Multimodal AI Virtual Doctor with Voice-Based Diagnostics

Vishwas Aithal K V[1], Khavya U S Reddy[2], Prof. Suresh Babu P[3]
[1,2,3] *Department of Artificial Intelligence and Data Science, BMS College of Engineering, Bengaluru, India*

Abstract*—The rapid integration of AI into healthcare is bringing about scalable, accessible, and intelligent diagnostic support systems. This paper presents an AI-driven Medical Voice Agent for automated symptom-based doctor allocation, real-time conversational consultation, and GPT-powered structured medical report generation. It integrates a multi-modal architecture with text-based symptom analysis, voice interaction via the Vapi API, speech transcription using AssemblyAI, and document understanding via a hybrid OCR pipeline. Computer-generated medical reports are processed using EasyOCR, while re-analyses of handwritten prescriptions or low-confidence scans leverage a vision-transformer model, TrOCR, for enhanced accuracy. Furthermore, the system includes an X-ray fracture detection module powered by a CNN and Grad-CAM++ for explainable image-based diagnostics. This ensures that the final clinical summary is generated solely from the recorded doctor–patient conversation to implement human-like medical reporting. This system also offers location-based recommendations for nearby hospitals and pharmacies to enhance post-consultation usability. This solution can thus be used to provide fast, preliminary, and accessible medical guidance, reduce waiting times, and increase accessibility, especially in resource-constrained areas. Experimental results revealed accurate doctor allocation, reliable OCR-based document extraction, robust fracture detection, and coherent GPT summaries, all which point to the potential effectiveness of this system as a supportive tool in healthcare.

Index Terms*—AI Medical Voice Agent, GPT Report Generation, Speech-to-Text, Vapi API, AssemblyAI, EasyOCR, TrOCR, Medical Document OCR, Symptom-Based Doctor Allocation, CNN, Grad-CAM++, X-ray Fracture Detection, Healthcare Automation, Telemedicine, Google Maps API, Hospital and Pharmacy Recommendation, NeonDB.

I. INTRODUCTION

Despite all the significant advancements in global healthcare infrastructure, accessibility, efficiency, and diagnostic accuracy continue to pose persistent challenges in both urban and rural medical ecosystems [2]. Conventional telemedicine systems-while gaining widespread adoption-continue to rely on manual doctor allocation, inflexible scheduling workflows, and resource-intensive video consultations, thus rendering them inadequate in high-demand environments. Delays in response time, restricted specialist availability, inconsistent clinical documentation, and increased patient burden are frequent outcomes of such inefficiencies, particularly for patients with mobility or low digital literacy challenges. With the exponential increase in demand for healthcare, scaling the traditional teleconsultation pipelines is difficult; this generates fragmented patient experiences and deepens disparities in timely medical care. Adding to these issues, the dependency on static text forms and template-based questionnaires that cannot capture the nuance, spontaneity, and contextual richness of real doctor–patient conversations should be taken into consideration. The lack of automated reasoning, multimodal diagnostic support, and contextual comprehension drastically limits its capacity for effective triage or precise summarization, leading to information gaps and incomplete medical records across millions of consultations. Addressing these challenges, the project "AI Medical Voice Agent for Automated Doctor Allocation, GPT-Driven Clinical Report Generation, X-ray Fracture Detection Using CNN and Grad-CAM++, and Medical Document Understanding with EasyOCR and TrOCR" proposes a fully automated AI-centric multimodal tele-health ecosystem. Advanced speech technologies-Vapi for real-time conversational interaction and AssemblyAI for high-fidelity speech-to-text translation-form the backbone of the system, which enables seamless, uninterrupted doctor–patient communication. Transcribed dialogues will be then interpreted by a GPT-based clinical reasoning engine capable of generating structured medical reports as would be done by standard physician documentation. Beyond voice-based assessment, a dual-stage OCR pipeline is incorporated into the platform: EasyOCR for extracting text from computer-generated medical reports and TrOCR [11], a Vision Transformer–based model, to enable robust recognition of handwritten prescriptions. This unified OCR framework ensures accurate digitization of diverse medical records, further enriching the clinical context available for downstream reasoning. In addition, the diagnostic support through a CNN + Grad-CAM++ module for X-ray fracture detection provides interpretable visual evidence without influencing the GPT-generated summary [3], [2], [12]. For improved post-consultation care, a location-aware recommendation engine powered by the Google Maps API recommends nearby hospitals, clinics, and pharmacies based on the patient's geographic location. The platform is deployed on a secure and scalable technology stack, comprising Next.js, React, Clerk authentication, and NeonDB, ensuring data stored is trustworthy, the interaction is seamless, and previous consultations are quickly accessed. In tune with key Sustainable Development Goals-SDG 3 of Good Health, SDG 9 of Innovation, and SDG 10 of Reduced Inequalities-the proposed system offers a scalable, intelligent, and just approach toward digital health, providing quick triaging, increased access, and more value-informed medical decision-making for varied populations.

II. RELATED WORK

A. Voice-Based Telemedicine and Speech Processing

Early telemedicine systems employed rule-based chatbots and menu-driven conversational agents with limited natural dialogue, offering scant support for diagnosis. Recent speech technologies, in particular cloud-based ASR platforms such as Google Speech, Whisper, and AssemblyAI, demonstrate major improvements in medical speech recognition [8]. Research has established that voice-based medical assessment is more accessible to elderly users, people with visual impairments, and patients with low digital literacy. Moreover, studies that have incorporated APIs like Vapi for real-time interaction have found evidence that continuous, back-and-forth conversations increase patient engagement significantly and result in much lower miscommunication compared to systems with text-based conversation [9]. These contributions set the base for developing AI voice agents capable of emulating real doctor–patient interactions in medicine.

B. Deep Learning for Medical Imaging and Fracture Detection

Over the last few years, CNNs have evolved as the leading paradigm for medical image interpretation, especially in radiology. Previous research on fracture detection has shown that deep CNN architectures-ResNet, DenseNet, and MobileNet-often outperform conventional radiographic interpretation in both sensitivity and specificity [3]. Also, the recent development of explainable AI techniques, especially Grad-CAM and its enhanced form, Grad-CAM++, has been crucial in instilling confidence among clinicians by visually identifying regions responsible for model predictions. Research within orthopedic imaging and trauma assessment emphasizes that heatmap visualization bears great significance to clinical adoption; hence, CNN + Grad-CAM++ serves as an apt choice for real-time fracture identification in AI-assisted diagnostic platforms.

C. LLM for Medical Summarization and Clinical Reasoning

Large language models like GPT-4, MedGemma, and BioGPT have shown excellent performance in generating human-like clinical summaries, answering medical questions, and even performing preliminary diagnostic reasoning. Studies clearly indicate that LLMs outperform earlier NLP pipelines by leveraging contextual understanding from long-form medical dialogues and multimodal instructions. However, most of the existing healthcare LLM applications are focused on either question-answering or summarization from structured EHR notes, rather than real-time conversational transcripts [6]. Very few prior works integrate the generation of LLM-based reports right after a voice consultation. Further, few current systems offer a unified workflow that combines triage, doctor allocation, voice interaction, and automated documentation, which justifies the gap the proposed system aims to bridge.

D. Automated Triage and Intelligent Doctor Allocation Systems

Automated triage systems have evolved from symptom-checker algorithms to advanced ML-driven classifiers that are able to predict categories of disease, urgency, and appropriate medical specialties. Research in clinical decision support clearly shows that ML-based triage reduces the patient load in emergency departments while improving the accuracy of referrals [10]. Some incorporate Bayesian networks or supervised learning models for symptom-classification-based specialty classification. However, most of these work as separate modules and remain disconnected from conversational agents or imaging systems. Different from these fragmented solutions, the proposed platform will integrate triagedriven doctor selection directly into a live-voice consultation workflow, thus ensuring an uninterrupted digital clinical experience from symptom entry to final report creation.

E. Location-Aware Healthcare Recommendation Systems

Geospatial health informatics has been garnering ever-increasing attention in the last few years, wherein studies have used Google Maps APIs and GIS tools to spot medical resources nearby for quick emergency response. Most of the previous works basically revolve around ambulance routing, hospital capacity detection, and urban healthcare distribution analysis [7]. Although these are very useful, very few systems incorporate location-based recommendations as part of an AI-driven teleconsultation pipeline. The proposed system incorporates geolocation at the last stage of consultation to add a pragmatic layer of assistance in order to direct the patients towards nearby hospitals, diagnostic centers, and pharmacies; such steps are very crucial for real-world clinical applicability at both urban and rural levels.

F. Multimodal AI Systems for Integrated Healthcare

Multimodal AI, integrating speech, structured text, and sensor data, has emerged as a very strong paradigm in healthcare. Prior art has focused on the use of multimodal fusion for disease detection, patient monitoring, and clinical decision support. However, most systems so far only implement partial multimodality-for example, text + imaging or imaging + sensor data [8]. An end-to-end pipeline combining voice consultation, summarization with LLM, triage logic, and medical imaging remains largely unexplored. The proposed system differentiates from the current state of the art by integrating these components into one seamless workflow, using conversational AI, CNN-based fracture detection, and LLM-driven medical reasoning on a single operational platform.

G. OCR for Medical Document Digitization and Handwritten Prescription Analysis

OCR plays a crucial role in the conversion of medical documents into structured digital text, including prescriptions, laboratory reports, and discharge summaries. Traditional OCR systems often fall short when dealing with noisy scans, irregular layouts, and handwritten clinical notes; therefore, deep learning–based methods have been widely adopted for improved accuracy. EasyOCR, based on the CNN–LSTM architecture with CTC decoding, performs reliably on printed or computer-generated medical reports but shows reduced accuracy for handwritten content [11], [12]. This calls for transformer-based OCR models such as Microsoft's TrOCR, which follows a Vision Transformer encoder and Transformer decoder to allow for complex handwriting with varied stroke patterns. Unlike previous systems that handled printed and handwritten documents separately, the proposed platform will integrate both through a dual-stage OCR pipeline, namely EasyOCR for high-confidence printed text extraction and TrOCR as a fallback in cases of handwritten or low-confidence inputs. This hybrid approach ensures more robust medical document digitization and strengthens the overall multimodal intelligence of the system.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

Overall Architecture and System Flow:

The system architecture is a unified multimodal AI healthcare pipeline, integrating voice consultation, medical document OCR, X-ray analysis, and LLM-powered clinical reporting. A user starts from the AI Consultation Website by authenticating through Clerk API and accessing the Medical Console Dashboard to initiate a voice consultation or upload medical reports and X-rays. The real-time voice interaction is handled by VAPI, and speech is converted into text using AssemblyAI for further downstream processing. Uploaded medical documents are interpreted using a dual OCR framework: EasyOCR for printed reports and TrOCR for handwritten prescriptions. This ensures high-quality text extraction across diverse formats. X-ray images are analyzed by a CNN enhanced with Grad-CAM++ for fracture detection and providing visual explainability. All the data extracted—speech transcript, OCR outputs, and imaging results—are accumulated in the Model Integration layer and passed to the LLM-driven Medical Report

Generator, generating a structured, doctor-style clinical summary. Additionally, the Google Maps API provides practical guidance on nearby hospitals and pharmacies. All the consultation records, diagnostic findings, and generated reports are securely stored in NeonDB through Drizzle ORM, completing the integrated end-to-end automated digital healthcare workflow.

Fig 1: Overview of Overall Architecture

1. User Authentication and Dashboard Interaction Layer

This layer controls the authentication of users and offers access to patients to the AI medical console. The Clerk API is responsible for authentication, hence session tokens are encrypted, and login is secure. After authentication, it will load personalized health data on the dashboard and provide entry points such as voice consultation, symptom submission, X-ray upload, and medical report upload [6]. On submitting symptoms, these are normalized first into structured form and then handed over to the triage subsystem along with the request for specialization mapping.

2. Symptom-Based Doctor Allocation Engine

The system features an automated triage classifier to map the symptoms to appropriate specialist categories. It uses a rule-based scoring vector:

$$D^{*} = argmax_{d \epsilon D} S(d|x) \qquad (1)$$

where X is the symptom vector and S(d|x) is the rule-based similarity score [10]. The engine finds whether the patient needs an Orthopaedic specialist, General Physician, Dermatologist, or Cardiologist depending on the keyword clusters. The output specialisation is used in customizing the conversational prompt of the AI Doctor before the actual voice consultation begins.

3. Real-Time Voice Consultation Pipeline

The voice-based medical interview is conducted through Vapi, a real-time streaming conversational agent, while AssemblyAI performs low latency Automatic Speech Recognition (ASR) [8]. Speech signals are transformed into text by using end-to-end ASR modelling:

$$\hat{Y} = argmax_{Y}\, P(Y|X) \qquad (2)$$

where X is the audio sequence, and Y is the decoded transcript. AssemblyAI's medical speech models make sense of accents, variations in pace, and environmental noise to let the transcript capture symptom descriptions, details about the injury, lifestyle facts, and medication intake accurately [9]. The final transcript then acts as an input for LLM-based medical reasoning.

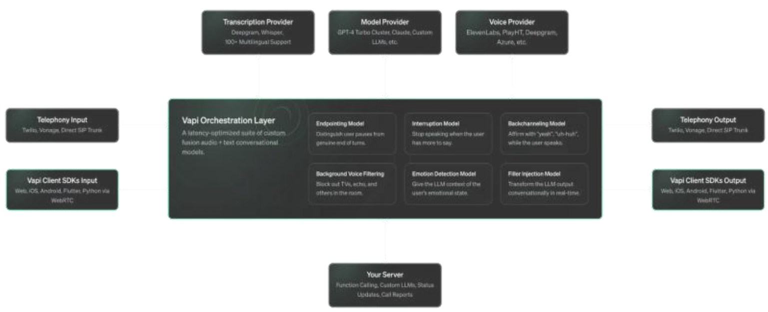

Fig 2: Vapi AI orchestration layer.

Vapi acts as a real-time orchestration layer that connects telephony and web inputs with the top-tier AI providers for transcription, language modeling, and voice synthesis. Core value comes in because it deals with the complicated nuances around human conversation: distinguishing pauses from turn-taking, dealing with interruptions, and injecting natural fillers to ensure lowlatency fluid interactions. By acting as the central hub that routes audio data between these external AI services and your own server for business logic, Vapi enables the deployment of responsive emotionally intelligent voice agents without requiring developers to build the underlying audio processing infrastructure.

4. OCR-Based Medical Report Understanding (EasyOCR + TrOCR)

The system uses a dual-stage OCR pipeline for the extraction of details from uploaded medical documents like prescriptions, lab reports, and discharge summaries.

A. EasyOCR Processing for Printed Reports

EasyOCR uses the CNN-LSTM-CTC architecture. CNN extracts the spatial features f(x), the bi-directional LSTM models the sequential dependencies, and CTC decodes the best character sequence:

$$\hat{Y} = CTCDecode(P(Y|f(x))) \tag{3}$$

It is only really performing well when these are clean, computer-generated medical documents with uniform fonts and formatting [11]. In your system, EasyOCR is the first-pass OCR model for all uploaded PDFs/images.

B. TrOCR Processing for Handwritten Reports

If EasyOCR confidence falls below a threshold, the system switches to Microsoft TrOCR (Handwritten).TrOCR uses a Vision Transformer encoder (splitting images into patches) and a Transformer decoder that autoregressively generates tokens:

$$h = ViTEncoder(x) \tag{4}$$

$$y_t = TransformerDecoder(h, y < t) \tag{5}$$

This architecture is highly effective for irregular handwriting, varying stroke patterns, and noisy prescriptions.

In this project, TrOCR is responsible for accurately transcribing handwritten prescriptions in a task that cannot be reliably done by conventional OCR [12]. Finally, GPT creates a complete medical summary based on transcript and OCR outputs. The LLM processes structured and unstructured inputs through attention mechanisms:

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{d_K})V \tag{6}$$

The generated report includes Chief Complaint, History of Present Illness, Differential Diagnosis, Doctor Observations, Assessment, Treatment Advice, and Follow-up Instructions. Where applicable, medical reports extracted via OCR and X-ray findings are incorporated. The final summary mimics the writing of a real doctor and is stored in NeonDB.

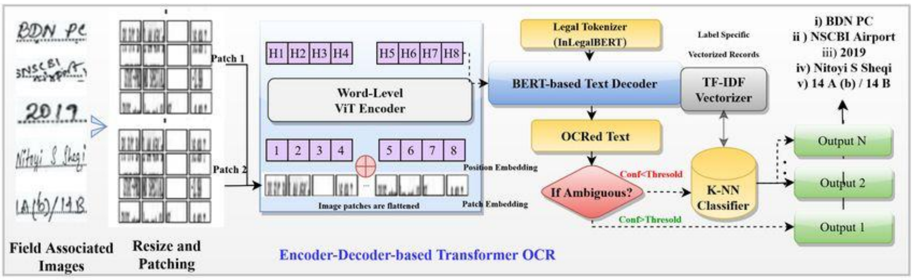

Fig 3: TrOCR Architecture

This hybrid OCR system segments handwritten images into patches that are processed by a Vision Transformer (ViT) encoder, together with a BERT-based decoder. It uniquely introduces a safety mechanism in which the low-confidence ("ambiguous") predictions are cross-referred automatically against a database using a K-NN classifier and TF-IDF vectorizer. This enables the model to exploit deep learning for general recognition, using a retrieval-based fall-back to correct errors on difficult or messy inputs.

5.. Diagnostic Modelling Layer (CNN + Grad-CAM++)

This module analyzes uploaded X-rays to detect fractures.

A. CNN-Based Fracture Detection

The CNN extracts hierarchical spatial features using convolution:

$$F_{i,j,k} = \sigma(\Sigma_{m,n} X_{i+m,j+n} . W_{m,n,k} + b_k) \quad (7)$$

where σ is ReLU. The final dense layer outputs a probability:

$$P(fracture|X) \quad (8)$$

B. Explainability via Grad-CAM++

Grad-CAM++ computes pixel importance using:

$$L^c_{CAM} = \Sigma_k \alpha^c_k A^k \quad (9)$$

where α^c_k are the weighted gradients and A^k are feature maps [1].This heatmap highlights suspected fracture regions and is overlaid on the original X-ray for visual clarity.

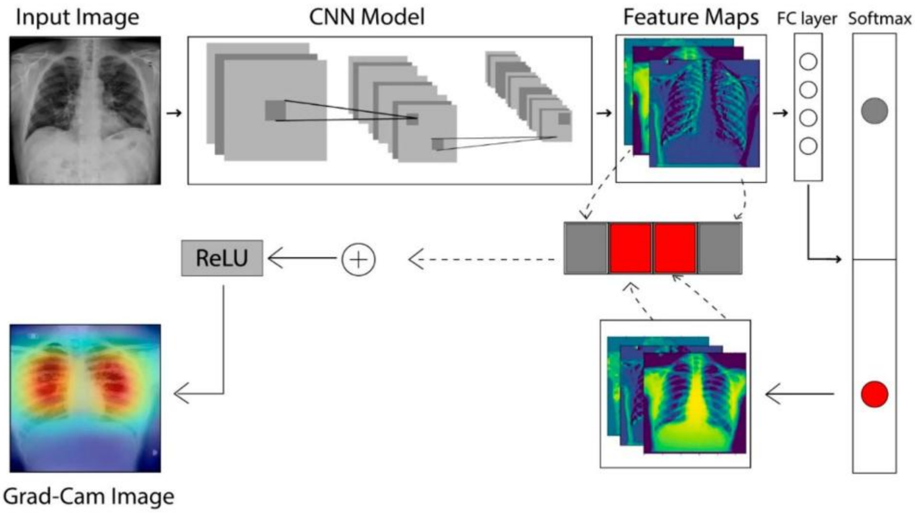

Fig 4: CNN + Grad-CAM++ Architecture

This is a diagram explaining one such technique known as Grad-CAM, used in showing where exactly a model looked to make its prediction by explaining deep learning decisions. After the CNN predicts one particular class (a disease in this case), working backwards to highlight and weigh the feature maps most active for that result: The relevant features will be combined and filtered through the ReLU function, which simply removes the negative values to generate the "heatmap". This heatmap now overlaid onto the original X-ray shows specific regions the AI focused on in reaching its diagnosis [2].

6. Location-Based Healthcare Recommendation Layer

It makes use of the Google Maps API in order to provide proximity-based medical assistance through the nearest hospitals, clinics, and pharmacies, taking into consideration the user's current geolocation coordinates (lat, lon). The distances between the patient and all the nearby facilities are calculated using the Haversine formula, which accurately computes real-world distances. The system calculates latitude and longitude differences to identify the care centers most relevant within a practical radius [3]. This list of recommendations is generated dynamically at the end of each consultation to facilitate timely medical follow-up. This layer adds great value to the platform by connecting digital diagnostics with real-world healthcare access.

7. Data Storage and Management Layer

This involves storing patient data, including consultation transcripts, medical summaries generated by LLM, X-ray fracture detection results, text extracted by OCR, and historical logs of visits in the Data Storage Layer using NeonDB. Drizzle ORM provides typesafe SQL operation support, which makes it reliable and defends against schema mismatches for database operations. All these records are stored in encrypted format to satisfy the demand of modern data protection and privacy provisions. Storage architecture will be optimized for fast retrieval to seamlessly access medical history from the dashboard. Thus, the structured storage pipeline will allow scalability in the long run with consistent performance across the healthcare ecosystem.

Development Environment:

Processor: Intel Core i5 (8th Gen) / AMD Ryzen 5

(sufficient for multimodal AI processing).
RAM: 16–32 GB (required for handling LLM interactions, OCR pipelines, and CNN inference).
GPU: NVIDIA RTX 3050 / 3060 or higher (for training and running CNN + Grad-CAM++ and accelerating TrOCR inference).
Storage: SSD with 512 GB–1 TB capacity for medical datasets, X-ray archives, and report images.
Deployment Setup:
Platform: AWS / GCP / Azure or Vercel + GPU-backed instances for scalable voice, OCR, and imaging workloads.
Compute Requirements: GPU-enabled VM (NVIDIA T4 or A100) for running CNN + Grad-CAM++ efficiently.
Data Storage: NeonDB (PostgreSQL) for structured medical data; object storage (AWS S3/GCP Storage) for X-ray and report images.
APIs & Services:
Vapi for real-time voice agent interaction
AssemblyAI for speech-to-text
Google Maps API for hospital/pharmacy recommendations
OCR Stack: EasyOCR + TrOCR.
Software Stack:
Frontend: Next.js, React, TailwindCSS
Backend: Node.js, Express, REST APIs
AI/ML Components:
Python 3.10+
PyTorch / TensorFlow for CNN and Grad-CAM++
HuggingFace Transformers for TrOCR-based OCR
EasyOCR for printed medical report extraction
Database Layer: NeonDB + Drizzle ORM
Model Serving: FastAPI or Flask for model endpoints.

IV. RESULT

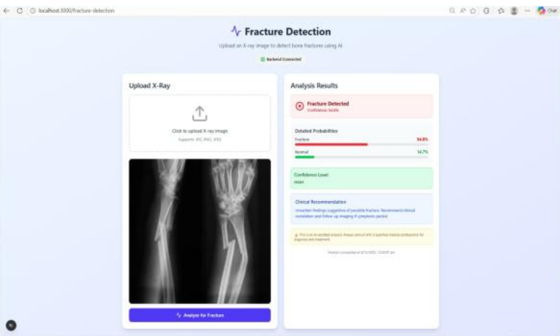

Fig 5: The fracture detection module will analyze the uploaded X-ray and show the area fracture probability with clinical recommendations. The interface clearly represents the confidence scores and diagnostic insights coming out of the CNN + Grad-CAM++ model.

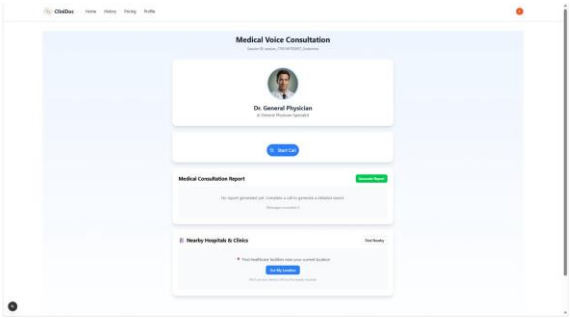

Fig 6: The Voice Consultation page allows for real-time communication with the AI doctor by using VAPI and AssemblyAI. It also provides options to generate the medical report and to view nearby hospitals based on the user's location.

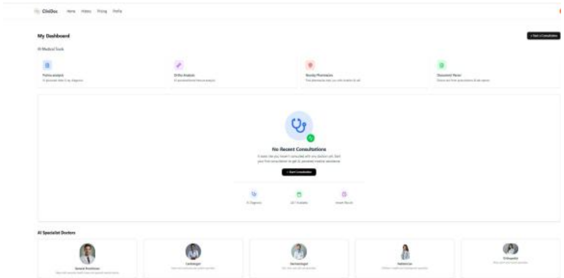

Fig 7: It gives access to all AI medical tools, such as X-ray analysis, fracture detection, nearby pharmacy searches, and document parsing. It also shows the available AI specialist doctors and allows the user to start new consultations.

V. LIMITATION AND FUTURE ENHANCEMENT

Limitation

Despite its advanced multimodal architecture, X-ray fracture detection, automated doctor allocation, and medical document analysis by AI Medical Voice Agent are burdened with several technical constraints that affect the real-world clinical reliability of this AI. Though the system integrates speech-based consultation, LLM-driven clinical summarization, dual-stage OCR, and CNN-based imaging diagnostics, performance remains bounded by the limitations inherent to current medical AI technologies. First, only preliminary medical guidance can be provided through this system, not replacing licensed clinical expertise. While GPT reports strongly resemble human-generated medical documentation, they remain sensitive to the quality of transcripts and completeness in patient responses during voice consultations. Any transcription error, missed context, or ambiguous phrasing can result in partial or inaccurate clinical summaries. Second, the automated doctor allocation mechanism uses a rule-based symptom classification-based approach. This keyword-driven approach could misinterpret vague, overlapping, or multi-symptom descriptions, assigning an inappropriate specialist, thereby impacting consultation quality. Third, the document analysis module, which leans on EasyOCR for printed content and TrOCR for handwritten prescriptions, may still face challenges for extremely noisy scans, low-resolution uploads, or highly stylized handwriting. EasyOCR's CNN–LSTM architecture may generate low-confidence outputs for handwritten content, whereas TrOCR-although transformer-based and hence robust-can struggle with irregular image orientation or severe shadow artifacts. These OCR limitations may affect downstream reasoning accuracy. Fourth, the CNN-based X-ray module has its functionality restricted exclusively to fracture detection and cannot identify a broader range of radiological abnormalities, such as infection, malignancies, degenerative changes, or soft-tissue conditions. In addition, Grad-CAM++, though useful for interpretability, may produce imprecise heatmaps when the X-ray quality is low, or where contrast levels vary significantly. Fifth, the voice consultation can vary greatly with environmental noise, accent variations, background disturbances, and network instability. Although AssemblyAI provides state-of-the-art speech recognition, real-time ASR may fail under fast speech, code-switching, or emotionally distressed patient communication. Finally, the lack of deeper clinical context integration, such as laboratory value interpretation, medication-safety cross-checking, or longitudinal health tracking, restricts the ability of this system to deliver holistic, end-to-end clinical decision support comparable with traditional care settings.

Future Work and Road Map

The short-term roadmap focuses on increasing the diagnostic accuracy and broadening the multimodal capabilities of the platform. First, we would extend the CNN fracture detector to be a multi-pathology model that can detect pneumonia, osteoarthritis, soft tissue abnormalities, and early degenerative patterns on X-rays, with possible future extension to MRI and CT imaging. Extend Grad-CAM++ to support multi-region interpretability and increase transparency clinically. For document processing, improve the OCR pipeline by domain-specific fine-tuning of EasyOCR on medical fonts and TrOCR on diverse handwriting patterns to make document extraction consistent across prescriptions, lab reports, and discharge summaries. Integrate a medical-NLP module powered by models such as ClinicalBERT or BioBERT on extracted text analysis. This allows for multimodal document-aware clinical reasoning without solely relying on summaries derived from conversation. The medium-term expansion will introduce symptom severity scoring with a probabilistic model and supervised learning-based refinement for enhanced triage prioritization. Multilingual ASR will allow regional languages to be understood, thus increasing accessibility in diverse populations. Emotion/stress detection using prosody, sentiment, and vocal biomarkers will be used to detect panic, pain, or distress for better risk assessment during voice consultations. The long-term goals will transform the system into a holistic telemedicine ecosystem. We will integrate wearable IoT devices for continuous monitoring of vital parameters such as heart rate, oxygen saturation, and sleep metrics. That allows for proactive risk alerts and longitudinal tracking. A telemedicine appointment bridge seamlessly escalates cases to human doctors when needed. More importantly, the platform will undergo clinical validation and be aligned with the UN SDG goals, notably SDG 3 (Good Health and Wellbeing) and SDG 9 (Industry Innovation and Infrastructure), to solidify its impact on healthcare accessibility, early detection, and AI-aided patient empowerment.

VI. CONCLUSION

The AI Medical Voice Agent showcases the novel, comprehensive, and intelligent approach toward overcoming persistent problems in accessibility, early diagnosis, and resource limitations across modern healthcare systems. By embedding advanced artificial intelligence paradigms, including real-time conversational consultation via Vapi and AssemblyAI, GPT-based clinical reasoning for structured medical report generation, a dual-OCR pipeline through EasyOCR for printed medical documents and TrOCR for handwritten prescriptions, and a Grad-CAM++ enhanced CNN for explainable X-ray fracture detection, the system goes beyond traditional telemedicine solutions while establishing a unified, multimodal framework for automated medical consultations. This will transform healthcare from static symptom input into a dynamic and contextual consultation capable of eliciting clinically meaningful insights. On this platform, triage accuracy increases with automated doctor allocations; moreover, diagnostic outputs in both textual and visual formats facilitate transparency, and usability increases due to the geolocation-based suggestion of hospitals and pharmacies. NeonDB allows the safe and scalable storage of medical data. Without replacing professional medical examinations, the system reduces diagnostic delays significantly, supports underserved communities, and provides direct AI-driven guidance to steer patients toward appropriate care pathways. Setting a new benchmark in building intelligent, integrated, and scalable digital health infrastructure, the AI Medical Voice Agent becomes a foundational technology for next-generation telemedicine and contributes to an accessible, responsive, and equitable global healthcare ecosystem.

REFERENCE

[1] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.

[2] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.

[3] Y. L. Thian et al., “Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs,” Radiology: Artificial Intelligence, vol. 1, no. 1, 2019, Art. no. e180001.

[4] T. Inoue et al., “Automated fracture screening using an object detection system in radiographs,” Scientific Reports, vol. 12, 2022.

[5] P. H. S. Kalmet et al., “Deep learning in fracture detection: a narrative review,” Acta Orthop. Belg., 2020.

[6] D. Van Veen et al., “Clinical Text Summarization: Adapting Large Language Models,” 2023.

[7] J. D. Oliveira et al., “Development and evaluation of a clinical note generation system using large language models,” Communications / Nature Digital Medicine, 2025.

[8] A. Ermolina et al., “Voice-Controlled Intelligent Personal Assistants in Health: Scenarios and Future Directions,” J. Med. Internet Res., vol. 23, no. 6, 2021.

[9] (Survey) “Speech Recognition for Medical Conversations,” various authors, ResearchGate, covering ASR challenges and solutions for healthcare deployments.

[10] T. Jadczyk et al., “Artificial Intelligence Can Improve Patient Management at Scale,” NPJ Digital Medicine, vol. 4, 2021.

[11] M. Li et al., “TrOCR: Transformer-Based Optical Character Recognition with Pre-Trained Vision and Language Models,” *Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[12] J. Jaided et al., “EasyOCR: A Modern Optical Character Recognition System Based on Deep Learning,” *arXiv preprint arXiv:2209.14132*, 2022.