DermAssist: An AI-Powered Dermatology Support System

Mehmet Akif Özgür, Bengü Barış Balkan, Furkan Özyurt

June 18, 2025

1 Abstract of the Project

Dermatological diseases represent a significant global health challenge that requires timely and accurate diagnostic tools. Due to shortages in dermatological specialists and difficulties accessing immediate medical consultations, automated diagnostic systems have become increasingly necessary. Although current image classification methods achieve high accuracy, they often lack transparency, interpretability, and user interaction capabilities. Additionally, general-purpose Large Language Models (LLMs) alone face challenges such as generating incorrect medical information (hallucinations) and an inability to integrate effectively with visual diagnostic input.

This project addresses these limitations by developing an interactive web-based dermatological assistant that integrates image classification, Retrieval-Augmented Generation (RAG), and LLM methodologies. Users upload an image and a free text query through the web interface. The uploaded image is first processed by a Vision Transformer (ViT-Base) model pre-trained on the DermNet data set and subsequently fine-tuned on our custom 30-class dermatology dataset, achieving approximately 91% accuracy.

After classification, the system retrieves relevant contextual information using the RAG module. The user's original question is reformulated into formal medical language by the selected LLM model. This refined question is then embedded using `all-MiniLM-L6-v2` sentence embeddings and compared against a set of 50 question-answer (QA) pairs per disease class, generated using GPT-o3. The retrieval step involves computing a hybrid similarity score, combining both question-question and question-answer semantic similarities to select the most relevant QA pairs.

In the final stage, both the original and refined user questions, along with the selected relevant QA pairs, are fed back into the same LLM, which generates a medically accurate, empathetic, and user-friendly response. This response is then displayed interactively on the website's chat interface.

Preliminary evaluations demonstrate that the fine-tuned ViT-Base classifier performs robustly in dermatological image classification. Early qualitative assessments also indicate significant improvements in the contextual relevance and accuracy of responses generated by integrating the RAG module. The final LLM model to be used in this project is selected by the LLM-as-a-judge evaluation methodology. Based on these assessments, the `MedGemma-8b Q8` model has shown the best performance.

The anticipated impact of this project is to provide a scalable, reliable, and interpretable preliminary dermatological consultation tool. Particularly in regions with limited access to dermatological specialists, this tool aims to reduce diagnostic delays and assist patients in making more informed decisions regarding their dermatological health.

Future directions include incorporating multi-label classification support, extending multi-turn dialogue capabilities, integrating larger and more diverse dermatological datasets, and performing comprehensive clinical validation studies to assess real-world effectiveness.

2 Introduction, Problem Definition & Literature Review

Introduction

Dermatological diseases represent a major problem for global healthcare systems and require timely and accurate diagnostic methods. Due to a shortage of specialized dermatologists and barriers to immediate medical consultation, the development of intelligent and automated diagnostic tools has become increasingly essential. These systems are expected not only to classify skin conditions from images with high accuracy but also to explain the results in medically valid and user-friendly ways, thereby improving trust and usability.

The primary problem addressed in this project is the lack of interpretability, interaction, and medical reliability in existing dermatological image classification systems and general-purpose language models. Traditional deep learning-based classifiers typically function as black boxes and offer no reasoning for their outputs. Meanwhile, Large Language Models (LLMs), although successful in various natural language processing (NLP) tasks, often generate hallucinated medical content and lack direct integration with visual inputs, limiting their utility in clinical or consumer-facing scenarios.

Literature Review

Several studies have demonstrated promising results in skin condition classification using deep learning. Esteva et al. [EKN+17] showed that convolutional neural networks (CNNs) could achieve dermatologist-level performance in skin cancer detection. Han et al. [HKL+18] extended this by developing CNN-based classifiers for a broader range of dermatological diseases, while Javid [Jav22] and Kumari [Kum24] contributed curated datasets to support further model development. However, these CNN-based systems typically lack transparency and fail to provide contextual explanations to end users.

With the advent of transformer-based architectures, particularly Vision Transformers (ViTs) introduced by Dosovitskiy et al. [DBK+21], deep learning models began to exploit self-attention mechanisms to capture long-range dependencies in images. Subsequent works [LHZ+23], [MHW+23], [GCC+24] have demonstrated the utility of ViT and multimodal extensions in dermatology and other clinical imaging tasks. Despite their superior accuracy, these models still suffer from poor interpretability and limited user interaction.

In the field of natural language generation, domain-specific LLMs such as BioGPT [LSX+22], PubMedBERT [GTC+21], and Med-PaLM [SAT+22] have shown improved performance over general-purpose language models in biomedical QA tasks. These models generate more clinically aligned responses, though they are not always grounded in retrievable, verifiable context, leaving them vulnerable to hallucinations. Jeong et al. [JGLO24] and Shrestha et al. [SAK+23] further highlight the gap in adapting LLMs to real-world medical applications.

To address this, Retrieval-Augmented Generation (RAG), introduced by Lewis et al. [LPP+21], offers a compelling solution by combining dense retrievers with generative language models. RAG frameworks enable context grounding, thereby improving factual accuracy and consistency. Recent applications in biomedicine [YJT+23], [vSDN+23] confirm the potential of this approach, particularly when paired with instruction-tuned models such as InstructBLIP [DLL+23] or MedCLIP [WWAS22].

Despite these advancements, the integration of visual classification pipelines with retrieval-augmented language generation remains underexplored—especially in dermatological domains. While models such as BLIP [LLXH22], Mini-InternVL [GCC+24], and BioMedCLIP [ZXU+25] show promising results in general medical vision-language tasks, their deployment for fine-grained, patient-specific dermatology applications still lacks sufficient validation and interpretability mechanisms.

Hypothesis and Research Question

Hypothesis: An integrated architecture that combines a fine-tuned image classifier with a Retrieval-Augmented Generation module and a medical large language model will give more accurate, interpretable, and user-centric responses compared to traditional black-box classifiers or LLM-only systems.

Research Question: Can a pipeline composed of a visual classifier, a retrieval-based context gathering system, and a LLM produce medically reliable, context-aware answers to user questions about dermatological conditions?

Preliminary Approach: Vision-Language Model Architecture

Prior to finalizing the classification-based architecture, we initially experimented with a multimodal design inspired by Vision-Language Models (VLMs). The objective was to develop a unified system capable of jointly processing both the input image and the user's question within the same language model to produce a context-aware, medically accurate response.

The proposed architecture consisted of three components:

- **Vision Encoder:** Extracts high-level feature embeddings from the input image.
- **Mapper:** Projects the visual embeddings into the latent space of the language model to enable joint representation.
- **Language Model (LLM):** Receives both the text query and the mapped visual features to generate a coherent and medically informed response.

Using this setup, we evaluated multiple vision encoders including ResNet, EfficientNet, and ViT. All vision models were fine-tuned on our custom 30-class dermatology dataset. The results across these encoders were comparable, indicating that the choice of vision backbone had limited impact on overall system performance.

On the language modeling side, we experimented with two prominent LLMs: `T5` and `GPT-2 XL`. Both models were fine-tuned end-to-end on our dataset, using the combined visual and textual input for training.

Performance was assessed using standard natural language generation (NLG) metrics that evaluate the quality of generated text based on various linguistic dimensions:

- **BLEU-4:** Measures the precision of 1-to-4-gram overlaps between the generated text and reference answers, commonly used for evaluating fluency and grammatical consistency.
- **ROUGE:** Focuses on recall-based n-gram overlaps, indicating how much of the reference content is captured by the generated output—often used in summarization tasks.
- **METEOR:** Considers synonymy, stemming, and word order to provide a more semantically-aware measure of sentence-level alignment, making it useful for medical and contextual QA tasks..

The results are summarized as follows:

Model	BLEU-4	ROUGE	METEOR
T5	0.27	0.63	0.56
GPT-2 XL	0.28	0.24	0.18

Table 1: Performances of T5 and GPT-2 XL models by standard NLG metrics.

Discussion: Although the vision encoder type had little effect on the final outputs, significant differences were observed between the two language models. `T5`, being an encoder-decoder model optimized for sequence-to-sequence tasks like summarization and QA, consistently produced more fluent and contextually aligned responses. In contrast, `GPT-2 XL`, while capable of generating diverse and high-coverage outputs due to its scale, struggled with maintaining consistency and coherence in medical settings, as reflected by its lower ROUGE and METEOR scores.

These experiments demonstrated the potential of VLM-based designs but also highlighted the critical role of the language model in determining overall system effectiveness. This insight ultimately guided our shift toward a more modular and interpretable pipeline—combining image classification, RAG-based retrieval, and a dedicated LLM—which provided improved control, transparency, and performance for medical QA generation.

Proposed Solution

Our proposed system is composed of three primary components: image classification, context retrieval, and natural language generation.

We evaluated several pretrained image classifiers including ResNet-50, Swin Transformer, ViT-Base (ImageNet), and ViT-Base pretrained on the DermNet dataset. After comparison, the ViT-Base DermNet-pretrained model was fine-tuned using our custom 30-class dermatology dataset. It achieved the highest classification accuracy ($\sim$91%) and was selected as the final image classification model.

Once a user uploads an image and a question through the web interface, the image is classified by the ViT-Base model to predict the most likely skin condition. Based on the predicted class, the system accesses a domain-specific knowledge base (in JSON format) containing a brief disease description and 50 question-answer (QA) pairs generated via GPT-o3.

The user's original question is then reformulated into formal medical language by the selected LLM. Both the original and refined versions are embedded using `all-MiniLM-L6-v2` to produce semantic vectors. These vectors are compared to the questions and answers in the knowledge base using cosine similarity. A hybrid similarity score—a weighted average of question-question and question-answer similarities—is used to rank and select the top matching QA pairs above a similarity threshold.

Finally, the original question, the refined question, and the top-relevant QA pairs are passed again to the same LLM, which generates a concise, empathetic, and medically accurate response. This response is then presented to the user via a chat interface, enabling an interactive and informative consultation experience.

The proposed system bridges the gap between accurate image classification and contextual, explainable user communication. Preliminary results confirm that the fine-tuned ViT classifier achieves strong performance across various dermatological conditions. Additionally, the inclusion of a RAG module significantly improves the contextual accuracy of LLM-generated responses. The final deployment used `MedGemma-8b Q8`, selected through systematic LLM-as-a-judge evaluations for its superior overall performance.

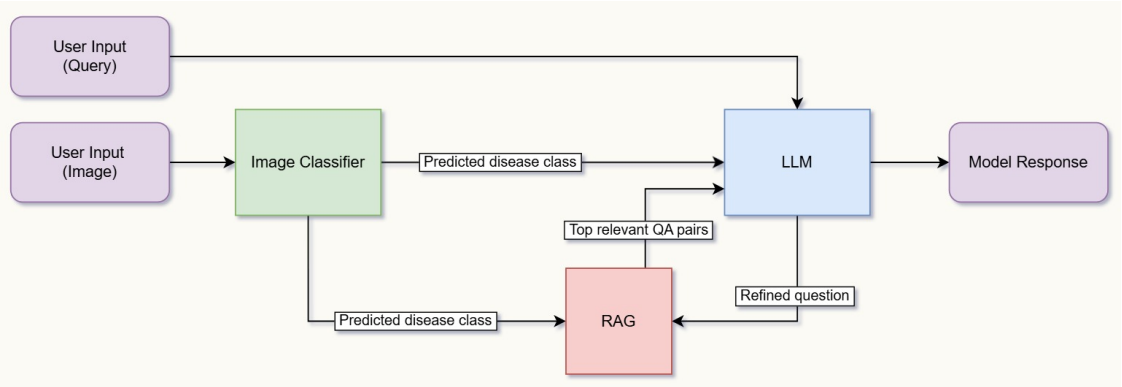

Figure 1: Overview of the proposed system architecture combining image classification, retrieval, and natural language generation.

Expected Impact and Future Work

The expected impact of this project is to deliver a scalable, reliable, and interpretable first-line dermatological consultation tool. In particular, it aims to reduce diagnostic delays and help users make more informed decisions about their skin health, especially in regions with limited access to specialists.

Future work will focus on supporting multi-label classification, extending the system to multi-turn dialogue interactions, integrating larger and more diverse dermatology datasets, and conducting clinical validation studies to assess the system's real-world effectiveness.

3 Methodology

Datasets

The primary data source for this project was a custom dermatology image dataset covering 30 distinct dermatological conditions. This dataset was specifically prepared for fine-tuning the Vision Transformer (ViT-Base) model, which was initially pretrained on the DermNet dataset. While the dataset reflects realistic clinical diversity, it is inherently imbalanced, with certain disease classes significantly underrepresented.

To address the class imbalance issue during training, a weighted loss function was employed. Class weights were calculated based on the inverse frequency of each class, ensuring that minority classes had a stronger influence on the learning process. This strategy helped improve model generalization and reduced bias toward dominant classes.

In addition to the image dataset, a textual knowledge base was generated using GPT-o3 for each disease class. For every class, the knowledge base includes a brief medical description and 50 question-answer (QA) pairs designed to reflect potential user queries. This resource was essential for supporting the RAG module, allowing the system to provide contextually relevant and medically grounded responses.

Data Pre-processing and Featurization

Image data went through standardized and systematic preprocessing steps, including:

- **Image resizing:** All images were resized to dimensions compatible with the model input requirements (224×224 pixels).
- **Color channel conversion:** All images were converted to RGB format to ensure uniform three-channel inputs.
- **Normalization:** Images were normalized using per-image mean and standard deviation values, according to the following formula:

$$x_{\text{norm}} = \frac{x - \mu}{\sigma} \tag{1}$$

where x represents the original pixel value, μ the mean pixel value, and σ the standard deviation.

The textual data was generated automatically by GPT-o3 and stored in JSON format. This structured knowledge base serves to retrieve contextually relevant QA pairs when answering user questions.

Computational Models and Algorithms

Image Classifier Models

Several pretrained image classifiers were evaluated during the project:

- ResNet-50 (pretrained on ImageNet)
- Swin Transformer (pretrained on ImageNet)
- ViT-Base (pretrained on ImageNet)
- ViT-Base (pretrained on DermNet)

Following comprehensive evaluations, the ViT-Base model pretrained on the DermNet dataset was fine-tuned on our custom 30-class dermatology dataset, achieving the highest classification accuracy (∼91%). Accordingly, it was selected as the final image classification model.

To address class imbalance and improve overall performance, the fine-tuning process utilized a *weighted Cross-Entropy Loss* function. In this setting, class weights were calculated based on the inverse frequency of each class, such that underrepresented classes had a proportionally greater influence on the loss. This strategy helped the model to generalize better across all categories.

The loss function used during training is defined as:

$$\text{Loss} = -\sum_{i=1}^{N} w_i \cdot y_i \log(p_i) \tag{2}$$

where y_i represents the ground-truth label, p_i denotes the predicted probability for class i, and w_i is the weight assigned to class i based on its frequency in the training set.

Retrieval-Augmented Generation (RAG)

The RAG module was developed to understand user queries and retrieve contextually relevant information through the following steps:

1. The user's original question is reformulated by the selected Large Language Model (LLM) into formal medical language.
2. Both original and reformulated questions are embedded into semantic vectors using an embedding model (denoted as `all-MiniLM-L6-v2`).
3. These vectors are compared with GPT-o3-generated QA pairs using cosine similarity:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \tag{3}$$

where A and B represent the semantic vectors being compared.

A hybrid similarity score is computed by combining question-question and question-answer similarities using the weighted average formula:

$$\text{Hybrid}_{\text{score}} = \alpha \cdot \text{Sim}_{\text{QQ}} + (1 - \alpha) \cdot \text{Sim}_{\text{QA}} \tag{4}$$

Here, α was set at 0.7. The highest-ranking QA pairs surpassing a predefined threshold were selected for generating user responses.

Large Language Model (LLM)

The LLM module operates in two distinct phases:

- **Question Reformulation:** The user's original query is reformulated into formal medical terminology by the selected LLM.
- **Answer Generation:** Both the original and reformulated questions, along with contextually relevant QA pairs retrieved by the RAG module, are again provided to the LLM, generating medically accurate, empathetic, and comprehensible responses to user queries.

Using an LLM-as-a-judge methodology, three models—`Medgemma-4b Q8`, `Meta-Llama-8b Q8`, and `Gemma3-4b Q8`—were systematically evaluated. `Medgemma-4b Q8` was selected as the best-performing model.

System Training and Testing Process

The overall system comprises two core components: an image classification module and a language generation module. The image classifier, based on a ViT-Base architecture, was trained using 80% of a custom dermatological dataset, with the remaining 20% reserved for validation. Model performance was evaluated using standard metrics such as accuracy, sensitivity (recall), specificity, precision, and F1-score to ensure balanced classification across 30 skin condition classes.

For the language generation component, a large language model was used to produce answers based on retrieved medical content. These responses were evaluated using an LLM-as-a-Judge approach, in which the model was prompted with the original question, a reference answer, and the generated output. It then rated the output on three criteria—accuracy, emphaty, and medical consistency—each on a scale from 0 to 10.

Principles of Model Evaluation

Evaluation of Image Classification

Model evaluation was based on the following metrics:

Accuracy: Measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total number of predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Precision: Indicates the proportion of true positive predictions among all positive predictions made by the model — useful when the cost of false positives is high.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

Recall (Sensitivity): Represents the ability of the model to identify all relevant (true positive) cases — important when minimizing false negatives is critical.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

F1 Score: The harmonic mean of precision and recall, providing a balanced measure when there is an uneven class distribution or when both false positives and false negatives are important.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Where TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) represent the components of the confusion matrix.

Cross-validation methods and threshold variation analyses were employed to ensure robustness and generalizability of the model's performance.

Evaluation of LLM-Generated Responses (LLM-as-a-Judge)

To assess the quality of language model outputs in medical QA generation, we employed an LLM-as-a-Judge evaluation strategy.

In this setup, a dermatology-specialized language model is prompted with the following inputs:

- The original user question
- A reference answer (original answer)
- A model-generated answer

The evaluating LLM is asked to score the generated answer in terms of three criteria:

- **Accuracy:** Does the generated answer correctly and completely address the user's question in alignment with the reference answer?
- **Emphaty:** Does the answer communicate in a human-like, supportive, and patient-aware tone, especially appropriate for medical contexts?
- **Medical Consistency:** Is the response factually aligned with clinical guidelines and free of contradictions, hallucinations, or unsafe suggestions?

All scores are given on a scale of 0 to 10. The evaluation prompt explicitly instructs the LLM to consider both the original question and the reference answer as context, and then judge only the generated answer — avoiding any additional commentary in its reply.

This format enables structured, scalable, and domain-aware judgment of open-ended language model responses, addressing limitations of traditional string-based metrics like BLEU or ROUGE.

4 Results and Discussion

This section presents the detailed outcomes of the project, including training, validation, and optimization of image classification models, performance evaluations, outputs from the RAG module, comparative analysis of different LLMs and use-case scenarios.

Training and Evaluation of Image Classification Models

Throughout the project, four pretrained image classification models were tested: ResNet-50, Swin Transformer, ViT-Base (ImageNet), and ViT-Base pretrained on DermNet. All models were trained on the same dataset and evaluated using a dedicated validation set. The table below summarizes the training and validation performance of the models:

Model Name	Training Accuracy	Test Accuracy
ResNet-50 (ImageNet)	0.8970	0.80
Swin Transformer (ImageNet)	0.9693	0.87
ViT-Base (ImageNet)	0.8562	0.89
ViT-Base (ImageNet & DermNet)	0.9756	0.91

Table 2: Training and Validation Metrics for Image Classifiers

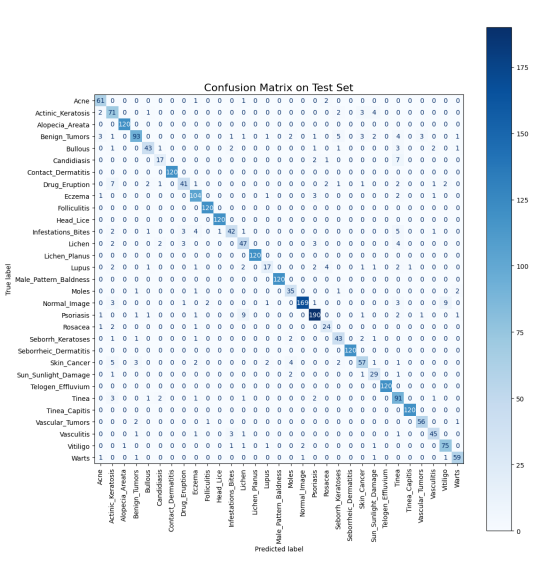

Figure 2: *
Confusion Matrix of ViT-Base (ImageNet & DermNet)

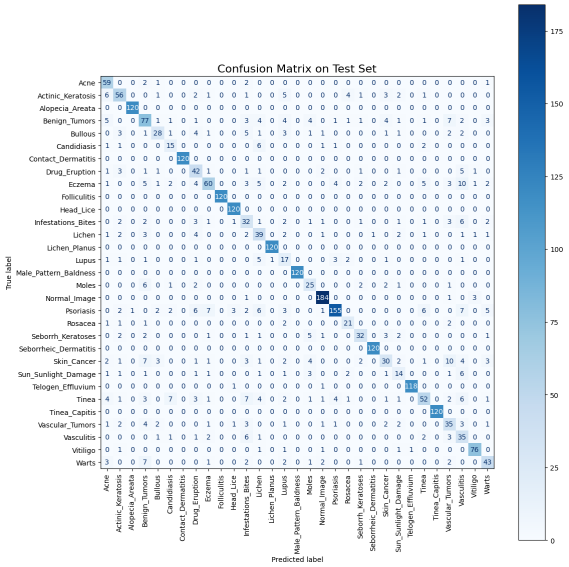

Figure 3: *
Confusion Matrix of ResNet-50 (ImageNet)

Discussion: The ViT-Base model, pretrained on the DermNet dataset and fine-tuned with our custom data, was expected to deliver the best results. This is likely due to the ViT architecture's capacity to capture long-range visual dependencies and its superior generalization ability on dermatological image features compared to traditional CNN-based models.

Class	Precision	Recall	F1-Score	Support
Contact_Dermatitis	1.00	1.00	1.00	120
Drug_Eruption	0.85	0.67	0.75	61
Eczema	0.87	0.93	0.90	204
Folliculitis	0.98	1.00	0.99	120
Head_Lice	0.99	1.00	1.00	120
Infestations_Bites	0.86	0.77	0.77	60
Lichen	0.73	0.77	0.75	61
Lichen_Planus	1.00	1.00	1.00	120
Lupus	0.74	0.50	0.60	34
Male_Pattern_Baldness	1.00	1.00	1.00	120
Moles	0.78	0.88	0.82	40
Normal_Image	0.98	0.98	0.94	189
Psoriasis	0.93	0.91	0.92	208
Rosacea	0.73	0.86	0.79	28
Seborrh_Keratoses	0.78	0.84	0.81	51
Seborrheic_Dermatitis	1.00	1.00	1.00	120
Skin_Cancer	0.83	0.74	0.78	77
Sun_Sunlight_Damage	0.72	0.85	0.78	74
Telogen_Effluvium	1.00	1.00	1.00	120
Tinea	0.71	0.89	0.79	102
Tinea_Capitis	0.99	1.00	1.00	120
Vascular_Tumors	0.93	0.93	0.93	60
Vasculitis	0.88	0.87	0.87	52
Vitiligo	0.86	0.91	0.89	80
Warts	0.91	0.92	0.91	64
Accuracy	-	-	**0.91**	2746
Macro Avg	0.88	0.87	0.87	2746
Weighted Avg	0.91	0.91	0.91	2746

Table 3: Per-class classification performance metrics of the ViT-Base (ImageNet&DermNet) model.

Performance Analysis of the RAG Module

The Retrieval-Augmented Generation (RAG) module is responsible for refining user questions and retrieving the most relevant QA pairs. Performance outputs obtained through this module are summarized below:

Disease Class	Original Question	Refined Question	Matched QA Pairs (Titles/Summaries)
Head Lice	Does this disease make me smell bad?	Can head lice cause body odor?	'question': 'Do lice cause a bad smell?', 'answer': 'No, but infected skin from scratching might develop an odor in rare cases.'
Acne	Is there any harm in shaving?	What are the potential risks associated with shaving?	'question': 'Can shaving reduce symptoms on the face?', 'answer': 'Shaving may help remove flaky skin but isn't a cure for the underlying condition.'

Table 4: Sample RAG Outputs: Question Refinement and Retrieval

Discussion: The effectiveness of the RAG module depends largely on the clarity and structure of question reformulation. Its performance is directly influenced by the quality of embeddings, the threshold selection, and the breadth of the QA database. The success of the RAG system significantly contributes to the factual and contextual strength of the final LLM-generated answer, improving user trust and interpretability.

Comparative Analysis of Large Language Models (LLMs)

Two LLM models were evaluated during the project: `Medgemma-4b Q8` and `Gemma3-4b Q8`. The evaluation was conducted using the LLM-as-a-judge method. The table below summarizes each model's performance across key qualitative metrics:

LLM Model	Accuracy	Empathy	Medical Consistency
Medgemma-4b Q8	8.43	8.63	8.43
Gemma3-4b Q8	8.31	9.10	8.29

Table 5: LLM Model Evaluation Results

Discussion: The evaluation results highlight that Medgemma–4b Q8, a domain-specific language model, consistently outperformed the general-purpose Gemma3–4b Q8 across all key qualitative metrics—accuracy, empathy, and medical consistency. This superiority can be attributed to Medgemma's specialized pretraining on medical corpora, which enables it to better understand domain-specific terminology and context.

Use Case and Demo Analysis

The project's web-based chatbot interface enables users to upload images and ask questions, receiving immediate, reliable dermatological consultation. Example use cases and demo screenshots will be provided as follows:

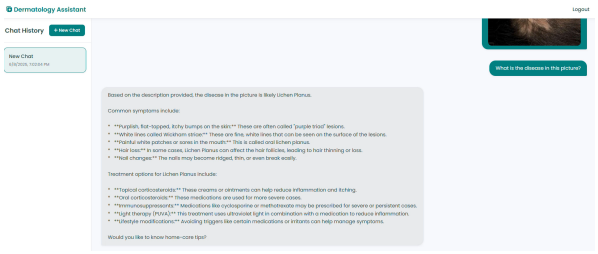

Figure 4: *
Use Case of The Project

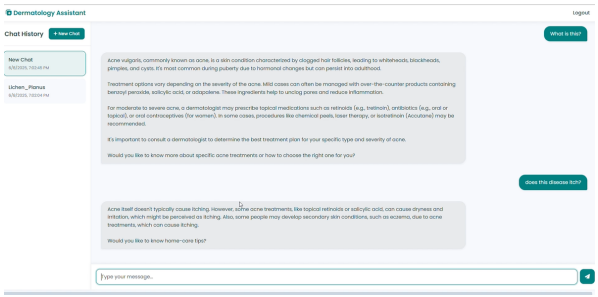

Figure 5: *
Use Case of The Project

Figure 6: Visual samples used during classification.

Discussion: This demo implementation and use-case simulation directly demonstrate the system's practical applicability and user experience. It is particularly expected to reduce the diagnostic gap in regions with limited access to dermatologists and support patients in making informed healthcare decisions.

5 The Impact and Future Directions

The methods and system architecture developed in this project provide a user-friendly, interpretable, and high-accuracy healthcare solution that facilitates early assessment of dermatological conditions. The prototype can serve as a pre-diagnostic assistant for patients, particularly in regions with limited access to dermatology specialists, by helping users better understand their skin concerns before clinical visits and empowering them to make more informed healthcare decisions.

Potential real-life impacts include:

- **Primary Care Support:** The system can function as a preliminary screening tool for early symptom detection.
- **Patient Education:** By providing medically accurate and empathetic responses, the system promotes health literacy among users.
- **Teledermatology Integration:** The assistant could be embedded into existing telemedicine platforms, acting as an intelligent support layer for clinical decision-making.

The results of this project may be of interest to researchers in digital health, AI, and NLP, and are suitable for publication in reputable venues such as *IEEE Access*, *Journal of Biomedical Informatics*, or *npj Digital Medicine*. Open-sourcing the web-based version of the system may also provide value to both academic and commercial communities.

If the project continues, several enhancements can be prioritized:

- **Multi-label classification:** Detecting multiple co-occurring dermatological conditions in a single image.
- **Multilingual support:** Enabling interaction with users in various native languages.
- **Clinical validation:** Evaluating the system's responses with feedback from dermatologists and healthcare professionals.
- **Multi-turn dialogue:** Supporting follow-up questions and contextual dialogue over time.
- **Mobile deployment:** Making the system available on mobile devices to increase accessibility.

Altogether, this project represents a technically robust and practically applicable digital health infrastructure that can be scaled, refined, and deployed for broader impact in the near future.

References

[DBK+21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[DLL+23] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[EKN+17] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, Feb 2017.

[GCC+24] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Jifeng Dai, and Wenhai Wang. Mini-internvl: A flexible-transfer pocket multimodal model with 5

[GTC+21] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, October 2021.

[HKL+18] Seung Seog Han, Myoung Shin Kim, Woohyung Lim, Gyeong Hun Park, Ilwoo Park, and Sung Eun Chang. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *Journal of Investigative Dermatology*, 138(7):1529–1538, 2018.

[Jav22] M. H. Javid. Melanoma skin cancer dataset of 10000 images. https://www.kaggle.com/datasets/mhasnain/melanoma-skin-cancer-dataset-of-10000-images, 2022. Kaggle.

[JGLO24] Daniel P. Jeong, Saurabh Garg, Zachary C. Lipton, and Michael Oberst. Medical adaptation of large language and vision-language models: Are we making progress?, 2024.

[Kum24] S. Kumari. Facial skin diseases dataset. https://www.kaggle.com/datasets/shwetakk/facial-skin-disease, 2024. Kaggle.

[LHZ+23] Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Xiaotang Gai, Yang Feng, and Zuozhu Liu. A chatgpt aided explainable framework for zero-shot medical image diagnosis, 2023.

[LLXH22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[LPP+21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[LSX+22] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), September 2022.

[MHW+23] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner, 2023.

[SAK+23] Prashant Shrestha, Sanskar Amgain, Bidur Khanal, Cristian A. Linte, and Binod Bhattarai. Medical vision language pretraining: A survey, 2023.

[SAT+22] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge, 2022.

[vSDN+23] Tom van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees G. M. Snoek, and Marcel Worring. Open-ended medical visual question answering through prefix tuning of language models, 2023.

[WWAS22] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text, 2022.

[YJT+23] Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training, 2023.

[ZXU+25] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025.