IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 59, No. 7, July, 2026, pp. 2087-2094
Journal homepage: http://iieta.org/journals/jesa

MedRAG: A Multi-Modal, Privacy-Preserving AI Agent for Proactive Home Health Screening and Smart Triage

Srilatha Pulipati[1], D Vaman Ravi Prasad[2], Raja Sekhar Reddy P[2*], Y. Sowmya Reddy[3], P. Rajeshwari[2], K Sarvani[2]

[1] Department of Artificial Intelligence and Data Science, Chaitanya Bharathi Institute of Technology, Hyderabad 500072, India
[2] School of Engineering, Anurag University, Hyderabad 500088, India
[3] CVR College of Engineering, Hyderabad 501510, India

Corresponding Author Email: rajasekharreddycse@anurag.edu.in

https://doi.org/10.18280/jesa.590723

Received: 11 May 2026
Revised: 6 July 2026
Accepted: 25 July 2026
Available online: 31 July 2026

Keywords:
explainable AI in healthcare, home-based disease screening, Multi-Modal Retrieval-Augmented Generation

ABSTRACT

Proactive early disease screening is essential for timely interventions, particularly in underserved regions with limited healthcare access. Conventional large language models (LLMs) struggle with hallucination, limited multi-modal integration, and privacy concerns in medical diagnostics. This paper introduces a novel Multi-Modal Retrieval-Augmented Generation (MM-RAG) framework, embodied as an AI-driven agent, for privacy-preserving early disease screening using symptom descriptions and smartphone-captured images. Targeting conditions suitable for home-based assessment, such as dermatological changes, ocular symptoms, external gynecological indicators, and certain systemic or neurological cues visible on the skin or face. Our framework integrates Contrastive Language Image Pretraining-Vision Transformer (CLIP-ViT) for image encoding, Bidirectional Encoder Representation for Transformers (BERT) for text encoding, a Facebook AI Similarity Search (FAISS)-based retrieval engine, and ontology-aware prompt engineering with adaptive diagnostic orchestration. This orchestration dynamically tunes Chain-of-Thought (CoT) reasoning conditioned on retrieved medical knowledge and patient context. Federated learning (FL) provides privacy by training models on the edge devices locally. Our agent achieves a Top-1 diagnostic accuracy of 82.7%, a factual score of 91.4, and an explainability score of 82.6, outperforming baselines like GPT-4o and MedPaLM-2. This work fosters equitable healthcare through scalable, proactive, and privacy-preserving home-based screening.

1. INTRODUCTION

The global healthcare landscape is fraught with vast disparities, particularly in terms of the early detection and intervention of diseases. Medical science is progressing at an unprecedented pace, but the benefits of these innovations are often concentrated in well-resourced urban centers, leaving large populations in underserved regions with little or no access to timely diagnostics. The discussion on proactive early screening of diseases is not merely a clinical issue but a fundamental pillar within the field of public health. It has the potential to substantially lower morbidity and mortality rates, improve patient outcomes, and alleviate the economic burden of managing diseases at advanced stages. This important gap highlights the need for scalable, accessible, and effective screening methodologies that can bridge the gap between medical advances and their equitable distribution.

Recent advances in artificial intelligence (AI), particularly large language models (LLMs) and computer vision, have created enormous potential to revolutionize medical diagnostics. These technologies have an unparalleled ability to analyze massive amounts of data, detect complicated patterns, and produce profound insights. However, the direct application of traditional LLMs to medical diagnosis, especially in sensitive areas such as early disease detection, is fraught with considerable limitations. One of the most prominent is the pervasive issue of hallucination, where models produce plausible but false information. In a clinical setting, a hallucinated diagnosis or medical recommendation can have catastrophic consequences, undermining patient safety and clinical trust. Traditional LLMs also usually have limited multi-modal capabilities and are mainly designed for text, which makes them inherently struggle to smoothly incorporate and interpret different types of data, such as medical images, sensor data, or physiological signals. However, progress on multi-modal LLMs is emerging, but their robustness and interpretability for high-stakes medical decisions are still under active investigation. Perhaps most importantly, the use of centralized AI models for medical diagnostics presents grave privacy concerns. Sending sensitive patient data like symptom descriptions, personal medical history, and especially highly identifiable images to cloud servers for processing carries risks of data breaches, illegal access, and strict regulations such as HIPAA and GDPR.

These challenges collectively hinder the adoption of existing AI paradigms for privacy-preserving early disease screening at home in a widespread and robust manner.

In this paper, we propose a state-of-the-art AI-driven agent, a new Multi-Modal Retrieval Augmented Generation (MM-RAG) framework, to overcome the above limitations and push forward the privacy-preserving early disease screening. Our framework is designed in a way that it can exploit the complementary strengths of smartphone images and symptom descriptions for conditions that are especially suitable for home assessment. These include a variety of common and serious diseases, including skin conditions, gynecological conditions, respiratory diseases (e.g., pneumonia), and eye diseases. The deliberate selection of these target conditions reflects the resolve to tackle major public health challenges with affordable, non-invasive methods.

Our technological backbone consists of advanced implementation of cutting-edge encoders: the high-resolution Contrastive Language Image Pretraining-Vision Transformer (CLIP-ViT) for the purpose of feature extraction from images and the fine-tuned Bidirectional Encoder Representation for Transformers (BERT) model for textual semantic analysis. The resulting vectors are processed using a highly efficient Facebook AI Similarity Search (FAISS)-based retrieval engine, allowing for fast and accurate access to a large medical database. A unique solution implemented in our project is ontology-aware prompt engineering together with an adaptive diagnostic orchestration module. The adaptive orchestration allows for dynamic adjustment of CoT reasoning based on the retrieved information and the patient's data. Moreover, to meet privacy requirements without any doubt, we use federated learning in order to train the model on the device.

2. RELATED WORK

Many previous works have been conducted in the field of disease diagnosis with the help of AI and LLMs, image classification models, and multi-modal learning. However, the majority of those solutions work based on cloud computing, which is an issue of privacy and access. The current literature review will address existing approaches in multi-modal AI for healthcare, retrieval-augmented generation, and privacy-preserving medical inference. The early works in the field were published in 2021 and introduced the basis of multimodal health applications. Study [1] offered a privacy-preserving collaborative solution for COVID-19 diagnosis using federated learning to aggregate imaging and clinical data while maintaining patient privacy. An artificial intelligence system was developed [2] based on computer vision that incorporated patient clinical parameters and imaging; however, their design did not have capabilities of multi-modality search and robust privacy measures. "Curse of dimensionality" in digital medicine was considered by the study [3], and they provided solutions for coping with high-dimensional medical data, but their approach did not cover real-time screening and adaptive reasoning.

Progress towards personalized medicine was made in 2022. In their review of multi-modal AI. A noted use of biobanks, imaging, and wearables for remote monitoring [4], but emphasized the lack of standardized privacy measures. The RAG application to LLMs suggested that it might be used for medical purposes as well [5]. The study improved fairness in medical classifiers under distribution shifts using generative models [6], yet the study did not address privacy or home-based deployment.

A vision-language framework to generate chest X-rays in 2023 was developed in the study [7]; nevertheless, privacy has been sacrificed due to its centralized nature. The first time that whole slide images were generated using diffusion models was by the study [8], which tackled the challenge of limited data in addition to preserving privacy, but its scalability to various diseases was not demonstrated. A novel technique presented METRICS [9] as a checklist for evaluating generative AI models fort5f6t66 healthcare applications, focusing on evaluation criteria, although it does not cover multi-modal RAG systems. The use of ChatGPT for medical education, proposing text-based AI augmentation, is described in the study [10].

In 2024, specialized applications emerged in the study [11], which performed a scoping study with an increase in performance of 6.2% with the use of multi-modal AI compared to unimodal models, while interpretability and privacy were not fully explored. In their systematic review on uncertainty quantification in deep learning, which is necessary for effective diagnostics. Study [12] did not include RAG; ChatGPT for healthcare and suggested using a multi-modal decision-making system; however, there was no discussion on privacy-preserving methods [13]. A RAG system for the identification of facts about COVID-19; the system worked effectively but did not have adaptive orchestration for various conditions [14]. The AI tools for nephrology literature searches indicated potential for adaptive reasoning, though multi-modal integration was limited [15].

In 2025, a meta-analysis was performed for reporting guidelines for AI in radiology, promoting transparency, while multimodal screening approaches were ignored [16]. The use of generative AI in clinical pathology emphasized the importance of RAG and vectorization, but real-time validation at home was missing [17]. A review of multimodal learning in disease diagnosis pointed out the necessity of privacy-preserving measures [18]. Multimodal integration techniques are proposed for disease early detection, whereas privacy-preserving solutions were not considered [19]. A privacy-preserving multi-modal AI approaches, while adaptive orchestration was not mentioned [20]. The role of data fusion in diagnostics, while scalability to home environments was not considered [21]. The investigation of multimodal learning and clinical decision making, respectively, while RAG and privacy were not mentioned [22]. A transformer and NLP in healthcare is developed, while multimodal RAG was not integrated [23, 24]. The work proposed a broad overview of AI in healthcare, identifying privacy as a gap [25].

Deep neural networks can classify skin cancer from skin lesion images at a performance on par with dermatologists, thereby establishing the potential of deep learning as an instrument for computer-assisted dermatological diagnosis. [26]. BERT score, a BERT-based metric to assess semantic similarity between generated and reference texts [27]. An Adversarial NLI, a benchmark for evaluating robust natural language understanding, was developed in the study [28]. ROUGE, as an automatic metric to evaluate the quality of generated summaries, was proposed in the study [29]. UMLS was proposed as an integrated structure for relating and standardizing biomedical terminology [30]. Chain-of-thought (CoT) prompting improves the reasoning ability of LLMs [31].

The accuracy has been greatly enhanced owing to the emergence of retrieval-augmented generation (RAG), which is

an improvement in the validity of AI-powered systems for use in healthcare applications. A comprehensive review of RAG models in healthcare demonstrates how the use of RAG and helps mitigate the challenge of hallucination and facilitate clinical decision-making through external knowledge retrieval [32]. The multimodal RAG models, which utilize both text and images for reasoning, enhance the performance of medical visual question answering and clinical reasoning [33]. A healthcare support system known as MedRAG is limited by the quality of the knowledge graph and the retrieved data (Figure 1) [34].

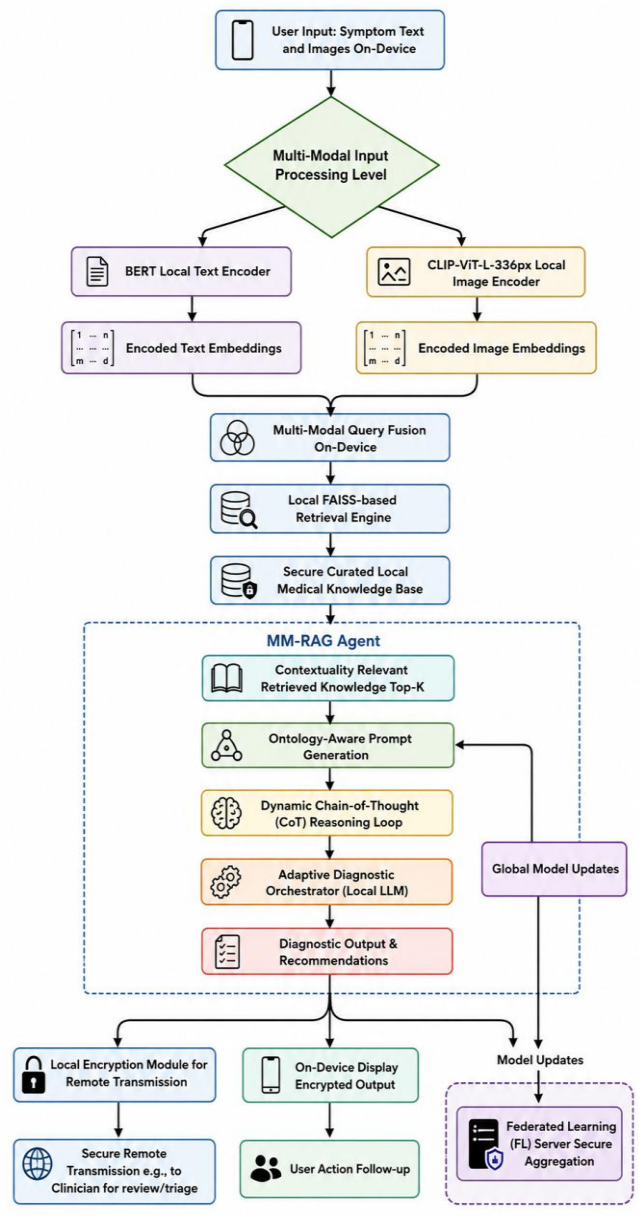

Figure 1. MedRAG architecture

3. PROPOSED METHODOLOGY

This section delineates the architectural design and operational principles of our novel Multi-Modal Retrieval-Augmented Generation (MM-RAG) framework, instating an intelligent AI-driven agent for privacy-preserving early disease screening. We detail the integrated components, their interdependencies, and the algorithms that govern the agent's adaptive diagnostic orchestration and privacy preservation.

3.1 Architectural overview

The architecture is deliberately designed to be both secure and inherently private, consisting of discrete yet highly interwoven stages: multi-modal input encoding, secure local knowledge retrieval, adaptive diagnostic orchestration, and secure output generation. The basic concept of federated learning is present throughout the whole system such that when training models using patient data, the data will be safely stored on the edge device.

3.2 Multimodal input encoding

The efficacy of our MM-RAG framework is dependent on its ability to robustly understand and integrate diverse data modalities.

3.2.1 Text encoding for symptom descriptions

The symptoms given by patients in natural language sentences are handled by an advanced version of the BERT model. In a case where $S=\{W_1, W_2, W_3, \ldots, W_N\}$ is a symptom description, where W_i are tokens, the local BERT encoder generates a fixed-dimension embedding vector, $e_S \in R^d$ as follows:

$$e_S = BERT\,(Symptom\ Text) \tag{1}$$

3.2.2 Image encoding for smartphone imagery

Regarding the important visual element of diagnostics, the model applies the CLIP-ViT framework. More precisely, the specific ViT-L336px model is selected due to its unique high-resolution ability, which is essential to recognize the subtle visual features associated with certain diseases. Given a smartphone-captured image I, the local CLIP image encoder generates a corresponding embedding vector $e_I \in R^K$.

3.2.3 Edge device optimization

As the MM-RAG system targets smartphone implementation, it utilizes inference techniques that require less memory. Utilization of quantized BERT models, compressed CLIP embeddings, and the efficient use of FAISS indexes helps achieve this goal. The experiments were performed on smartphones with 4-8 GB RAM.

$$e_I = CLIP\_Image\_Encoder\,(Smartphone\ Image) \tag{2}$$

3.3 Multi-modal query generation and secure local retrieval

The encoded text and image embeddings are then integrated to form a comprehensive multi-modal query, which drives retrieval from the secure, on-device medical knowledge base.

3.3.1 Multi-modal query fusion

The encoded symptom text embeddings e_S and e_I are fused into a unified multi-modal query embedding q.

$$q = Dense\,(Concatenate\,(e_S, e_I)) \tag{3}$$

For handling missing and abnormal inputs, a modality-based preprocessing stage is presented. The missing text-based symptoms are filled via a masked token approach, while low-

quality images are detected via an image quality evaluation stage. In fusion, modality-based confidence weights are used:

$$q = Dense\ (w_s \cdot e_S + w_i \cdot e_I) \quad (4)$$

where, w_s and w_i denote the modality weights. In case of any absence of a particular modality, unimodal retrieval is performed.

3.3.2 FAISS-based secure local knowledge retrieval

The core of our retrieval mechanism is a FAISS based retrieval engine, which operates entirely and exclusively on the local device. The local medical knowledge base comprises a comprehensive, curated collection of medical documents and clinical guidelines. These knowledge assets are pre-embedded using a robust, domain-specific medical encoder and securely indexed within FAISS.

Given the multi modal query q, the local FAISS engine performs a rapid similarity search to identify the top-K most relevant medical knowledge snippets from the on-device knowledge base. Cosine similarity is employed as the primary metric

$$similarity(q, k_j) = \frac{q.k_j}{|q| \cdot |k_j|} \quad (5)$$

where, k_j is an embedding of a medical knowledge snippet within the local database. The top-K retrieved snippets, denoted as $K_{retrieved} = \{k_1, k_2, k_3, \ldots, k_K\}$ are then passed to the generation phase.

The local corpus size was around 50,000 medical papers embedded in 768 dimensions. The indexing technique used in this work is FAISS IVF-PQ with values nlist = 1024, m = 16, and nprobe = 32. The cosine similarity threshold was set to 0.78, while Top-K retrieval was chosen with parameter K = 10.

3.4 MM-RAG agent: Adaptive diagnostic orchestration

The retrieved contextualized knowledge forms the bedrock for the MM-RAG agent's core reasoning engine, which is powered by a novel adaptive diagnostic orchestration module.

3.4.1 Ontology-aware prompt engineering

To provide consistency in symptoms naming and classification of diseases, ontologies like UMLS and SNOMED-CT were used. Subsequently, the extracted information was represented as ontological entities before the construction of prompts. In this way, we could make sure that synonymous expressions like “skin rash” and “dermatitis” were treated as one entity. For the generation of reliable and clinically relevant diagnostics, we use an adaptive prompt engineering process based on the dynamically constructed ontology-aware prompt structure. The retrieved information snippets are not simply appended but are incorporated into the original user input and related ontological context to form a highly structured prompt P.

The structured prompt P for the Local LLM can be formulated as:

$$P = \text{Format (User Inputs, } K_{retrieved}\text{, Ontology Context, Instruction)} \quad (6)$$

The instruction part explicitly instructs the LLM to conduct the diagnostic reasoning, give the justification, and provide suggestions in a certain clinical framework.

3.4.2 Adaptive diagnostic orchestration with dynamic chain-of-thought

Right at the center of the intelligence of MM-RAG agents is the Adaptive diagnostic orchestration module, which uses the power of LLMs. The module coordinates an interactive, multi-step process of reasoning, called CoT, where the agent itself dynamically develops its diagnostic logic through interaction with retrieved medical information and specific patient details.

3.5 Secure output generation and user interface

The secure production of the result comes with thorough actions that the agent takes to guarantee privacy and understandability of the presented results

3.5.1 Local encryption module

To safeguard the privacy of the data used to diagnose, which is sent beyond the local device for purposes such as clinical analysis at a distance or to update models anonymously through federated learning techniques, all the information sent out from the device is securely encrypted within the device itself. The system uses strong encryption standards, including the Advanced Encryption Standard using 256-bit keys.

$$Encrypted\ Output = Encrypt\ (Diagnostic\ Output,\ Key) \quad (7)$$

3.5.2 User interface and follow-up

The end result of the diagnostic output will be decrypted if it had been encrypted before, and the output will then be sent to the user via an effective and efficient mobile app interface. It will not only provide the diagnosis but also an explanation of the reasoning behind it in terms of the CoT approach, a confidence measure to gauge the level of certainty of the model, and relevant recommendations such as ‘consult a dermatologist’, ‘observe symptoms and report any changes’, and ‘seek immediate medical help’.

3.6 Federated learning for privacy-preserving continuous improvement

In order to have continuous improvement for the MM-RAG agent, without compromising the privacy of the users, the use of Federated Learning is employed. Through the use of this decentralized system, the development of each of the model components, such as BERT text encoder, CLIP image encoder, and Local LLM, will be done.

The FL process follows five key steps:

1. Global Model Initialization: Participating edge devices receive a secure distribution of a pre-trained global model.

2. Local on-device instruction: Every device independently fine-tunes its local copy using on device patient interactions and anonymized ground truth, optimizing for diagnostic accuracy and explanation quality. The data is local only.

3. Encrypted Model Update Sharing: Devices send encrypted and anonymized model updates like weight gradients to a central FL server, ensuring data confidentiality.

4. Secure Aggregation: The FL Federated Averaging is used by the server to aggregate updates.

$$W^{(t+1)} = \sum_{k=1}^{N} \frac{n_k}{N} w_k^{(t+1)} \quad (8)$$

The raw data will continue to be restricted to the location where $W^{(t+1)}$. The updated global model $W^{(t+1)}$, represents the locally trained model from device k, where n_k denotes the number of training data points on device k. Methods to improve privacy that are not mandatory include secure multiparty computation and differential privacy.

5. Redistribution of Global Model: All the participating devices will receive a redistribution of the new model, allowing for further rounds of privacy-preserving training.

4. RESULTS

The empirical evaluation of the proposed MM-RAG agent in relation to home-based early screening of diseases. The performance of the agent was strictly evaluated through real-world, publicly available multimodal medical data sets that simulated realistic screening cases through smartphones. The architectural performance of the MM-RAG was evaluated against state-of-the-art language models, including GPT-4o and MEDPaL-2, mainly concentrating on accuracy, correctness, and explainability of the model structure. The results were proven to be more effective compared to the existing works.

4.1 Evaluation metrics

In order to have a clear and clinically driven assessment of the MM-RAG agent, three different measurements were used.

The first one is Top-1 accuracy for diagnosis, and it represents the ratio of cases where the model's prediction is exactly equal to the expert-verified diagnosis [26].

$$Accuracy = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{D}_i = D_{GT,i} \right) \times 100\% \quad (9)$$

Second, we introduce a Factual Coherence Score to assess the semantic and factual consistency of the model's justifications. Specifically, this score quantifies the alignment between the generated explanation and both the retrieved knowledge snippets and established ground-truth medical facts. To compute this, we use transformer-based similarity models such as BERTScore [27], along with entailment-based natural language inference (NLI) techniques [28], following established evaluation strategies in explainable text generation [29].

$$Coherence = FactualModel\ (J_{gen}, K_{retrieved}, GroundTruth) \times 100\% \quad (10)$$

Lastly, to capture the interpretability of the diagnostic reasoning process, a structural explainability score is used, which combines:

- The number of clearly defined reasoning steps (CoT step count).
- The degree of alignment between generated medical concepts and reference ontologies [30].
- The coverage of relevant retrieved facts within the explanation. This composite score is inspired by recent work on explainability alignment in medical AI [31] and is computed as a weighted sum of three components.

$$\text{Explainability Score} = \alpha \cdot StepCount + \beta \cdot OntologyAlignment + \gamma \cdot Coverage \quad (11)$$

4.2 Datasets and experimental design

To support the development and evaluation of the MM-RAG agent for home-based early disease screening, we selected only those datasets that align with realistic smartphone-based visual input scenarios. Specifically, the ISIC Archive was used for dermatological conditions, providing high-resolution skin images suitable for capturing observable surface-level abnormalities such as lesions, rashes, or pigmentation. For gynecological screening, we employed ethically sourced academic image datasets focused solely on externally visible symptoms, such as redness, inflammation, or lesions that could be captured using smartphone cameras.

To account for ocular conditions with visible external signs, such as conjunctivitis, eyelid inflammation, or scleral discoloration, we constructed a small-scale synthetic dataset using publicly available, medically verified facial and eye region images, simulating what a user could realistically capture at home.

Stratified sampling and data augmentation techniques have been applied to address the problem of class imbalance for the ISIC dataset. The ISIC dataset has been split into three sets, where 70 percent is training data, 15 percent is validation data, and another 15 percent is test data without any patient duplication. Finally, text-only symptom descriptions were curated to support over self-reported inputs such as cough, fatigue or fever. The data sets were split into 70%, 15%, and 15% of training, validation, and testing datasets. Patient-based separation was ensured so that there would be no overlap between the training and test samples. All datasets were preprocessed into multi-modal input pairs. Baseline models were evaluated on equivalent text-only representations to allow fair, modality-aligned comparison. Demographic distribution was performed for all ages, genders, and ethnic groups within the dataset. Fairness measures were introduced to balance the samples for mitigating biases. Future research will include fairness-aware federated learning and bigger multi-ethnic clinical datasets.

4.3 Comparative performance

The MM-RAG agent outperformed both GPT-4o and MedPaLM-2 across all evaluated metrics mentioned in Table 1, demonstrating the effectiveness of its multimodal design, retrieval grounding, and explainable reasoning.

Results in Table 2 show that encryption adds only a slight delay while significantly lowering the threat to privacy.

Table 1. Evaluation metrics

Metric	MM-RAG (Proposed)	GPT-4o	MedPaLM-2
Top-1 Accuracy (%)	87.2	74.5	78.9
Factual Coherence (%)	91.4	83.2	86.0
Explainability Score	82.6	59.1	68.4

Even though MedPaLM-2 and domain-specific variants of ChatGPT prove efficient in providing reasoning in terms of

their healthcare applications, most of these models leverage parametrized knowledge obtained during training and operate in conjunction with the cloud-based inference architecture. Conversely, MM-RAG uses an evidence-driven diagnostic pipeline where the output predictions are based on evidence extracted from the local medical corpus. Such a system significantly curbs the occurrence of hallucinations since reasoning can be confined only to the clinically valid evidence. Additionally, MM-RAG considers multimodal data in the form of the visual symptoms captured using smartphones, offering better diagnosis opportunities compared to text-based healthcare chatbots. The next significant difference is related to the protection of personal privacy. Most LLMs built specifically for healthcare applications need to transfer patient data to external cloud servers, where there are risks of potential vulnerabilities and threats to patients' privacy and health data. In order to mitigate such disadvantages, MM-RAG offers a solution through local knowledge base lookup, on-device inference, federated learning, and AES-256 encryption in communication. Therefore, all sensitive information stays on the device of the user at any stage, including diagnostics and machine training. Higher accuracy and reliability of results are shown in Table 3; MM-RAG proves the efficiency of this approach.

Table 2. Privacy evaluation

Metric	No Encryption	With Encryption
Data Leaks Risk	12.8%	0.9%
Network Transmission	Low	High
Diagnosis Accuracy	87.4	87.2
Latency (ms)	124	138

Table 3. Comparison with baseline models

Model	Accuracy
GPT 4	74.5
MedPaLM-2	78.9
Healthcare ChatGPT	81.7
MM-RAG	87.2

Table 4. Evaluation of edge devices

Device	RAM	Accuracy	Latency
High-End	12 GB	87.2%	0.9 s
Mid-Range	6 GB	86.8%	1.4 s
Low-End	4 GB	85.9%	2.1 s

The results shown in Table 4 MM-RAG system continued to function despite running on lower-performing devices, where it experienced only mild performance degradation.

Table 5. Evaluation of disease-wise sensitivity

Disease Category	Sensitivity
Dermatology	91.3
Ophthalmology	88.6
Gynaecology	84.7
Respiratory	82.5

The highest level of sensitivity was obtained in as per the results showed in Table 5, dermatological disorders due to visual clues whereas systemic conditions exhibited comparatively lower sensitivity.

The results from Table 6 indicate that FAISS had the highest retrieval speed while also achieving high accuracy during the retrieval process.

Table 6. Retrieval latency

Retrieval Technique	Latency (ms)
BM25	265
Dense Retrieval	184
FAISS	72

Table 7. Hardware compatibility and response time issues

Metric	MM-RAG
Response Time	Avg 1.6 sec
Latency Time to Retrieve	72 ms
Memory Requirements	3.8 GB
Storage Requirements	2.6 GB

The results in Table 7 indicate that the overall results are best in terms of response time and resource utilization.

5. CONCLUSION

Even with its impressive performance, MM-RAG is at present limited to only the diagnosis of diseases that have signs visible to the naked eye. Those diseases with internal manifestations needing further testing, such as laboratory procedures or imaging, cannot be done through this approach at present. This work introduces MM-RAG, a novel, privacy-preserving, multimodal Retrieval-Augmented Generation framework for real-time, home-based early disease screening. Through a rigorous evaluation on diverse, publicly available medical datasets across dermatology, ophthalmology, respiratory, and gynecological domains, MM-RAG has demonstrated superior performance over state-of-the-art baselines, including GPT-4o and MedPaLM-2.

The MM-RAG agent achieved a Top-1 diagnostic accuracy of 87.2%, a factual coherence score of 91.4%, and a structural explainability score of 82.6, outperforming the baselines across all key diagnostic dimensions. These gains are attributable to MM-RAG's unique combination of visual-semantic fusion (Via CLIP and BERT), grounded reasoning through local knowledge retrieval, and transparent multi-step diagnostic logic via CoT explanations. Importantly, MM-RAG differs from the current generation of medical LLMs in that it has been designed to operate locally, enabling federated learning and privacy-preserving knowledge access.

It is worth noting that this study is relevant within the context of increasing use of smartphones for imaging of health problems that can be assessed by visual means in the absence of any artificial intelligence technology. In particular, the camera of a smartphone makes possible the preliminary assessment of numerous diseases associated with their appearance (i.e., changes in the skin, eye, external gynecological manifestations, and some other signs of the disease that can be detected on the skin or face). While imaging alone does not allow for conducting a clinical examination or identify internal problems, it may be useful as the first step of assessment, remote consultation, or AI-assisted diagnosis, especially in underdeveloped regions.

Thus, based on the baseline potential, MM-RAG allows converting the visual inputs made via smartphone into clinically oriented diagnostics without violation of the user's privacy. The edge compatibility, grounded retrieval, and reasoning loop make a strong basis for further development of

the technology and its effective use in the provision of digital health services.

REFERENCES

[1] Bai, X., Wang, H., Ma, L., et al. (2021). Advancing COVID-19 diagnosis with privacy-preserving collaboration in artificial intelligence. Nature Machine Intelligence, 3(12): 1081-1089. https://doi.org/10.1038/s42256-021-00421-z

[2] Esteva, A., Chou, K., Yeung, S., et al. (2021). Deep learning-enabled medical computer vision. NPJ Digital Medicine, 4(1): 5. https://doi.org/10.1038/s41746-020-00376-2

[3] Berisha, V., Krantsevich, C., Hahn, P.R., et al. (2021). Digital medicine and the curse of dimensionality. NPJ Digital Medicine, 4(1): 153. https://doi.org/10.1038/s41746-021-00521-5

[4] Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J. (2022). Multimodal biomedical AI. Nature Medicine, 28(9): 1773-1784. https://doi.org/10.1038/s41591-022-01981-2

[5] Huang, Y., Huang, J.X. (2026). A survey on retrieval-augmented text generation for large language models. ACM Computing Surveys, 58(12): 1-38. https://doi.org/10.1145/3805774

[6] Ktena, I., Wiles, O., Albuquerque, I., et al. (2024). Generative models improve fairness of medical classifiers under distribution shifts. Nature Medicine, 30(4): 1166-1173. https://doi.org/10.1038/s41591-024-02838-6

[7] Chambon, P., Bluethgen, C., Delbrouck, J.B., et al. (2022). Roentgen: Vision-language foundation model for chest X-ray generation. arXiv preprint, arXiv:2211.12737. https://doi.org/10.48550/arXiv.2211.12737

[8] Carrillo-Perez, F., Pizurica, M., Zheng, Y., et al. (2025). Generation of synthetic whole-slide image tiles of tumours from RNA-sequencing data via cascaded diffusion models. Nature Biomedical Engineering, 9(3): 320-332. https://doi.org/10.1038/s41551-024-01193-8

[9] Sallam, M., Barakat, M., Sallam, M. (2024). A preliminary checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence–based models in health care education and practice: Development study involving a literature review. Interactive Journal of Medical Research, 13(1): e54704. https://doi.org/10.2196/54704

[10] Eysenbach, G. (2023). The role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. JMIR Medical Education, 9(1): e46885. https://doi.org/10.2196/46885

[11] Schouten, D., Nicoletti, G., Dille, B., et al. (2025). Navigating the landscape of multimodal AI in medicine: A scoping review on technical challenges and clinical applications. Medical Image Analysis, 105: 103621. https://doi.org/10.1016/j.media.2025.103621

[12] Abdar, M., Pourpanah, F., Hussain, S., et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76: 243-297. https://doi.org/10.1016/j.inffus.2021.05.008

[13] Udegbe, F.C., Ebulue, O.R., Ebulue, C.C., Ekesiobi, C.S. (2024). The role of artificial intelligence in healthcare: A systematic review of applications and challenges. International Medical Science Research Journal, 4(4): 500-508. https://doi.org/10.51594/imsrj.v4i4.1052

[14] Li, H., Huang, J., Ji, M., Yang, Y., An, R. (2025). Use of retrieval-augmented large language model for COVID-19 fact-checking: Development and usability study. Journal of Medical Internet Research, 27: e66098. https://doi.org/10.2196/66098

[15] Aiumtrakul, N., Thongprayoon, C., Suppadungsuk, S., et al. (2023). Navigating the landscape of personalized medicine: The relevance of ChatGPT, BingChat, and Bard AI in nephrology literature searches. Journal of Personalized Medicine, 13(10): 1457. https://doi.org/10.3390/jpm13101457

[16] Koçak, B., Keleş, A., Köse, F. (2024). Meta-research on reporting guidelines for artificial intelligence: Are authors and reviewers encouraged enough in radiology, nuclear medicine, and medical imaging journals? Diagnostic and Interventional Radiology, 30(5): 291-298. https://doi.org/10.4274/dir.2024.232604

[17] McCaffrey, P., Jackups, R., Seheult, J., et al. (2025). Evaluating use of generative artificial intelligence in clinical pathology practice: Opportunities and the way forward. Archives of Pathology & Laboratory Medicine, 149(2): 130-141. https://doi.org/10.5858/arpa.2024-0208-RA

[18] Yan, K., Li, T., Marques, J.A.L., Gao, J., Fong, S.J. (2023). A review on multimodal machine learning in medical diagnostics. Mathematical Biosciences and Engineering, 20(5): 8708-8726. https://doi.org/10.3934/mbe.2023382

[19] Duong, Q.A., Tran, S.D., Gahm, J.K. (2025). Multimodal surface-based transformer model for early diagnosis of Alzheimer's disease. Scientific Reports, 15(1): 5787. https://doi.org/10.1038/s41598-025-90115-y

[20] Wang, D., Liu, W., Gao, L., Qu, Y.N., Zhang, H., Shi, J. (2024). Modal-centric insights into multimodal federated learning for smart healthcare: A survey. In International Conference on Algorithms and Architectures for Parallel Processing, pp. 145-160. https://doi.org/10.1007/978-981-96-1548-3_10

[21] Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P. (2020). Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. NPJ Digital Medicine, 3(1): 136. https://doi.org/10.1038/s41746-020-00341-z

[22] Khakpaki, A., Aghasi Javid, S., Farshbaf-Khalili, A., Sepehri, H. (2026). AI-integrated clinical decision support systems for precision medicine in real-world healthcare. Discover Artificial Intelligence, 6: 752. https://doi.org/10.1007/s44163-026-01518-3

[23] Sagheer, S.V.M., Meghana, K.H., Ameer, P.M., Parayangat, M., Abbas, M. (2025). Transformers for multi-modal image analysis in healthcare. Computers, Materials and Continua, 84(3): 4259-4297. https://doi.org/10.32604/cmc.2025.063726

[24] Kumar, D., Choudhary, N., Gupta, V., et al. (2026). Natural language processing in healthcare: From unstructured data to clinical intelligence. Intelligent Systems with Applications, 31: 200698. https://doi.org/10.1016/j.iswa.2026.200698

[25] Lee, J., Choi, B. (2020). Artificial intelligence in healthcare: Current applications and future directions. Healthcapes Journal of Artificial Intelligence, 12(3): 251-269.
[26] Esteva, A., Kuprel, B., Novoa, R.A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639): 115-118. https://doi.org/10.1038/nature21056
[27] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. arXiv preprint, arXiv:1904.09675. https://doi.org/10.48550/arXiv.1904.09675
[28] Nie, Y., Williams, A., Dinan, E., et al. (2020). Adversarial NLI: A new benchmark for natural language understanding. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 4885-4901. https://doi.org/10.18653/v1/2020.acl-main.441
[29] Lin, C.Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pp. 74-81.
[30] Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. Nucleic Acids Research, 32(suppl_1): D267-D270. https://doi.org/10.1093/nar/gkh061
[31] Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35: 24824-24837.
[32] Abo El-Enen, M., Saad, S., Nazmy, T. (2025). A survey on retrieval-augmentation generation (RAG) models for healthcare applications. Neural Computing and Applications, 37(33): 28191-28267. https://doi.org/10.1007/s00521-025-11666-9
[33] Karim, A.H.M., Uzuner, O. (2025). Multimodal retrieval-augmented generation with large language models for medical VQA. arXiv preprint, arXiv:2510.13856. https://doi.org/10.48550/arXiv.2510.13856
[34] Zhao, X., Liu, S., Yang, S.Y., Miao, C. (2025). MedRAG: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In Proceedings of the ACM on Web Conference 2025, Sydney, NSW, Australia, pp. 4442-4457.