voice, reducing barriers posed by complex digital interfaces. Healthcare workers can use the platform as a preliminary triage tool before formal consultation. The platform also serves as a practical demonstration of multimodal AI integration for academic research and educational settings in AI and healthcare engineering.

IX. FUTURE SCOPE

The AI Doctor – Voice & Vision platform has been designed as a scalable and modular system allowing for continuous enhancement. Several significant improvements are planned for future versions.

- Real-Time Data Integration: Connecting to hospital EHR systems, pharmacy APIs, and wearable health devices to enable live health data analysis and personalised continuous healthcare recommendations.
- Advanced Disease Prediction Models: Integrating time-series health data analysis and training models on larger, more diverse medical datasets to improve prediction accuracy and disease coverage.
- Telemedicine and Video Consultation: Adding real-time video consultation for doctor-patient communication and enabling online appointment scheduling and emergency consultation support.
- Multi-Language Support: Implementing voice recognition and text-to-speech in regional and international languages to expand accessibility for diverse global user populations.
- Enhanced Medical Image Analysis: Extending Computer Vision capabilities to support X-ray, MRI scan, CT scan, and blood report analysis using advanced Deep Learning architectures.
- Mobile Application Development: Developing Android and iOS mobile applications with real-time health alerts, voice interaction, and report download capabilities.

International Journal of
AI ELECTRICAL CIVIL AND MECHANICAL ENGINEERING
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0434 www.zesterapublications.com Original Research Paper

- Cloud Deployment and Scalability: Docker containerisation, Kubernetes orchestration, and scalable vector databases such as Pinecone or Weaviate for large-scale medical knowledge bases.
- AI-Based Mental Health Support: Detecting stress, anxiety, and depression symptoms through voice pattern analysis and conversational interaction, providing mental wellness guidance.

X. CONCLUSION

The AI-Powered Multimodal Healthcare Assistant (AI Doctor – Voice & Vision) presents an effective and scalable solution for automating preliminary healthcare guidance and improving healthcare accessibility in modern digital environments. The system successfully integrates Retrieval-Augmented Generation with a multi-agent AI architecture to deliver data-grounded, multimodal healthcare assistance through an interactive voice and image-capable dashboard interface. The project demonstrates the successful implementation of a full-stack AI pipeline using Speech Recognition for voice processing, spaCy and NLTK for NLP-based symptom analysis, OpenCV and CNN models for medical image classification, ChromaDB for semantic medical knowledge retrieval, and LLM integration for recommendation generation.

The Gradio frontend provides an accessible and functional interface for both technical and non-technical users, including elderly patients, with natural voice interaction, medical image upload, AI-powered disease prediction, and text-to-speech voice response output. Results from system testing validate the platform's ability to deliver healthcare guidance with an end-to-end pipeline execution time of approximately 4 seconds and consistently high RAG retrieval relevance scores. The system acts as a smart preliminary healthcare assistant capable of providing valuable guidance anytime and anywhere, representing an innovative step toward intelligent healthcare solutions and highlighting the growing importance of AI-powered multimodal digital healthcare systems in bridging the healthcare accessibility gap.

REFERENCES

[1] Stuart Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, Pearson Education, 4th Edition, 2020.
[2] World Health Organization (WHO), 'Digital Health and Artificial Intelligence,' Available: WHO Official Website.
[3] National Institutes of Health (NIH), 'Artificial Intelligence in Healthcare,' Available: NIH Official Website.
[4] Aurélien Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media, 2022.
[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, Deep Learning, MIT Press, 2016.
[6] P. Lewis et al., 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,' NeurIPS, 2020.
[7] OpenCV Documentation, 'Open Source Computer Vision Library,' Available: OpenCV Official Website.
[8] TensorFlow Documentation, 'Machine Learning Framework,' Available: TensorFlow Official Website.
[9] Hugging Face Documentation, 'NLP and Transformers,' Available: Hugging Face.
[10] Gradio Documentation, 'Python Library for ML Interfaces,' Available: Gradio Official Website.
[11] Christopher Bishop, Pattern Recognition and Machine Learning, Springer, 2006.