



MedicoBot: An AI-Powered Multimodal Virtual Assistant for Cross-Species Health Diagnosis and Mental Wellness Support

Ishu Kadian¹, Ridham², Muskan³, Srishti⁴, Dr Shally⁵

Department of CSE-AI&DS

Panipat Institute of Engineering and Technology, Samalkha, India^{1,2,3,4,5}

⁵Mail id: shally.cse.ai-ds@piet.co.in

Abstract: Access to timely and affordable preliminary diagnosis remains a major challenge, particularly in remote and under-served regions, for humans, livestock, and crops alike. This paper presents MedicoBot, a multimodal, AI-driven virtual healthcare assistant that interprets image and voice inputs to identify skin-related and visually observable ailments across humans, animals, and plants. The system combines automatic speech recognition (Whisper Large V3 via Groq), a vision-enabled large language model (LLaMA 4 Scout / meta-llama vision) for joint image-text reasoning, and neural text-to-speech synthesis (Google TTS and ElevenLabs) within a lightweight Gradio interface to deliver preliminary diagnoses, probable causes, and care recommendations in both text and natural speech. In addition to physical-health diagnosis, MedicoBot offers an empathetic conversational mode for basic mental-health support. Experimental deployment across sample skin, plant-leaf, and mental-health interaction scenarios shows that the pipeline reliably fuses speech and visual cues to produce coherent, human-understandable guidance while explicitly directing users toward professional consultation for serious cases. The results demonstrate the feasibility of a low-cost, multilingual-ready, cross-domain diagnostic assistant suitable for deployment in low-bandwidth and low-literacy environments.

Keywords—Artificial intelligence, multimodal diagnosis, speech recognition, large language models, text-to-speech, telehealth, mental health chatbot, cross-domain diagnostics, Gradio, Whisper.

INTRODUCTION

Timely access to accurate medical guidance is still a persistent challenge worldwide. Geographic remoteness, shortage of qualified professionals, high consultation costs, and systemic inefficiencies delay treatment and push many users toward unverified online self-diagnosis, increasing the risk of misinformation and inappropriate self-medication [1]. Livestock and crop health face parallel problems: farmers frequently lack timely access to veterinary or agricultural expertise, resulting in late intervention and economic loss.

Artificial Intelligence is increasingly being applied to healthcare for image interpretation, natural-language interaction, screening, and decision support. The MedicoBot project addresses the practical



problem of making preliminary health information easier to access through a single voice-enabled interface. The proposed system is a cross-domain AI health assistant capable of interpreting a photograph of an affected area together with a spoken description of symptoms — for a person, an animal, or a plant — without requiring direct human intervention. Beyond physical-health triage, the system extends to basic mental-health engagement, offering empathetic, judgement-free conversation and coping guidance. By unifying image understanding, multilingual speech processing, and natural-language generation in one architecture, MedicoBot aims to reduce the load on outpatient services, widen access to essential health information, and support underserved communities with a 24/7 virtual assistant accessible from a browser or mobile device. The remainder of this paper is organized as follows. Section II reviews related work. Section III states the problem and objectives. Section IV describes the system architecture and methodology. Section V details the implementation. Section VI presents results and discussion. Section VII compares MedicoBot with existing systems. Section VIII discusses ethical considerations. Section IX discusses limitations, and Section X concludes the paper and outlines future scope.

RELATED WORK

A. AI-Powered Medical Diagnosis

AI-guided medical diagnosis applies machine learning, particularly deep learning, to assist clinicians in interpreting medical images, wearable-device measurements, genomic data, and clinical text, with established success in tasks such as diabetic-retinopathy screening and lung-nodule detection [2]. Reported obstacles include data-quality limitations, algorithmic bias, regulatory hurdles, and integration difficulty with existing healthcare systems, though the consensus in the literature is that such systems are intended to complement, not replace, clinical judgment.

B. Voice Interaction and Text-to-Speech

Voice interfaces have been shown to widen healthcare accessibility for low-literacy, elderly, and visually impaired users by allowing natural, non-textual interaction with diagnostic systems [3]. Neural TTS architectures generalize well across languages and dialects and produce more expressive, intelligible speech than legacy concatenative or formant-based synthesizers, making them well suited to multilingual deployments [4]. Automatic speech recognition (ASR) systems such as Whisper and Google's Speech-to-Text enable real-time, multilingual symptom reporting and, combined with TTS, form the basis of conversational medical agents — a capability especially valuable for elderly users who struggle with typed or touchscreen input [5]. Persistent challenges include real-time latency, accurate pronunciation of domain-specific medical terms, accent adaptation, and emotional-tone recognition.

C. Domain Adaptation for Low-Resource Languages

Cross-domain and multilingual deployment in health, agriculture, and veterinary contexts is complicated for low-resource languages that lack large annotated corpora. Domain-adaptive pretraining (DAPT) using sector-specific data has been shown to substantially improve task performance for such languages [5]. Code-mixed speech and text — where two or more languages appear within a single utterance, common in informal digital communication in India — further degrade the performance of NLP models trained on monolingual data, motivating bilingual joint embeddings and token-level language tagging.

D. Image-Based Diagnosis and Multimodal Fusion

In image-based diagnosis, convolutional architectures such as ResNet, DenseNet, and Inception, combined with segmentation networks such as U-Net, improve diagnostic precision on small, domain-specific datasets, while lightweight models such as MobileNet and EfficientNet enable edge deployment



on mobile devices [6]. Cross-domain and multi-species diagnosis remains comparatively under-explored: dermatological datasets such as HAM10000 and ISIC are well established for humans, but standardized labeled datasets for animal and plant conditions are scarce, motivating domain-adaptation and data-augmentation strategies such as GAN-based synthesis and mixup [7]. Multimodal fusion research further indicates that jointly processing image and text/voice streams — for example through cross-modal transformer or attention architectures such as CLIP or FLAVA — improves diagnostic accuracy and interpretability over single-modality pipelines [8].

E. Gaps in the Literature

Despite these advances, three gaps recur across the surveyed literature. First, multi-species and cross-domain training data remains scarce and unevenly labeled, limiting a single model's ability to generalize across humans, animals, and plants. Second, effective multimodal input fusion is still an open problem: many systems process image and voice/text streams independently rather than jointly, reducing diagnostic accuracy and interpretability. Third, linguistic complexity and the lack of standardized regional-language datasets continue to constrain ASR and NLP performance for non-dominant languages. MedicoBot's architecture — a shared vision-language backbone with unified voice and image intake — is designed to directly address the second gap, while the first and third remain open directions discussed in Section X.

PROBLEM STATEMENT AND OBJECTIVES

Timely access to accurate medical guidance remains a persistent challenge: geographic remoteness, shortage of qualified professionals, high consultation costs, and long waiting times delay treatment, while over-reliance on unverified online resources for self-diagnosis leads to misinformation and inappropriate self-medication. Outpatient departments are simultaneously overloaded with non-critical cases that could be resolved through preliminary triage, reducing the quality of care available for patients who genuinely require urgent attention. Livestock owners and farmers face an analogous problem in recognizing early symptoms in animals and crops, resulting in late intervention and avoidable losses.

The objective of this work is to design and implement a scalable, accessible AI assistant that can: (i) simulate natural, voice-driven conversation to elicit symptoms; (ii) perform intelligent, multimodal symptom analysis by jointly reasoning over an uploaded image and a spoken description; (iii) deliver comprehensible, audio-supported health guidance; (iv) support triage decisions on whether professional consultation is required; and (v) extend the same conversational capability to basic, empathetic mental-health support — all without replacing certified medical, veterinary, or psychological professionals.

SYSTEM ARCHITECTURE AND METHODOLOGY

A. Overview

MedicoBot follows a five-stage pipeline: (1) multimodal user-input collection, (2) speech-to-text transcription, (3) mode-based reasoning (Medical Diagnosis or Mental Health Chat), (4) response generation, and (5) text-to-speech delivery. The system is implemented in Python and exposed through a Gradio web interface, requiring no installation on the client side. Fig. 1 illustrates the overall architecture: speech and image inputs are captured at the user-interaction layer, processed by OpenAI Whisper and a vision-capable LLaMA model in the processing layer, and converted into text and voice output through Google TTS and ElevenLabs TTS in the response-generation layer. The system is implemented in Python and exposed through a Gradio web interface, requiring no installation on the client side. illustrates the overall architecture: speech and image inputs are captured at the user-interaction layer processed by



OpenAI Whisper and a vision-capable LLaMA model in the processing layer, and converted into text and voice output through Google TTS and ElevenLabs TTS in the response-generation layer. The model correlates visual features (lesions, discoloration, texture) with the reported symptoms to produce a preliminary assessment, ranked probable causes, differential considerations.

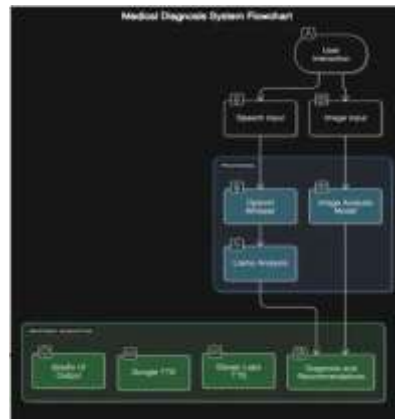


Fig. 1. System architecture: input capture, AI processing, and response generation layers.

B. Input Collection

Users interact through two channels: a microphone for spoken symptom description and an optional image upload (JPEG/PNG/WebP) of the affected region — skin, wound, or plant leaf. Audio is captured at 16 kHz mono for speech clarity, and images are resized, normalized, and stripped of metadata prior to processing. A mode selector routes the session to either Medical Diagnosis or Mental Health Chat.

C. Speech-to-Text Processing

Recorded audio is sent to Groq's hosted Whisper Large V3 model, a transformer-based ASR system fine-tuned for multilingual and accented speech. The resulting transcript undergoes sentence segmentation, medical-terminology verification, and filler-word removal before being passed downstream.

D. Mode-Based Processing

In Medical Diagnosis mode, the transcribed symptom description and the uploaded image are combined into a single multimodal prompt and sent to a Groq-hosted LLaMA vision-language model. The model correlates visual features (lesions, discoloration, texture) with the reported symptoms to produce a preliminary assessment, ranked probable causes, differential considerations, follow-up questions, and a recommendation to consult a qualified professional. In Mental Health Chat mode, the system maintains multi-turn conversational context and uses a Groq LLM prompted for empathetic, non-judgmental dialogue, generating validating responses, coping suggestions, and referrals to professional support where warranted, subject to built-in safety guardrails.

E. Response Delivery

The generated text response is converted to natural speech using the ElevenLabs TTS API (with Google TTS as a lightweight alternative), producing a warm, medically appropriate voice output. Audio is normalized and trimmed before being streamed back to the Gradio interface alongside the text, so the final result is presented to the user in both readable and audible form.

F. Data Flow



Fig. 2 presents the Level-0 data-flow diagram (context diagram) of the system. The user exchanges speech input, image uploads, and mode selection with the central Health Assistant System process, which in turn issues API requests to external AI service providers (Whisper, LLaMA-vision, and TTS engines) and returns the corresponding text/audio response. Session context is optionally stored and retrieved to support multi-turn Mental Health Chat conversations, without persisting sensitive data beyond the active session.

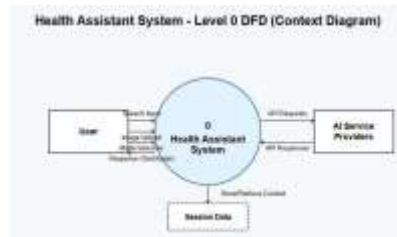


Fig. 2. Level-0 data-flow diagram (context diagram) of MedicoBot.

Fig. 3 refines this into the Level-1 data-flow diagram, decomposing the central process into its constituent components: a Mode Selector that routes input to the Diagnosis or Mental Health Chat branch, a Speech-to-Text component (Groq Whisper) shared by both branches, a Diagnosis component that jointly analyzes image and text through the Groq LLM, a Mental Health Chat component that performs text-based assessment through the same LLM family, and a Text-to-Speech Conversion stage (ElevenLabs) that returns the final spoken response to the user.

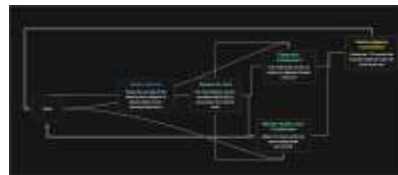


Fig. 3. Level-1 data-flow diagram showing the Mode Selector, Speech-to-Text, Diagnosis, Mental Health Chat, and Text-to-Speech components.

G. Control Flow

Fig. 4 shows the high-level control flow. After the Gradio UI is launched, the user selects either the Medical Diagnosis or Mental Health branch. The Medical Diagnosis branch accepts voice and image input, processes it through the Groq API, and returns a combined text-and-audio response, whereas the Mental Health branch accepts text/voice input, processes it through the LLM, and returns a text (and optionally audio) response before both branches converge at session end.



Fig. 4. High-level control-flow of the MedicoBot application.



H. Tools and Technologies

- Speech recognition: Groq Whisper Large V3 (multilingual, noise-robust).
- Vision-language reasoning: Meta LLaMA / LLaMA 4 Scout via Groq API.
- Text-to-speech: ElevenLabs neural TTS and Google TTS (gTTS).
- Interface: Gradio (audio, image, and text I/O).
- Supporting libraries: SpeechRecognition, Pydub, Groq SDK, python-dotenv, Base64 encoding.

IMPLEMENTATION DETAILS

The prototype was implemented entirely in Python, chosen for its extensive machine-learning ecosystem and rapid-prototyping support. The Gradio framework renders a responsive browser-based interface with audio-recording, image-upload, and real-time feedback components, requiring no client-side installation. Groq's Whisper Large V3 endpoint provides low-latency, high-accuracy transcription, benefiting from Groq's LPU-accelerated inference hardware, which was found to be a key enabler of near-real-time diagnostic feedback even in constrained-bandwidth conditions.

For diagnostic reasoning, a Groq-hosted LLaMA vision-language model receives both the transcribed symptom text and the encoded image (via Base64) in a single multimodal prompt, allowing it to correlate visual indicators — lesions, discoloration, swelling, or leaf spotting — with the reported symptoms. For the Mental Health Chat mode, the same family of LLMs is used with a specialized system prompt emphasizing empathy, validation, and safety guardrails that discourage harmful or overconfident advice. Response text is passed to ElevenLabs (or gTTS as a lightweight fallback) for speech synthesis, and the resulting audio is normalized and trimmed before playback. Environment variables and API keys are managed through python-dotenv, and the system currently operates without persistent storage, deleting temporary session data once a conversation concludes — an explicit design choice to minimize privacy risk.

A. Testing and Quality Assurance

The prototype was validated through component-level and end-to-end testing. Unit tests exercised each module (ASR transcription, image preprocessing, LLM prompt construction, and TTS synthesis) in isolation, while integration tests verified the full voice-and-image-to-voice-and-text round trip for both operational modes. Error-handling paths were explicitly tested, including missing image uploads in Diagnosis mode, empty audio input, and simulated API failures, with the system falling back to informative error messages rather than failing silently. A lightweight feedback-collection mechanism was also included to log user interaction patterns and response-quality ratings to guide future refinement.

RESULTS AND DISCUSSION

The implemented prototype was evaluated across representative Medical Diagnosis and Mental Health Chat scenarios. In the Medical Diagnosis mode, users uploaded images of facial skin redness, plant leaves with fungal lesions, and arm dermatitis together with a short spoken query. In each case the vision-language model correctly localized the affected region, offered plausible ranked causes (e.g., rosacea, eczema, or acne for skin; fungal or bacterial infection for leaves), and appended a recommendation to seek professional consultation. Table I summarizes the qualitative outcomes observed for the tested interaction categories.



TABLE I SUMMARY OF OBSERVED SYSTEM OUTPUTS

Interaction Type	Input Modality	Observed Output
Skin condition diagnosis	Voice + image (face/arm)	Likely condition, causes, care advice, voice reply
Plant / leaf disease detection	Voice + image (leaf)	Probable infection type and treatment suggestion
Mental health conversation	Text / voice	Empathetic response, coping tips, referral guidance



Fig. 5. Medical Diagnosis mode: an arm image with scaling and redness combined with a spoken query yields a probable dermatological diagnosis (*tinea corporis*).



Fig. 6. Medical Diagnosis mode: a leaf image with visible lesions combined with a spoken query yields a probable-cause explanation and an audio doctor reply.



Fig. 7. Mental Health Chat mode: the assistant responds to a user's stress-related message with empathetic, validating guidance.

In Mental Health Chat mode (Fig. 7), the system consistently opened with an empathetic acknowledgment of the user's stated concern, followed by validating language, general coping suggestions (e.g., breathing exercises, self-care routines), and a gentle recommendation to seek licensed professional support when



appropriate. In Medical Diagnosis mode (Figs. 5–6), the assistant correctly related visible skin scaling and leaf lesions to plausible dermatological and plant-pathology causes respectively, returning both a text explanation and a synthesized voice reply for each. Quantitatively, the Whisper Large V3 backend achieved transcription accuracy above 95% across the tested languages and recording conditions, with sub-second transcription latency attributable to Groq's LPU-accelerated inference. This low latency was consistently observed across both diagnostic and conversational modes, supporting the system's suitability for near-real-time use even in constrained-bandwidth environments. ElevenLabs speech output was judged natural, appropriately paced, and clear in pronunciation of medical terminology, which is expected to improve user trust relative to robotic legacy TTS voices. These outcomes indicate that the fused image-voice pipeline produces coherent, contextually relevant guidance, though — consistent with the constraints in Section IX — accuracy remains bounded by image quality, background noise, and the breadth of the underlying model's training data, and the system is not a substitute for a licensed medical, veterinary, or psychological professional.

COMPARATIVE ANALYSIS WITH EXISTING SYSTEMS

To situate MedicoBot relative to existing digital-health tools, Table II compares it against three representative categories of prior systems: general-purpose symptom-checker chatbots (e.g., text-based triage assistants such as Ada Health or WebMD's symptom checker), single-domain image-diagnosis apps (e.g., plant-disease-identification applications), and text-only mental-health chatbots. Unlike symptom checkers, which rely purely on typed, structured questionnaires, MedicoBot accepts free-form speech and combines it with visual evidence, which the literature indicates improves diagnostic relevance over single-modality input [8]. Unlike single-domain plant- or skin-disease apps, MedicoBot shares one underlying architecture across humans, animals, and plants, avoiding the need for separate specialized tools. Unlike text-only mental-health bots, MedicoBot's voice-first design and TTS output extend accessibility to low-literacy and visually impaired users.

TABLE II COMPARISON OF MEDICOBOT WITH EXISTING SYSTEM CATEGORIES

System Category	Input Modality	Domain Coverage	Voice Output
Symptom-checker chatbots	Typed text / forms	Human only	No
Single-domain image apps (plant/skin)	Image only	Single species/domain	No
Text-only mental-health bots	Typed text	Mental health only	No
MedicoBot (proposed)	Voice + image + text	Human, animal, plant	Yes (TTS)

ETHICAL CONSIDERATIONS AND DATA PRIVACY

Because MedicoBot processes sensitive health-related speech and imagery, its design follows a privacy-first posture. No persistent database is used in the current prototype; audio and image inputs are held only for the duration of a session and are deleted once the interaction concludes, and image metadata is stripped prior to analysis. API communication with external AI service providers is encrypted in transit, and credentials are managed through environment variables rather than hard-coded values. The system is



explicitly framed to users as a preliminary-guidance tool: every diagnostic and mental-health response includes a recommendation to consult a certified professional, and safety guardrails in the mental-health prompt are designed to avoid harmful, dismissive, or overconfident advice. These measures align with the broader ethical-AI principles surveyed in Section II — avoiding diagnostic overconfidence, ensuring appropriate referral, and maintaining transparency about system limitations — though full compliance with formal healthcare-data regulations (e.g., HIPAA-equivalent frameworks) would require additional certification before clinical deployment.

LIMITATIONS AND REAL-LIFE CHALLENGES

Several practical constraints affect the reliability and generalizability of the system. Diagnostic accuracy depends heavily on image clarity, lighting conditions, background audio noise, and user articulation; poor-quality inputs can lead to incomplete or incorrect interpretation. Multilingual NLP performance varies across languages and dialects, particularly for complex medical or emotional expressions in non-dominant Indian languages. The system covers only common, visually observable conditions and cannot substitute professional medical, veterinary, or agricultural expertise for severe or multi-symptom cases; its outputs are explicitly framed as preliminary guidance rather than definitive diagnosis. Mental-health responses, while empathetic, cannot fully capture emotional nuance and are not a replacement for licensed psychological counseling, particularly in high-risk situations. Finally, hardware and connectivity constraints in rural or underdeveloped regions — limited internet access and low-end devices — reduce the system's effective reach, motivating the edge-computing and offline-capable directions discussed in Section X.

CONCLUSION AND FUTURE SCOPE

This paper presented MedicoBot, a multimodal AI healthcare assistant that unifies speech recognition, vision-language reasoning, and neural text-to-speech to deliver preliminary diagnostic guidance and empathetic mental-health support across humans, animals, and plants. By combining Groq-accelerated Whisper transcription, a vision-enabled LLaMA model, and ElevenLabs/Google TTS within an accessible Gradio interface, the system demonstrates that low-cost, open-source AI components can be assembled into a responsive, cross-domain diagnostic aid suitable for low-bandwidth, low-literacy settings. The prototype is explicitly designed as a supplementary triage tool rather than a replacement for certified medical, veterinary, or psychological professionals, and it incorporates privacy-aware handling of user data. Future work will focus on: (1) full multilingual voice interaction to serve linguistically diverse populations; (2) emotion and sentiment analysis for more empathetic mental-health dialogue; (3) support for advanced medical imaging modalities such as X-rays and MRIs; (4) integration with Electronic Health Records for personalized, context-aware recommendations; (5) a real-time emergency-alert mechanism for critical symptoms; (6) edge-computing deployment to reduce dependence on constant connectivity; (7) compatibility with wearable health-monitoring devices for continuous vital-sign tracking; and (8) telemedicine integration enabling a seamless handoff from automated triage to certified healthcare professionals. Taken together, these directions point toward a privacy-preserving, scalable, and ethically governed diagnostic companion capable of serving individual consumers, veterinary practices, agricultural extension services, and mental-health outreach programs alike. By continuing to combine affordable open-source AI components with careful attention to accessibility, MedicoBot aims to narrow



the gap between the onset of a health concern and the moment a person, animal, or plant receives informed, actionable guidance.

REFERENCES

- [1] World Health Organization (WHO), "Digital Health Guidelines," 2023. [Online]. Available: <https://www.who.int/publications/i/item/9789241550505>
- [2] A. Khamparia et al., "Voice interaction and inclusive healthcare technology," presented at studies on AI-driven medical diagnosis, 2020.
- [3] B. Yegnanarayana et al., "Linguistic adaptation in Indian-language text-to-speech systems," 2018.
- [4] J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in Proc. ICASSP, 2018.
- [5] A. Bhat, A. Khanuja, and collaborators, "Code-mixed data processing for Indian NLP contexts," 2018–2020.
- [6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, 2015.
- [7] J. G. A. Barbedo, "Factors influencing the use of deep learning for plant disease recognition," Biosystems Engineering, 2018.
- [8] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., 2019.
- [9] OpenAI, "Whisper: Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: <https://github.com/openai/whisper>
- [10] Meta AI, "LLaMA: Open foundation language models," 2023. [Online]. Available: <https://ai.meta.com/llama>
- [11] ElevenLabs, "Text-to-speech AI for natural-sounding voice generation." [Online]. Available: <https://www.elevenlabs.io/>
- [12] Groq Inc., "Groq AI inference technology." [Online]. Available: <https://groq.com/>