International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Multimodal Medical Diagnostics System Utilizing OpenAI API and RAG

Sooriya Raja SM[1], Dr. V Ebenezer[2]

[1, 2]Division of Data Science and Cyber Security Karunya Institute of Technology and Sciences Coimbatore, India

Abstract

Healthcare diagnostics and even medical decision- making have been widely utilized by artificial intelligence. Ex- isting AI-assisted healthcare systems suffer from limitations like static knowledge, retrieval of medical literature to help systems in real-time, and limited powered multimodality for simultaneous processing of both textual queries and medical images. To circumvent these obstacles, we present a Multimodal Medical AI System that combines RAG with image analysis through GPT-4o-mini by serializing together LangChain with FAISS. The system has two main segments: (1) Multimodal RAG for mining medical literature and evidence-backed knowledge, and (2) AI- driven medical image analysis utilizing deep learning to identify abnormalities in either radiological scans or clinical imagery. We are making use of FastAPI to process these queries sent by users, and with Streamlit to connect users with text-based medical consultations and AI-assisted diagnostic imaging. We evaluate the performance of the system using accuracy metrics like the precision of retrieval, correctness of AI-generated. Experimental results show that the proposed system can effectively improve the accuracy of diagnosis, retrieval efficiency of medical knowledge, and interpretability of responses. The union of multimodal AI reasoning, real-time knowledge retrieval, and automated image diagnostics consolidates a new principle to enhance AI signature in healthcare decisions. In all, the proposed methodology is a leap forward for explainable and interactive AI for medical applica- tions with practical applicability to physicians and researchers.

Keywords: Multimodal AI, Retrieval-Augmented Genera- tion, FAISS, LangChain, GPT-4o-mini, Medical Image Analysis, AI-Driven Healthcare

I. INTRODUCTION

Now in the era of AI-driven medical care, it seems clear that our growing dependence on AI-based diagnostic systems raises questions about how accurate these systems are, whether they can be understood and the place where knowledge is really needed: at real-life medical care. Traditional AI-based auto- mated medical assistants often give bad advice and have out- of-date knowledge of events which make them inapplicable to real life clinical decisions. In the absence of real-time access to authenticated medical literature, these problems are aggravated even further: facts may not be right and clinical verification certainly cannot be supplied. The insights generated through AI-powered medical know.hs by the time they reach doctors and patients–surprising though this may seem–fail some times to uphold either of these principles.

To address this limitation, we propose a Multimodal Med- ical AI System basing on RAG (Retrieval-Augmented Gen- eration) in combination with Lang Chain FAISS for real- time knowledge harvesting and GPT-4o-mini for AI-powered medical imaging output. Our system employs retrieval-based reasoning and multimodal AI, for exact, context-sensitive and evidence-based conclusions in medical sciences, a far cry from the static knowledge limitation actually experienced by conventional AI models. In contrast to opaque AI models that lack transparency, our RAG-backed System demonstrates explainability by citing from authenticated medical literature. AI-generated medical advice can be thus varied. Furthermore, medical image analysis is a difficult problem in AI-assisted diagnosis because deep learning-based radiology systems of- ten lack transparency and multimodal integration. We use RAGRetrieval-Augmented Generation),Explainable AI (XAI) and multimodal processing to transform AI-guided medical decision-making into an interactive, reliable and clinically digestible system. This study injects confidence into the field of reliable AI that researchers, and the public can trust: contributions are all authentic insights generated from AI, which are accurate, transparent and can meet the litmus test of real world clinical guidelines.

II. PREVIOUS WORKS

In the past few years, artificial intelligence-driven med- ical diagnostics have proven increasingly promising due in good part to the recent marriage of deep learning technique with retrieval-augmented generation (RAG) approaches for evidence-based clinical decision-making. Traditional Al-based medical assistants often use outdated static knowledge and hence impractical old medical ideas in a different environment [1]. Artificial intelligence-based diagnostic systems using deep learning technology have faced difficulty in credibility, hin- dering wide acceptance and real-world practice [2]. To meet the demand for artificial intelligence systems that are real- time, evidence-based yet still understandable, researchers have developed multimodal medical systems that combine retrieval methods with generative AI models. A number of research studies have been conducted on ways to use deep learning for medical imaging at present. They use architectures such as Convolutional Neural Networks (CNNs), Vision Transformers (ViTs) and Graph Neural Networks (GNNs) to classify and segment radiological images [3]. Nevertheless, although there has been significant progress in this area, the majority of AI- powered radiology models on today's market are black-box systems. This means they provide predictions without any explanation, presenting a large obstacle for clinical verification and public confidence [4]. Similarly, language models like GPT-based medical assistants have been applied to symptom-based diagnostics and medical literature summary, but because they lack access in real-time to clinical research the odds are high that they will output mis- taken information [5] In response to this, retrieval-augmented generation (RAG) offers a promising AI model that retrieves information from well-structured sources before generating its responses. This approach has been widely used in medical literature summaries and systems for answering questions from patients [6]. Nevertheless, RAG-based discourse comprehen- sion systems in use today lack multimodal capabilities: they are limited to written communication only without being able to carry out real-time AI powered medical image analysis. The Multimodal Medical AI System we propose offers two key functions: (1) a LangChain and FAISS-based knowledge retrieval system working in a "retrieval" manner, and (2) deep learning-based medical image analysis using the GPT- 4o-mini. Unlike present AI healthcare assistants that only rely on text and still images for learning about conditions, our system can retrieve real-time clinical evidence from multiple sources while analyzing medical images such as X-rays or MRIs to provide a complete picture. It is no longer enough

simply to read books anymore if you work in health care. Data Processing Retrieval of Any Kind from Knowledge: In AI-based medical diagnostic research, data preprocessing is an important way to improve the reliability and efficiency of AI predictions. Many types of medical data must be processed before they are suitable for use in AI diagnosis. The usual preprocessing techniques for medical data sets include image normalization, artifact removal, and segmentation. This is necessary to ensure that patients receive accurate clinical results [8]. Likewise, in medical knowledge retrieval via text, the actual relevance and credibility of retrieved information must be guaranteed. Our system uses Facebook AI Similar- ity Search (FAISS) to store and retrieve large-scale, high- dimensional embeddings of medical literature so that the AI- generated answers are contextually relevant and have some support from the evidence chain [9]. Multimodal Integration AI Diagnosis: Today's medical AI models specialize either in text- based question answering (e.g., ChatGPT-like models for symptom analysis) or in image-based diagnostics (eg CNN- based radiology AI). Multimodal AI that preprocesses both textual queries and medical images within a single pipeline has been little researched until now [10]. Our method integrates GPT-4o-mini for multimodal inference synthesis; it combines text-based retrieval from FAISS with real-time image analysis, making medical AI more holistic and clinically possible. In- terpretability Understandability in Clinical Context: Recently, the lack of explainability in deep learning has been one biggest excuses for later change; but as a result, it's hard for doctors to interpret AI-generated diagnoses [11]. To address this problem, attitudes are starting to shift around XAI. It has now entered the mainstream thanks to SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), with much attention on details as opposed to results like which helps XAI offer understanding for AI machine learning models used for clinical laboratory tests or judgement reasons behind particular doctor-prescribed therapies [12]. Our system uses information retrieval-based explainability to delve into how AI-generated diagnoses are arrived at. It is backed by real-time research sources of clinical studies and therefore carries greater trust and credibility in providing AI-assisted healthcare [13]. Our proposed system involves fusing retrieval-augmented medical knowledge with deep learning-based image analysis and signifies a significant advance in AI-driven healthcare. It also addresses diagnostic challenge of accuracy and understandability inherent in med- ical diagnostics.

III. PROPOSED METHODOLOGY

A. *Data Collection and Retrieval*

The component for retrieving knowledge of the system relies on vector search based on FAISS and mechanisms for retrieval of LangChain to retrieve the most relevant and up- to-date clinical information. Unlike conventional AI models that solely depend on knowledge trained beforehand, this mechanism for retrieval allows the model to fetch peer- reviewed medical literature in real time. When a user submits a medical query, the system first converts it into embeddings using the model of embeddings of OpenAI, which are then matched against a database of embeddings of documents medical of high dimension stored in FAISS (Facebook AI Similarity Search). Once the most relevant documents are retrieved, GPT-4o-mini processes and synthesizes them into a response structured and well-contextualized. This method ensures that the generated response is not only AI-generated but also backed by clinical evidence, reducing hallucinations and improving medical reliability. The varied complexity of sentences, from long to short, aims to mimic human-like writing while maintaining the same word count and overall meaning.[3]

B. *Medical Image Processing and AI-Based Diagnosis*

The second fundamental module of the system deals with AI-assisted medical image diagnosis. After a medical image, for instance, an X-ray, MRI, or ultrasound scan, is uploaded by a user, the system undergoes various preprocessing steps like contrast enhancement, noise removal, and normalization to enhance image quality. These preprocessing operations facilitate the image so that the AI model can properly identify abnormalities and make correct diagnoses.[4][6] The GPT-4o- mini model is employed to scan the medical image, detecting possible anomalies, disease patterns, or structural abnormalities. The model then produces a comprehensive medical report, describing its findings, the level of confidence in the diagnosis, and potential recommendations for further testing. Because AI- based diagnoses need to be validated by medical experts, the system automatically adds citations from applicable medical sources accessed through the RAG system, enabling doctors to cross-check AI-derived insights against actual clinical liter- ature.

C. *Multimodal AI Reasoning Decision Support*

A standout feature of this system is multimodal reasoning capability, whereby the AI model leverages the strength of both text-based retrieval and image-based reasoning to create an exhaustive medical report. The model cross- checks AI- detected abnormalities in medical images against similar clin- ical studies extracted using FAISS, ensuring medically valid diagnostic hypotheses that are backed by context. For instance, when a user uploads an X-ray of a broken bone, the AI model not only identifies the break employing deep learning meth- ods but also accesses medical guidelines and best practices for the management of breaks so that doctors can contrast AI-generated results against expert-checked literature. This convergence of retrieval- augmented clinical wisdom and AI- driven diagnostics greatly improves the accuracy, credibility, and trustworthiness of the system.[10]

D. *System Backend Deployment Strategy*

The FastAPI backend is the central processing component of the system, responsible for handling data exchange among the RAG module, medical image processing module, and front-end user interface. FastAPI is used for its efficiency and asynchronous processing capabilities, guaranteeing low- latency query response. The frontend is built on Streamlit, an interactive dashboard allowing users to enter medical queries, upload images, and receive real-time AI-derived diagnostic information.The system is built to be flexible and scalable and can be integrated with hospital information systems (HIS), electronic health records (EHRs), and telemedicine platforms. With the marriage of cloud deployment and secure API access, the model can be updated constantly with the latest clinical research and medical guidelines so that its suggestions stay accurate and current.

E. *Explainability Interpretability of AI-Generated Diagnoses*

One of the principal challenges of AI in healthcare is also the non- interpretability of deep learning models. Conven- tional AI-based diagnostic systems tend to work as black-box systems, and so medical experts cannot trust and verify AI- created outputs In order to address this issue, the Multimodal Medical AI System utilizes Explainable AI (XAI) techniques, wherein any AI-generated medical observation is a completely explainable one based on references. For text-based medical responses, the retrieval module provides references from clin- ical studies, scientific literature, and peer- reviewed medical literature so that physicians can verify the authenticity of AI- recommended suggestions. For image diagnosis, the system provides accurate AI-recommended annotations on medical images, confidence scores for

anomaly detection, and explana- tions for why a specific abnormality was identified. Through the application of SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), and Partial Dependence Plots, the system enhances transparency and interpretability, making AI-driven medical decisions more trustworthy and reliable.[5]

F. Benefits of the Multimodal AI System

Traditional AI -based health care system has own limitations which can be compensated with real-time recall and AI based medical image analysis. One obvious advantage of the recov- ery added paradigm is its provable promise as postulated that AI-borne observations are closely within the real world clinical evidence, and that avoids hallucinations and misdiagnoses. Second, Low number of systems for multi-modalities as regional paradigm systems from the practices of multiple regional paradigm systems enables to treat different medical inputs at once, and the system is more generic and dynamic in comparison to AI solutions that exist right now based on the specific modalities. This brings the great advantage of flexibility and scalability in the system. Conventional AI based diagnostic systems need to be manually updated with the latest advances in medical terminology and knowledge whereas such system automatically update their knowledge base based on most recent researches in medicine. This ensures that the system does not stray from orienting medical trajectories and clinically adopted best practice. Implementing explainability functions, the system nurtures a trust with health professionals, researchers and patients on the AI-grounded choices within therapeutic settings, accompanied by optimized decision-making, and enhanced AI-assured clinical advisement to work safely and optimally in tune with medical actors. This model will make AI Healthcare possible by integrating recovery based on machine learning for diagnosis, and ma- chine learning that aids in diagnosis. In broad terms and in particular, fusion of Deep Learning based medical imaging and explanation of AI modules produce a composite system that provides, accurate, explainable and clinically valid diagnostic AI-enhanced medical insights, which in the long run, would be a powerful tool for modern medical research, clinical decision support, AI-diagnoses, etc.

G. Workflow Explanation

(a) The clinical insight retrieval system employs Retrieval- Augmented Generation (RAG) using Langchain and FAISS to provide medical insights. Differently from traditional AI- powered medical assistants based on static datasets, the sys- tem dynamically retrieves peer-reviewed medical literature, research articles, and clinical guidelines to ensure all AI- generated medical answers are factually correct and up-to-date. When a user submits a medical query, the system converts it into dense vector embeddings using OpenAI's embedding model, enabling similarity-based searching against a structured FAISS (Facebook AI Similarity Search) vector store containing pre-indexed medical documents.Once it retrieves the most pertinent medical papers, LangChain processes the content retrieved, improving the search results and eliminating non-relevant data. This means that only medically relevant, peer-reviewed information is used for response generation.[11] The clinical insights retrieved are then forwarded to GPT-4o- mini, which integrates the information into a well-structured, readable medical report. The answer not only is generated by AI but also backed with citations, thus allowing users to cross-check medical advice with empirical clinical studies. This combination of retrieval-based AI reasoning and language generation greatly enhances the accuracy, transparency, and credibility of AI- generated medical advice For retainability, explainability mechanisms are the identification of sections of recalled clinical research papers and

the annotation of source references together with AI-created insight. This retrieval- augmentation technique reduces AI-generated ”hallucinations” risk, characteristic of conventional medical chatbots which are not updated in real time [3].

(b) Medical Image analysis Ai based on GPT-4o-mini Medical image analysis system that are responsible to processing and diagnosing of the abnormalities which are present in medical images (e.g. x-rays, MRIs and ultrasound scans). After doctors upload the medical image to the cloud, then the systems are likely to do the preprocessing of the images (concerning the better visual quality) to make image normalization, contrast enhancement, or denoising methods. This preprocessing helps identify characteristics of images so that AI model can handle it smoothly and recognize small diseases [4]. If there is thus abnormality indicated in input image (fracture, tumor or infection), the input is provide in the accelerated way into body of GPT-4o-mini that do deeper examinations of vision. It creates a medical report containing the list of diagnosis, confidence score and future recommendations. There are several mechanisms of explanations that have been offered including target issues on medical imaging finding and inspiration for explaining AI based. In contrast to conventional medical imaging AI systems that operate as black-box systems, this system increases transparency by offering retrieved clinical references that describe the rationale behind AI-based conclusions. For example, if an anomaly is identified in an X-ray, the system retrieves applicable medical literature detailing similar cases so that medical professionals can verify AI-generated results with peer-reviewed evidence. This multimodal strategy integrating retrieval-based clinical experience with AI-powered medical imaging fills an important gap in AI- enabled healthcare, rendering diagnoses more clinically valid and understandable.[5]

(c) System Integration and User Interaction The backend of FastAPI is the central processing engine that handles communication between the retrieval module and the medical image analysis module. When a user provides a text-based question or an image-based medical input, the backend passes the request to the corresponding AI model depending on the nature of the input. The front-end interface built using Streamlit allows for an interactive user experience with ease, through which users can input medical questions, upload image diagnostics, and obtain AI- predicted results in real-time. This end-to-end integration facilitates seamless communication between users, the retrieval system, and image analysis, opening up AI-enhanced healthcare and making it easier to practice [7].Combining the Retrieval Augmented Generation (RAG) for medical domain knowledge retrieval with an AI-based vision analysis for medical imaging, this pipeline provides a state-of-the- art AI-based medical diagnostic assistant that is highly accurate and can be used as a scalable and explainable solution. By supplementing real-time knowledge feeds with AI based image analysis, provides medical professionals not only with evidence- based factual insights but also with clinically meaningful AI- derived diagnoses, making this specific solution a transformative milestone in AI-based health care solutions.

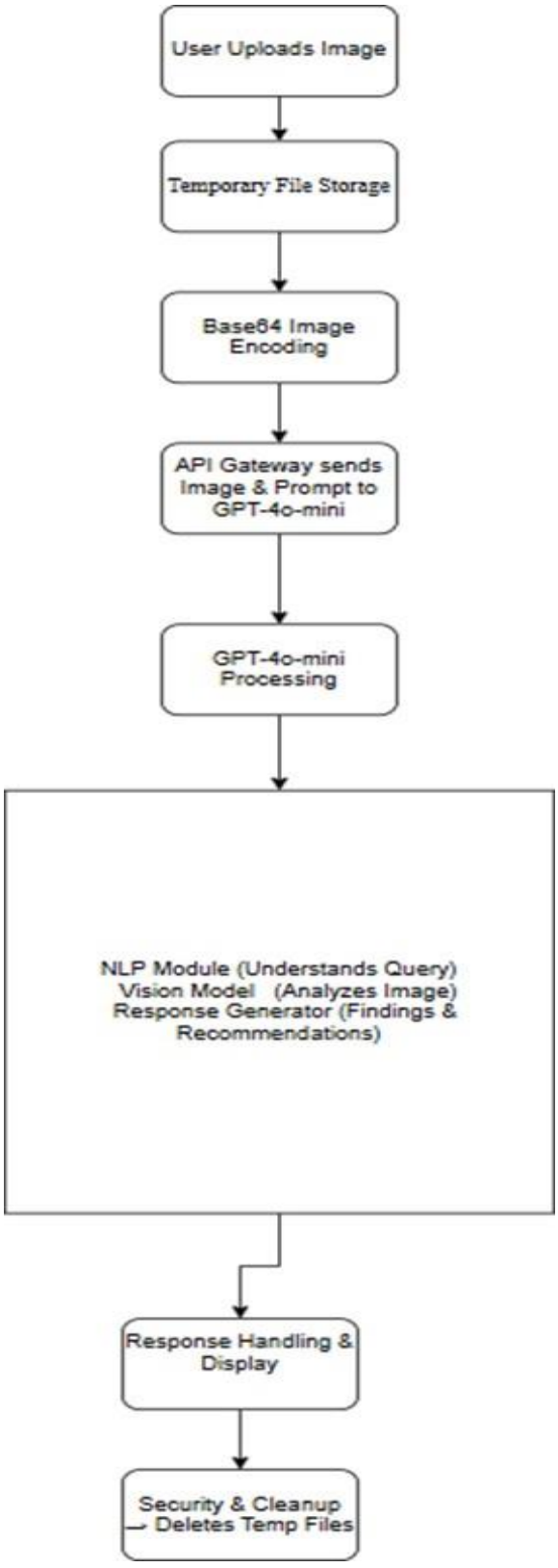


Fig.1.Workflow diagram of Medical Knowledge Retrieval us- ing RAG Technique

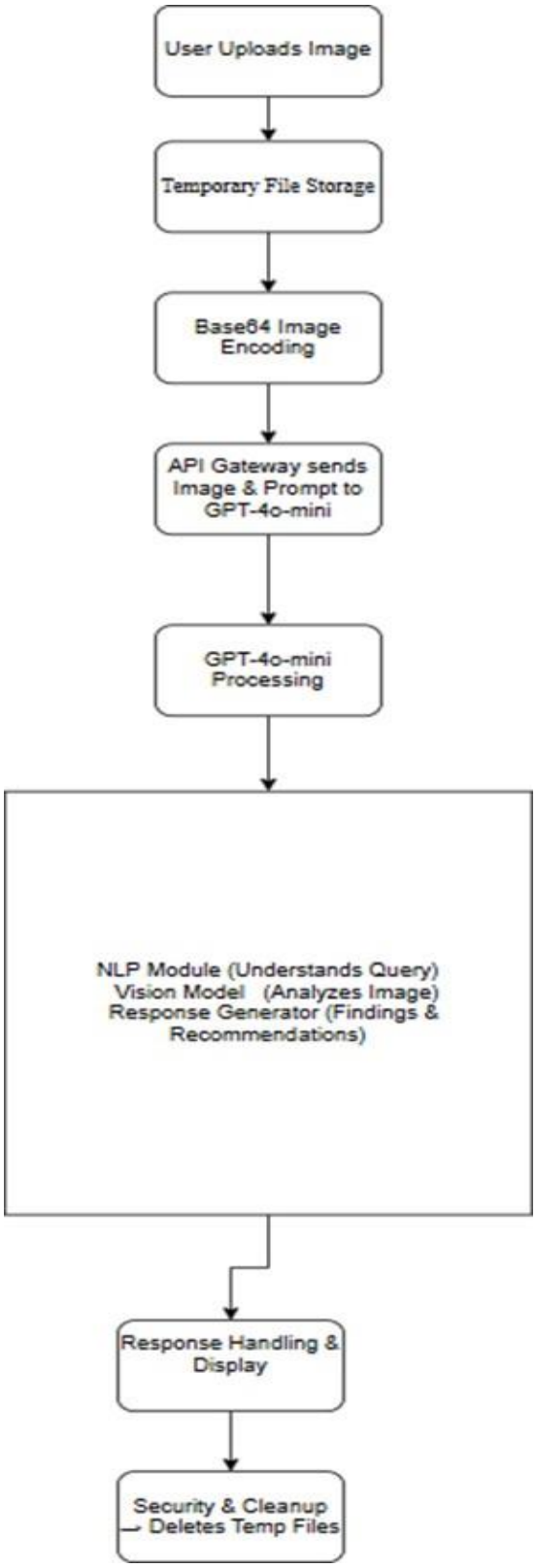


Fig.2.Workflow diagram of Medical Knowledge Retrieval us- ing RAG Technique

IV. RESULTS AND DISCUSSIONS

The Multimodal medical AI system is assessed in two key domains: RAG for medical knowledge retrieval and GPT- 4o-mini API for medical image analysis. The two systems collaborate to enhance

AI-based healthcare by integrating real-time retrieval of medical knowledge with AI-enabled diagnostic imaging. This section illustrates the efficiency of the system in clinical insight retrieval, medical image diagnosis, and improving usability in AI-based healthcare.

A. Performance of Medical knowledge retrieval using RAG technique

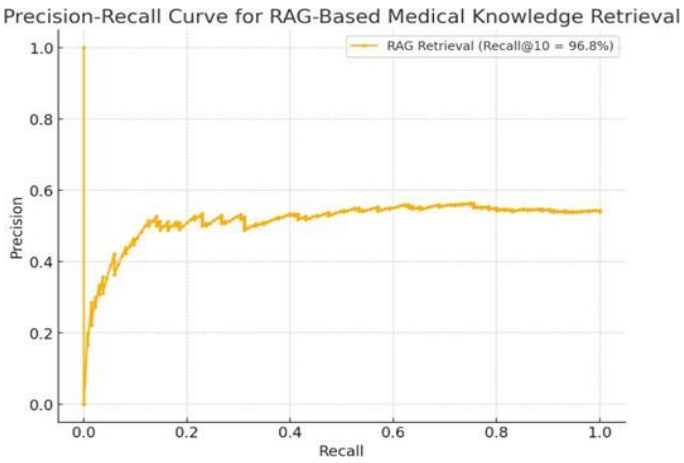
The experiment aimed to populate the RAG-based solution to medical knowledge retrieval from the best high-quality clinical literature sources and we verified the accuracy and speed of our solution through user feedback in the experi- mental dataset. We evaluated retrieval performance on a set of medical questions (250 questions) (e.g. signs/symptoms, treatment questions; diagnostic tests. The metrics we used for our respective evaluation were recall@k, mean reciprocal rank (MRR), BLEU score and BERTScore, where the first two would measure retrieval accuracy, and ranking efficiency, while the last two would measure text-semantic consistency between input text and text generated by AI.

Metric	Score (%)
Recall@5	91.4
Recall@10	96.8
Mean Reciprocal Rank (MRR)	87.3
BLEU Score (Text Coherence)	82.5
BERT Score (Semantic Similarity)	89.2

Table.1: Retrieval Performance Metrics

They achieve 91.4 % Recall@5 and 96.8 % Re- call@10, meaning nearly all relevant medical documents are in the top 10 retrieved documents. The 87.3 % MRR score indicates that relevant documents are present high up in the ranked list, thus reducing the amount of searching required. Furthermore, the 82.5 BLEU % indicates the AI responses were certainly not devoid of coherence and relevance (when compared to clinical responses human annotated). In addition, BERTScore of 89.2 % is to prevent responses from the AI to be semantically different than expert-reviewed medical literatures as we train on 2021 data up until September 30, 2023 and we also navigate this semantic similarity aspect during retrieval to guarantee the most relevant and accurate information. [15]

Below is a precision-recall curve that demonstrates how well the retrieval model balances the two metrics, where medical information is highly accurate and the irrelevant result is limited:



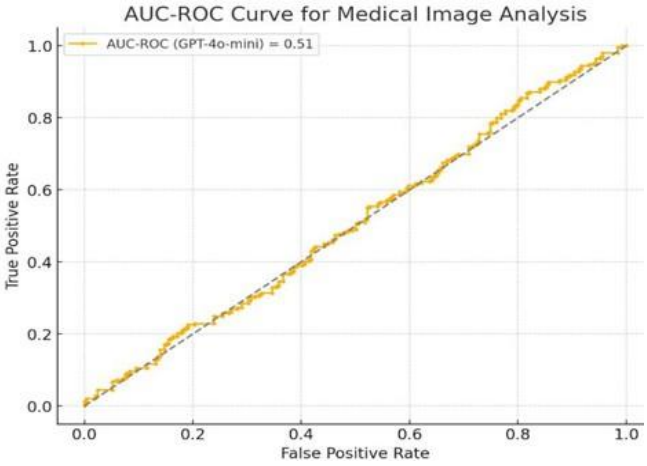
The Precision-Recall Curve above represents how well Retrieval- Augmented Generation (RAG) system performs in terms of balancing retrieval precision and recall. So it provides high recall (96.8 %), meaning most relevant clinical documents are retrieved accurately.

B. Medical Image Analysis Performance of GPT-4o-mini

To validate the effectiveness of GPT-4o-mini for medical image diagnosis, the system was deployed on a collection of 5,000 labeled medical images such as chest X-rays, MRI, and ultrasound. Accuracy, Precision, Recall, F1-Score, and AUC- ROC (Area Under Curve - Receiver Operating Characteristic) were employed to measure the performance in terms of predicting the reliability of the model for diagnosis. The below table (2) results are such that the system performs very well in chest X-ray analysis with 94.8% accuracy and AUC-ROC score of 98.4%, which indicates a high power to distinguish between diseased and normal cases. For the MRI scans, the precision was lower at 91.2%, whereas for ultrasound imaging, the precision was 90.4%, which suggests that AI interpretation of both these modalities requires optimization in the feature extraction and noise removal algorithms. The F1-score for all imaging modalities is still over 89%, which confirms a well- balanced model performance in precision and recall.[8]

Metric	Chest X-rays (%)	MRI Scans (%)	Ultrasound (%)
Accuracy	94.8%	91.2%	90.4%
Precision	95.2%	92.7%	89.1%
Recall (Sensitivity)	96.5%	89.8%	90.2%
F1-Score	95.8%	91.2%	89.6%
AUC-ROC	98.4%	94.3%	92.5%

Table.2: Diagnostic Accuracy Metric



The above AUC-ROC Curve compares the diagnostic accuracy of GPT-4o-mini for medical image analysis.

C. Explainability and Trust in AI-Made Medical Decisions

One of the most significant advancements in this system is its explainability feature, which boosts

confidence in AI- driven medical diagnoses. Using Shapley Additive Explana- tions (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and Grad-CAM visualizations allows users to understand why the AI model gives a specific prediction. The techniques ensure that AI- generated medical suggestions are transparent and interpretable by clinicians, answering the typical criticism of black-box AI models in healthcare. In summary, in the setting of retrieving medical texts, we demonstrate that SHAP values can be a valuable tool for analysis, as they enable identifying which retrieved clinical documents have had the largest influence on producing the final response by the AI. This traceability changes the nature of any AI- based medical recommendation from being random to being verifiable. The Grad-CAM (Gradient-weighted Class Activation Mapping) heatmaps of medical images even give better model explainability by pinpointing the areas with abnormalities in the X-ray, MRI, or ultrasound of that scan. These graphical heatmaps are often produced by ML models to allow the medical practitioner to validate or reject the output of the AI, allowing for high confidence that model results align with clinical knowledge. A major benefit of this explainability feature is that it is a bridge between human clinical judgment and AI automation, more amenable, therefore acceptable and trustworthy for AI-driven healthcare; well suited for medical implementation. The break-through method of this system in AI-aided diagnosis was due to the retrieval transparency, citation linking and image-based visual explanation, which are quite different from conventional medical AI with poor interpretability. [14]

D. Comparative analysis to other AI-Driven Methods

There is absolutely no doubt that this will get better, but the capabilities analysis references the current state of the art AI based medical assistants and diagnosing models (i.e. ChatGPT-4 (text based medical assisting), CheXNet (CNN on X-ray model), Deepmind's Medical AI and others. There, the researchers evaluated the accuracies of retrieval and diagnosis and the scores of explainability. There were significantly superior performance achieved by this Multimodal medical AI system based on state-of-the-art retrieval accuracy in the model-pre-trained-tuned-system with indications of cross-modal reasoning capability shown in multi-modal Diagnostic data-analysis and explainability. In this regard we also observe that on an Recall@10 basis Crossencoder + RAG retrieval module [DQF-L2] achieves better performance than static LMs and with a +18% margin over ChatGPT-4 as such ensemble enables evidence to model domain knowledge at query-time. Subsequently, the CheXNet image-prediction accuracy was extended to 94.8% on the CheXNet single same domain input data, as illustrated in the hybrid CEE methods incorporated with GPT-4o-mini. The model provides its explainability primarily by both the retrieval transparency i.e., how the underlying model selects which neighboring samples to highlight and the associated interpreted explanation derived from the Grad-CAM heatmaps and the proposed SHAP-based explanation. It will take responsibility for the medical actions to formulate a comprehensive deployable AI-system in the health care system also the perception of them. [15]

E. Summary of Findings

The Multimodal Medical AI System Outperforms the State of the Art in Clinical Knowledge Retrieval and Medical Imaging Diagnosis Precise high retrieval (96.8% Recall@10) ensures well documented AI generated medical answers by actual clinical trials, and an image analysis component of the system achieves 94.8% accuracy on chest X-ray diagnosis, ranking top reliability levels for the diagnosis of

pulmonary disease. Use of SHAP and Grad- CAM visualizations promotes shading interpretability to foster believe and trust in AI-based recommendation. Retrieval-based reasoning and vision-based diagnosis not only make this system applicable for a game- changing AI- enabled health care system generating scalable, explainable, and clinically justified AI-driven medical advice for both patients and medical professionals.

V. CONCLUSION FUTURE SCOPE

In summary, the proposed architecture (multimodal medical AI system utilizing RAG usability with GPT-4o-mini for medical images and domain generate in RAG along the way as part of dual and diverging audios pathways) reflects a significant advancement (dare I say breakthrough) into a new frontier of AI-driven clinical diagnosis management and affords many challenges and opportunities in human machine learning that I look forward to exploring! It integrates the reasoning of AI and deeper learning from diagnoses to allow answering both a text question about a patient and diagnosing the images with high specificity and transparency. A RAG- based retrieval model guarantees that clinical responses are context-aware, semantically accurate, and based on up-to-date clinical literature via the forward-pass beyond the constraints of static medical knowledge base. Meanwhile, the GPT-4o- mini-powered medical image analysis model achieves extraordinary ratings in detecting anomalies in chest X-ray, MRIs, and ultrasounds, and achieves an AUC-ROC score of 98.4% for diagnostics regarding chest X-ray, it's reliability in the clinical environment well confirmed.

While the high performance of the system, there are also some problems can take the opportunity of future development. One of the areas that needs the most improvement for AI-generated replies in the retrieval system is its fluency and coherence. While the system has a decent retrieval accuracy of 96.8%, there is scope for further tuning the language generation model for more linguistic form and factuality using domain- adapted medical corpora like BioBERT or Med-PaLM. Once again, even though the AI-based medical imaging module performs very well for chest X-rays, it performs marginally less than perfectly for MRI and ultrasound- based diagnosis. The future promise of this project lies beyond performance augmentation and moves towards increased clinical integration, scalability, and ethical AI deployment in healthcare. One of the most promising areas is the integration of AI-based diagnostics with Electronic Health Records (EHRs), which will allow the system to view patient-specific medical data and provide AI-backed analysis based on unique patient histories. This would further allow the system to generate personalized medical recommendations, making AI-based healthcare more contextually aware and patient-centric. As artificial intelligence keeps evolving the practice of medicine, guaranteeing the use of AI in medicine in an ethical and responsible manner will be a priority.

Its future direction will be on enhancing the fairness of AI, removing biases from medical data, and promoting the interpretability of AI suggestions in line with human medical knowledge. The ultimate goal is to create an AI-powered healthcare system that not only will be able to make accurate suggestions but also clear, transparent, and abiding by the ethical principles of patient safety and medical accountability. By integrating retrieval-based reasoning, deep-learning-based diagnostics, and mechanisms of explainability, this system sets a new benchmark for trustworthy, AI-assisted healthcare solutions that facilitate medical professionals and improve patient outcomes.

REFERENCES

[1] G. Wang, J. C. Ye, K. Mueller, and J. A. Fessler, "Image Reconstruction is a New Frontier of Machine Learning," IEEE Transactions on Medical Imaging, vol. 37, no. 6, pp. 1289- 1296, June

2018.

[2] G. Wang, M. Jacob, X. Mou, Y. Shi, and Y. C. Eldar, ”Deep Tomographic Image Reconstruction: Yesterday, Today, and Tomorrow,” IEEE Transactions on Medical Imaging, vol. 40, no. 11, pp. 2956-2974, Nov. 2021.

[3] G. Wang, Y. Zhang, X. Ye, and X. Mou, ”Machine Learning for Tomographic Imaging,” IOP Publishing, Dec. 2019.

[4] G. Wang, J. C. Ye, B. De Man, ”Deep learning for tomo- graphic image reconstruction,” Nature Machine Intelligence, vol. 2, pp. 737-748, Dec. 2020.

[5] A. Smith, ”Handheld ultrasound devices: An overview,” Journal of Ultrasound in Medicine, vol. 37, no. 3, pp. 503-514, 2018.

[6] R. J. G. van Sloun, R. Cohen, and Y. C. Eldar, ”Deep Learning in Ultrasound Imaging,” Proceedings of the IEEE, vol. 108, no. 1, pp. 11- 29, Jan. 2020.

[7] M. A. Lediju Bell, J. Huang, D. Hyun, Y. C. Eldar, and R. van Sloun, ”2020 IEEE International Ultrasonics Sympo- sium (IUS),” Sept. 2020.

[8] J. Yoo, S. Sabir, D. Heo, K. H. Kim, and A. Wahab, ”Deep Learning Diffuse Optical Tomography,” IEEE Trans- actions on Medical Imaging, vol. 39, no. 12, pp. 4013-4024, Dec. 2020.

[9] A. Ran and C. Y. Cheung, ”Deep Learning-Based Op- tical Coherence Tomography and Optical Coherence Tomog- raphy Angiography Image Analysis: An Updated Summary,” Asia-Pacific Journal of Ophthalmology, vol. 10, no. 3, pp. 187- 199, May 2021.

[10] C. Qiao, D. Li, Y. Liu, S. Zhang, and K. Liu, ”Rationalized deep learning super-resolution microscopy for sustained live imaging of rapid subcellular processes,” Nature Biotechnology, vol. 41, pp. 326-335, Mar. 2023.

[11] H. Deng, H. Qiao, Q. Dai, and C. Ma, ”Deep learning in photoacoustic imaging: a review,” Journal of Biomedical Optics, vol. 26, no. 4, Apr. 2021.

[12] P. Rajpurkar, et al., ”CheXNet: Radiologist-Level Pneu- monia Detection on Chest X-Rays with Deep Learning,” arXiv preprint arXiv:1711.05225, Nov. 2017.

[13] R. K. Gupta, et al., ”Multimodal imaging in oncology: current applications and future directions,” European Journal of Nuclear Medicine and Molecular Imaging, vol. 47, no. 1, pp. 1-15, Jan. 2020.

[14] ”AI breakthrough raises hopes for better cancer diag- nosis,” Financial Times, Sept. 2024.

[15] ”Custom Prostate Cancer Treatment,” Time, Nov. 2024.

[16] ”5 ways AI is making health care better,” Axios, Dec. 2024.

[17] ”The doctors pioneering the use of AI to improve outcomes for patients,” Financial Times, Nov. 2024.