World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

A Multi-Modal AI Approach for Dermatological Diagnosis

S. SUREKHA and VENNELA VILLA *

Department of Computer Science and Engineering, UCEK(A), JNTU Kakinada, Andhra Pradesh, India-533003.

World Journal of Advanced Research and Reviews, 2026, 31(02), 095–104

Publication history: Received on 22 June 2026; revised on 30 July 2026; accepted on 01 August 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.31.2.2013

Abstract

Skin is essential for maintaining it in good condition; identifying skin conditions through lesion images is insufficient; we require both visual and medical information. Although artificial intelligence has improved in recent days at predicting skin conditions from lesion images, most of the current systems are absent of medical content and instead focus on visual analysis. This research addresses the gap by building a diagnostic support framework that incorporates lesion image analysis with relevant medical knowledge. At the core of the system is a vision transformer (ViT), which analyses lesion images to predict possible skin diseases. From there, the system retrieves relevant medical information using RAG to support a more reliable diagnosis. It then goes a step further, generating clear diagnostic explanations, treatment suggestions and general healthcare guidance that the user can actually act on. The system is designed to be both trustworthy and practical-genuinely useful for real diagnostic decision-making

Keywords: Vision Transformer (ViT); Retrieval-Augmented Generation (RAG); LLaMA 3.2; ChromaDB; Dermatological Diagnosis

1. Introduction

Skin diseases are the most common health issues worldwide. Maintaining skin health is important for overall well-being. Conditions like psoriasis, acne, eczema, and skin cancer all require early diagnosis since delayed treatment can affect the whole body. The problem is that getting an accurate dermatological diagnosis usually depends on an experienced specialist and careful medical analysis resources that are not available in every healthcare centre [1]. When the diagnosis is delayed, or treatment goes wrong, the skin conditions can worsen, and patient outcomes can be negatively affected [2]. That's exactly why improving reliable skin disease diagnosis has become a critical priority.

Nowadays, AI has significantly advanced this field, particularly in medical image analysis and disease prediction. Models trained on lesion images can identify possible skin diseases with real effectiveness [3],[4], cutting down the manual work for healthcare professionals and speeding up the diagnostic process.

Even so, most existing systems identify the diseases but stop short of offering further explanation, treatment guidance, or recommendations to the patient. It is difficult to depend only on the predicted skin diseases in clinical decisions; we need something that gives more useful clinical insight. This gap is what motivates the need for a smarter dermatological framework; one that not only predicts the disease but also brings in contextual medical knowledge and practical healthcare guidance.

This paper proposes exactly that: a multi-model AI framework that uses the vision transformer (ViT) to classify different skin diseases, then draws on Retrieval-Augmented Generation (RAG) to retrieve the relevant medical knowledge. Beyond classification, the framework generates meaningful explanations and personalised recommendations, making the overall approach in a more comprehensive way.

* Corresponding author: VENNELA VILLA; Email: villavennela@gmail.com

This describes how the balance paper is arranged. The research on deep learning, AI, and RAG methods for dermatological diagnostics is reviewed in Section II. Section III proposes the methodology of the architecture phases: data collection, image preprocessing, disease prediction by vision transformer, medical information retrieval and report generation. Section IV presents the results and performance of all the models that are used in this framework. Section V summarises the research of this paper, which we have discussed for intelligent dermatological diagnosis and decision support.

2. Related work

Dermatological disorders are among the most common health issues worldwide, and getting the diagnosis right really matters — both for effective treatment and for avoiding complications down the line. Traditional computer-aided diagnosis approaches have leaned heavily on machine learning and deep learning to categorise skin conditions from pictures. Esteva et al. [2] showed early how deep learning could classify different skin cancers at a level comparable to dermatologists. Around the same time, Codella et al. [3] and Tschandl et al. [4] contributed benchmark datasets and evaluation frameworks that pushed automated skin lesion analysis forward considerably. This framework improved disease detection clearly, but they have some common limitations: it depends entirely on vision transformers, with some contextual understanding or explainability behind the predictions.“From there, transformer-based models look forward especially when they are needed for better feature extraction and building richer image representation."Take Liu et al. [5] — they built the Swin Transformer to make image understanding more efficient. Caron et al. [6] went a different route with DINOv2, using self-supervised learning to build visual representations without heavy labelling. Chen et al. [7], meanwhile, combined transformers with convolutional methods in TransUNet, mainly to get better medical image segmentation. All three moved the field forward, no question.

But here's the catch: most of these techniques still stop at image classification. They don't touch patient symptoms, medical history, or the kind of domain-specific knowledge a real diagnosis actually needs.That's the gap more recent research has tried to close, largely by pulling in large language models and outside knowledge sources to add context and generate real explanations. Lewis et al. [8] were among the first here, introducing Retrieval-Augmented Generation (RAG) — essentially pairing language generation with information retrieval so responses stay grounded rather than made up. Around the same period, [9] and Devlin et al. [10] showed just how far contextual language models could go in handling medical information and producing usable explanations.

This is really the foundation this work builds on: a Multi-Modal AI Framework for Dermatological Diagnosis, bringing together Vision Transformer-based image analysis, retrieval-augmented generation, and language models — not just to predict, but to explain and recommend treatment too. Where older image-only systems stop, this approach keeps going, pairing visual analysis with real medical context to make both the diagnosis and the decision-making behind it more dependable.

3. Methodology

This methodology presents the overall design and working of the proposed multi-model dermatological diagnosis and decision support framework, including the complete system architecture. It covers all the operational phases involved from image preprocessing and skin disease prediction, medical information retrieval and personalised report generation. It also outlines the performance metrics used to assess how well the disease classification model and the medical information retrieval process actually perform.

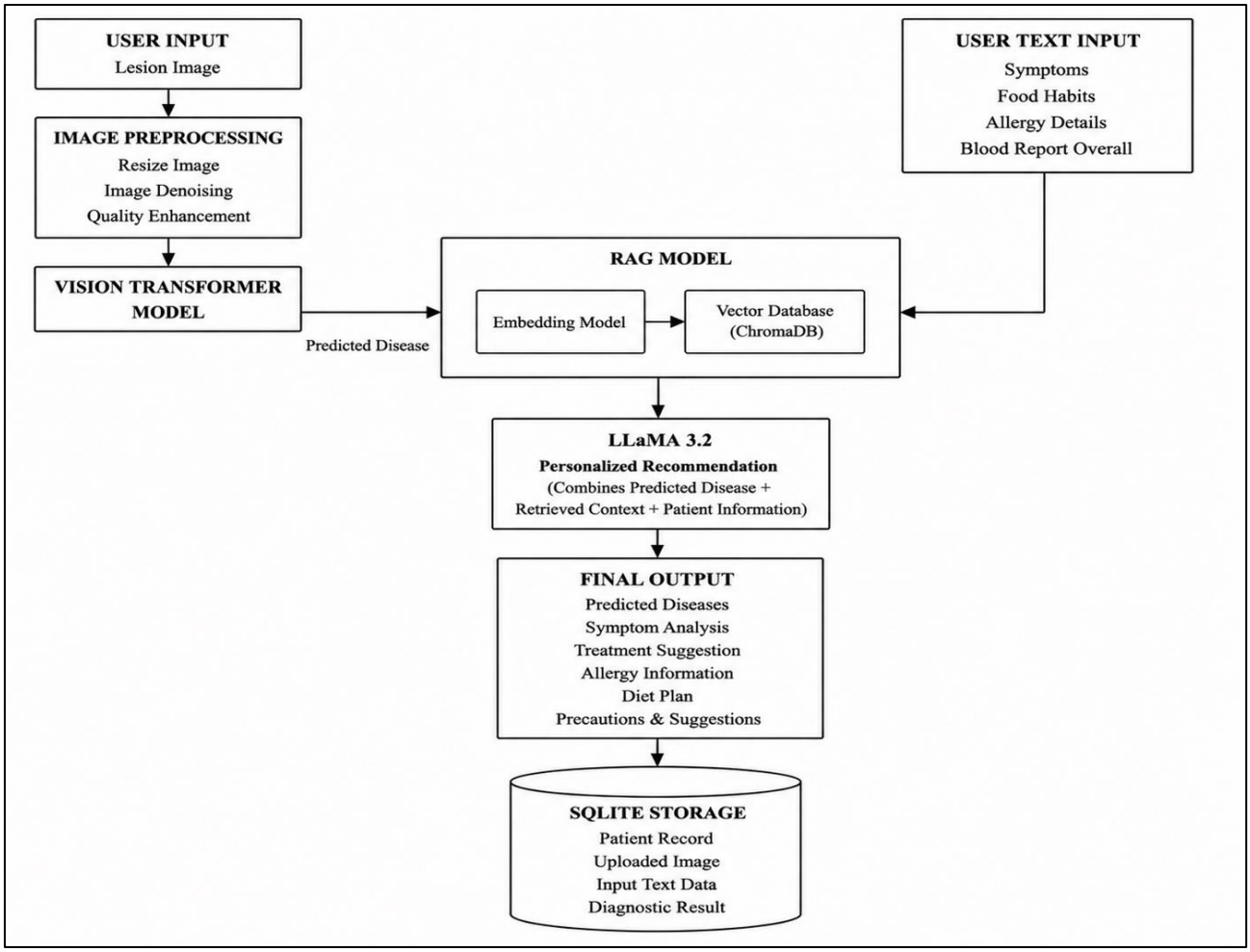

Figure 1 The proposed dermatological diagnosis system's architecture

3.1. PHASE 1: Input Data Collection and Image Preprocessing

The process starts with collecting two types of information: the lesion image and patient information, including symptoms, dietary habits, detailed allergies, and blood report summaries. So here is the lesion image itself can come from a phone camera, a digital camera or a dermatoscope, and gets uploaded through the interface in common formats like JPG, JPEG, or PNG. Because of this variety in sources, image quality naturally varies too: resolution, brightness, contrast, which is why the image needs to be preprocessed before analysis

The initial step in preprocessing is converting the lesion image into RGB format so colour representation stays consistent across all uploads. Since images come in all sorts of original dimensions, each one was resized to 224 ×224 ×3 pixels according to the input size required by vision transformers. This resizing is done by using Bicubic interpolation, which calculates new pixel values based on neighbouring pixels, helping preserve important lesion details like texture, colour variation, and border definition rather than blurring them away. After that, the image is then normalised using the standard ImageNet normalisation process to provide a consistent numerical representation.

By the end of this preprocessing, the original uploaded image has been transformed into a standardised, high-quality RGB image sized 224 ×224 ×3 pixels. This cleaned-up lesion image then goes to the vision transformer for feature extraction and disease prediction.

3.2. PHASE 2: Extraction of Features and Disease Prediction

The vision transformer (ViT) model uses the analyzed lesion image as input to extract features and predict or identify disease after image preprocessing. To identify different skin conditions, the model learns to see the texture of a lesion, the color of the lesion, its shape, its border, etc.

The vision transformer's self-attention mechanisms model interactions across multiple areas of lesion images, unlike traditional image classification. It takes visual patterns from local and global areas, which gives better classification skills and feature representation.

The model chooses with confidence based on which disease category it has the highest confidence. Then it labels the disease predicted for the patient with its data. Now, predicted diseases are thrown into the RAG for medical information retrieval.

3.3. PHASE 3: Contextual Medical Information Retrieval

By using a Retrieval-Augmented Generation (RAG) framework, the system pulls the relevant medical facts before it generates a diagnosis. Rather than depending purely on what the language models already know, it searches an external medical database for information that is helpful in identifying the skin condition. This one step matters a lot to add real clinical grounding and sharpen the quality of the matter. It cuts down the risk of the system offering incorrect and unverified content.

The retrieval process uses an embedding model and turns medical documents into numerical vectors. These embeddings capture the actual meaning of the text, allowing the system to connect different medical concepts instead of just matching exact keywords.

Those embeddings are stored in the ChromoDB, which essentially stores the searchable medical data. When a query comes into ChromoDB, it compares these numerical representations to run a semantic search to pull out the documents that genuinely match the context of the query rather than the wording.

Once the vision transformer predicts the skin diseases, that prediction gets bundled together with the patient details – symptoms, diet, allergies, blood report summaries to form a query embedding. From there, the system uses cosine similarity to measure how closely the given query matches the stored document, and anything scoring below 0.85 is filtered out, keeping only the genuinely relevant information.

The resulting medical document contains the disease characteristics, possible cause, treatment methods, dietary recommendations, allergy precautions, and health care measures.

Finally, this medical information goes to the LLaMA 3.2 model, which combines the patient's clinical data with the retrieved medical context to generate a comprehensive, context-aware decision support for the diagnosis report.

3.4. PHASE 4: Diagnostic Report Generation Using LLaMA 3.2

Once the relevant documents are retrieved, the framework runs the LLaMA3.2 through the Ollama platform, which gives us a full diagnostic report. In this takes the both predicted disease and the patient’s symptoms, diet, allergies, blood test summaries, and the relevant medical context. This model evaluates the patient’s condition from a clinical perspective.

While generating the report together, the model combines the predicted diseases along with retrieved medical facts, making sure everything links up with what is actually in the patient records. This allows the final report to deliver highly specific, actionable guidance rather than just generic medical information.

The report outlines the diagnosed condition along with educational treatment information, dietary advice, allergy precautions, and preventive care tips. All the medical data in the database was verified and all of it was relevant to the projected skin condition and the patient’s unique health profile.

The final report is in a structured format that is displayed on the user interface; it makes the diagnosis easy to read and understand. By combining predicted disease with relevant medical information, the final phase generates a diagnostic workflow and delivers a meaningful clinical medical decision.

3.5. PHASE 5: Patient Record Management Using SQLite

Once the final report is generated, the framework saves all the patient's details and diagnostic results into a local SQLite database for future reference. The patient's profile, predicted disease and clinical inputs, final report and exact timestamps are stored in this database.

SQLite provides a lightweight and stronger setup to store and retrieve patients' records quickly without needing a separate server. Because each diagnostic record follows a structured layout and provides past reports very quickly and reliably.

Maintaining the history of all diagnostic records isn't just for storage purpose also it supports the ongoing monitoring and compares the current findings with the past ones. This is genuinely useful for healthcare professionals to track the condition of patients. The framework manages the records efficiently and makes them accessible whenever they're needed again.

3.6. PHASE 6: Performance Evaluation Metrics

The proposed framework evaluates both classification and retrieval performance separately; it's made up of two distinct components: a vision transformer for disease prediction and an RAG pipeline for medical content retrieval. These metrics provide an assessment of the framework's predictive accuracy, retrieval quality and overall reliability.

3.6.1. Vision Transformer Evaluation Metrics

In disease classification, four standard metrics- Accuracy, Precision, Recall, and F1-Score were used to analyse the performance of the vision transformer. These values demonstrate how well we can predict whether we're right or wrong, if we're too optimistic or too pessimistic, and if we're right about this one but wrong about that one.

1) Accuracy

Accuracy figuring out percentage of all samples that were correctly predicted to have correct skin diseases. It represents that the vision transformer has better prediction from the overall classification performance.

$$\text{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \tag{1}$$

2) Precision

It looks at how accurately the model predicted positively. These are actually essential when predicted cases for a given class truly belong to it. It tells us how reliable the model classification is and how often raising the false alarms.

$$\text{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \tag{2}$$

3) Recall

It tells how well the model pulls the actual disease cases that are really these. This is most important because it reflects missing the diagnosis that it doesn't have.

$$\text{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{3}$$

4) F1-Score

It brings precision and recall together into a balanced single number, which is mainly useful when the skin condition shows more than the others. Without it, a model could look like a stronger metric than others.

$$\boldsymbol{F1} = \frac{\mathbf{2}\times(\text{Precision}\times\text{Recall})}{\text{Precision}+\text{Recall}} \tag{4}$$

3.6.2. RAG Retrieval Evaluation Metrics

The RAG performs on three standard metrics: Recall, NDCG@K and MRR@K. Together, these tell us whether we have taken the right document or not; they were ranked within the results.

1) Recall@K

Recall checks whether the relevant medical document is the top K result or not. The higher value are more successful modules captures text that is closer to the top of the list.

$$\boldsymbol{Recall@K} = \frac{|\boldsymbol{Relevant} \cap \boldsymbol{Retrieved@K}|}{|\boldsymbol{Relevant}|} \quad (5)$$

2) NDCG@K

Normalised Discounted Cumulative Gain (NDCG@K) classifies the quality of the document by assigning greater weight to the relevant text that appears first in the search result.

$$\boldsymbol{NDCG@K} = \frac{\boldsymbol{DCG@K}}{\boldsymbol{IDCG@K}} \quad (6)$$

3) MRR@K

Mean Reciprocal Rank (MRR@K) measures how rapidly the relevant document appears first in the search result. A higher MRR value indicates that it successfully gets the medical information at the top of the list.

$$\boldsymbol{MRR@K} = \frac{1}{|U|} \sum \frac{1}{\boldsymbol{rank_i}} \quad (7)$$

4. Results and analysis

The proposed system is a multi-modal AI architecture that uses the HAM10000 dataset, which contains 10,015 records of dermatoscope skin images along with clinical metadata. Here, we preprocessed the lesion images by using bicubic interpolation and verified the quality using the SSIM score. The vision transformer (ViT) evaluates the classification metrics: Accuracy, Precision, Recall, and F1-Score. For the RAG medical data retrieval, it uses the Recall, Normalised Discounted Cumulative Gain (NDCG), and Mean Reciprocal Rank (MRR) to assess search effectiveness and ranking quality. Finally, the system demonstrates how well it supports the diagnosis system.

4.1 Image Preprocessing

To choose the best resizing method, here, we preprocessed the lesion image to improve its quality using the structural similarity index (SSIM). SSIM measures how well the image matches the original by analysing its structural data, contrast and brightness. SSIM range from 0 to 1, with values nearer to 1 indicating better image quality and maximum retention of structural features.

In this, we compared three image resizing techniques to evaluate the structural preservation: Nearest Neighbour, Bilinear, and Bicubic interpolation. Nearest Neighbour achieved an SSIM of 0.82, but there was a loss of fine resolution due to pixelation. Bilinear interpolation achieved a higher SSIM score of 0.89, which was better than the previous approach; it produced smoother images, but blurred the fine lesion features significantly. Bicubic interpolation reached the highest SSIM score of 0.96, defeating both approaches.

This confirms that bicubic preserves the best critical lesion boundaries, texture patterns, and color variations. As shown in Figure 2, bicubic interpolation delivers better image quality and preserves the important features more effectively than the other two methods. This technique resizes the image to 224 ×224 ×3 pixels, which is useful for the next phase. After resizing the image by using bicubic interpolation based on the SSIM score,

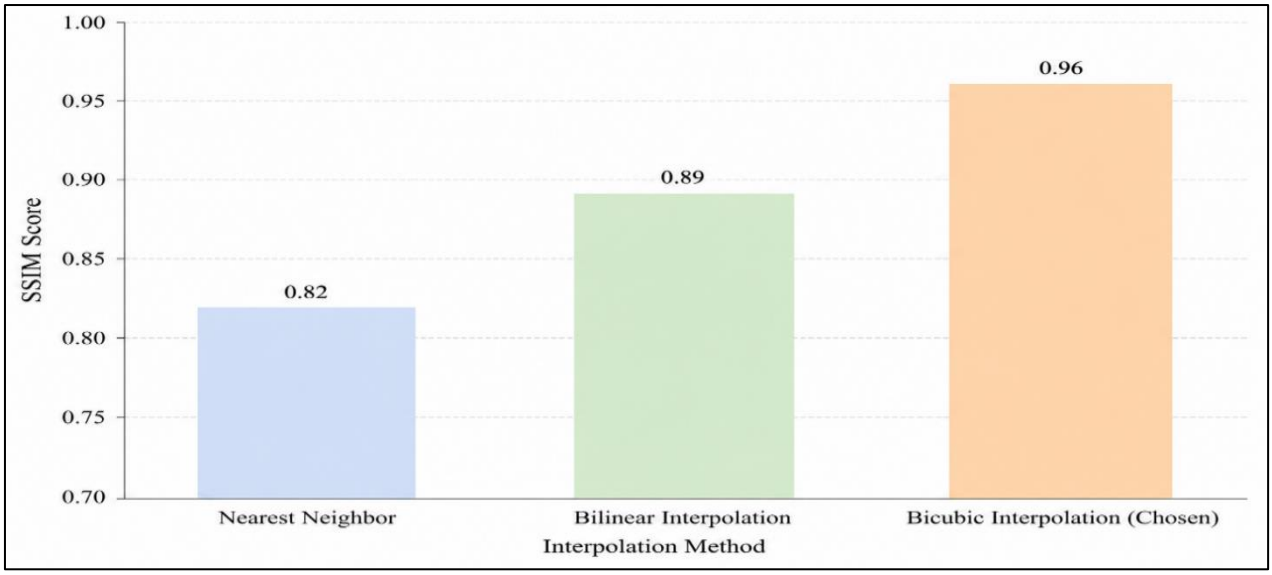

Figure 2 Image Quality Optimisation using SSIM Score

we forward this preprocessed lesion image to the vision transformer. There, we use the four standard metrics: Accuracy, Precision, Recall, and F1-Score to analyse the model's classification performance

4.2 Vision Transformer (ViT) Classification Performance Metrics

The Accuracy, Precision, Recall, and F1-Score are the four metrics used in the classification process to classify the Vision Transformers. Each metric scores between 0 to 1; metrics nearer to 1 have higher model predictions. Overall Classification Accuracy of the model was 93.5%, which shows that most of the evaluated lesion images were correctly predicted as the skin condition.

Precision of the model was 92.1%, which indicates a high confidence rate and rarely gives false alarms when it states that a lesion belongs to a specific disease class.

The model recall score was 94.2%, indicating a validated potential to reduce the frequency of missed diagnoses and detect accurate skin disease cases. This is critical in dermatological environments as neglected problems could jeopardise subsequent medical attention. Finally, the F1-Score was 93.1%, which

confirms the model’s capacity to maintain the ideal ratio of recall to precision, thus ensuring consistent performance during all phases.

As seen in Figure 3, the Vision Transformer achieves very accurate skin lesion classification, which forms a solid basis for the remainder of the framework.

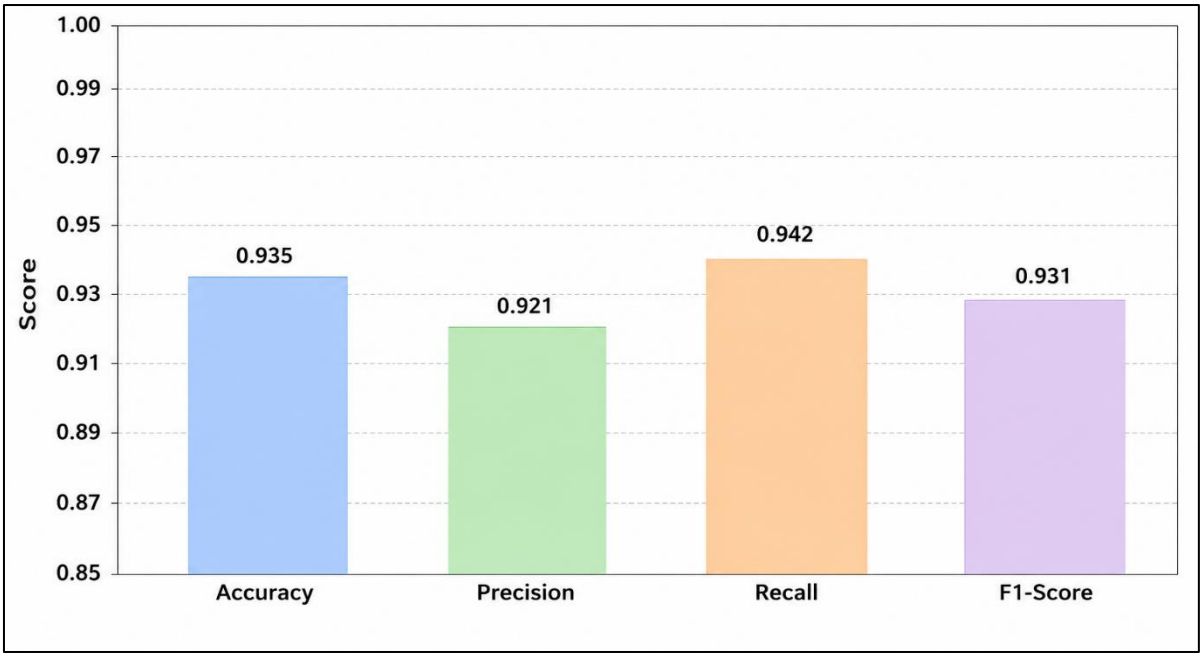

Figure 3 Vision Transformer (ViT) Classification Performance Metrics

The system sends the patient's data and the disease prediction from the Vision Transformer to RAG, so that the medical content can be retrieved. Then it measures using three standard metrics: Recall@K, NDCG@K, and MRR@K.

Table 1 Performance Evaluation of RAG Retrieval at K = 1, 3, 5

K	Recall@K	NDCG@K	MRR@K
1	0.8240	0.8240	0.8240
3	0.9410	0.8950	0.8820
5	0.9880	0.9120	0.8910

Here we analysed the RAG pipeline using Recall@K, NDCG@K, and MRR@K, where k represents the number of document segments retrieved from the ChromoDB for a given query. Increasing the value of k provides access to a broader pool of medical knowledge, thereby enhancing the quality of the diagnostic report.

Table 1 represents the retrieval metrics obtained at k = 1, 3, and 5. When we extract only one document chunk (k=1), the framework achieves identical Recall@K, NDCG@K, and MRR@K scores of 0.824, confirming that the system successfully recovers relevant medical content on the first attempt.

Now, these metrics improve as the number of retrieved document chunks increases to k=3; the framework achieved a Recall@K of 0.941, an NDCG@K of 0.895, and an MRR@K of 0.882. This increasing trend demonstrates that expanding the search capacity provides sufficient medical context for highly effective document ranking.

At k=5, the performance peak was achieved; Recall@K, NDCG@K, and MRR@K reached 0.988, 0.912, and 0.891. However, extracting more document chunks and broader context coverage, it also delays processing and retrieval of medical data, introducing peripheral data that adds little value to the final diagnosis.

Balancing retrieval quality and efficiency was needed, so we selected k=3 as the optimal configuration for the framework. The sample medical data for report generation without introducing unnecessary processing was taken from the threshold similarity, as classified by the comparison performance in Figure 4.

This overall performance proves that the RAG pipeline retrieves the medical data documents that have high priority and provides a reliable basis for generating informative diagnostic reports.

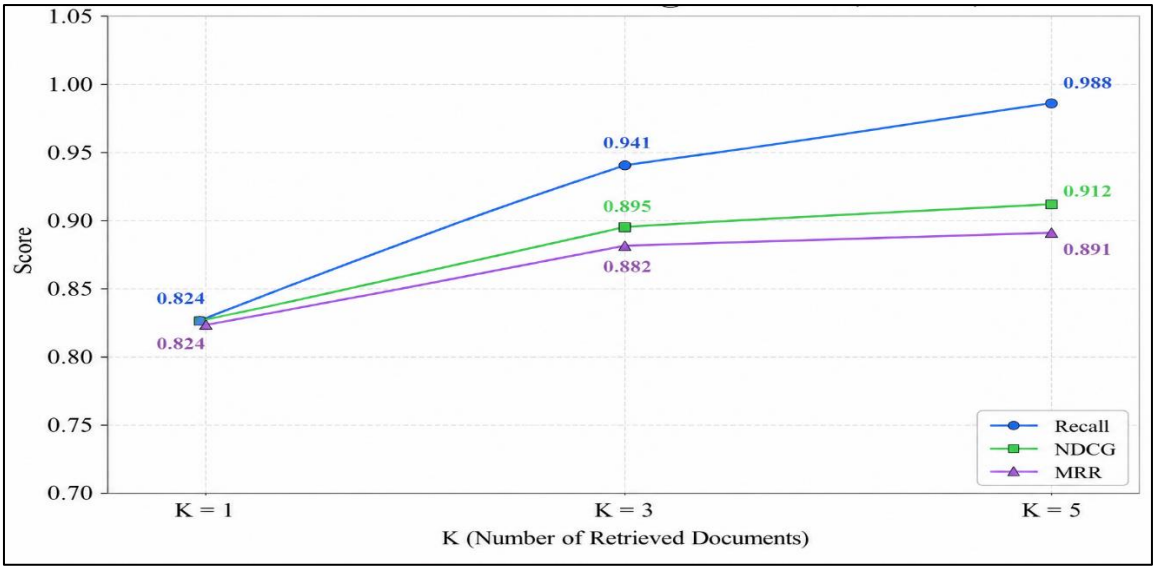

Figure 4 Performance of RAG Retrieval Metrics (Recall, NDCG, MRR)

This experiment confirms that all the components of the proposed framework worked effectively. Image preprocessing captures the important lesion features for reliable disease prediction. Concurrently, the classification model delivers strong predictive accuracy while the RAG pipeline retrieves highly relevant medical data for report generation. Together, these outcomes validate the overall efficacy of the proposed multimodal dermatological diagnosis framework.

5 Conclusion

In this research, we have presented a multimodal approach used in dermatological diagnosis and decision support. In this case, we have used the Vision Transformer (ViT) with Retrieval-Augmented Generation (RAG), a system that works in a two-stage process. First, the lesion image is preprocessed: bicubic interpolation, achieving a maximum SSIM score of 0.96, ensuring that lesion information is preserved for feature extraction. The image is passed to (ViT) for disease prediction, achieving classification accuracy of 93.5%, a precision of 92.1%, a recall of 94.2% and an F1-score of 0.931 when it comes to various skin diseases prediction. After that, RAG is used to retrieve relevant clinical data efficiently, achieving a Recall@5 of 0.988, NDCG@5 of 0.912 and MRR@5 of 0.891. Using this retrieved information, we can generate a personalised recommendation using LLaMA 3.2. This report is helpful from a clinical perspective and provides healthcare with a better understanding of the diseases and their treatment. The presented methodology proves that the combination of language models enhanced with medical knowledge retrieval and deep learning algorithms greatly benefits diagnosis.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] IEEE Access, vol. 13, pp. 201246–201256, 2025; E. A. Olca, "Professor X: Diagnosis and Treatment of Dermatological Diseases by Integration of Visual Diagnosis and Retrieval-Augmented Generation (RAG) Technologies."

[2] "Dermatologist-level Classification of Skin Cancer with Deep Neural Networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017; A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun.

[3] N. Codella *et al.*, "Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2018 International Symposium on Biomedical Imaging (ISBI)," in *Proc. IEEE Int. Symp. Biomedical Imaging (ISBI)*, 2018, pp. 168–172.

[4] "The HAM10000 Dataset: A Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions," Scientific Data, vol. 5, no. 1, Art. no. 180161, 2018, P. Tschandl, C. Rosendahl, and H. Kittler.

[5] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022.

[6] "DINOv2: Learning Robust Visual Features without Supervision," arXiv preprint, arXiv:2304.07193, 2023, M. Caron et al.

[7] "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv preprint, arXiv:2102.04306, 2021, J. Chen et al.

[8] "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," P. Lewis et al., Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.

[9] H. Touvron, L. Martin, K. Stone, et al., "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint*, arXiv:2302.13971, 2023.

[10] "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," by J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.