Lightweight Visual Question Answering (VQA) Model for Skin Disease Detection

by

Raiyan Habib Mahe
21301737
Abtahi Noor
21301304
Ridwan Noor Tasin
21101326

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

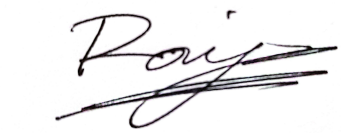
Department of Computer Science and Engineering
Brac University
June 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name and Signature:



Raiyan Habib Mahe
21301737

Abtahi Noor
21301304

Ridwan Noor Tasin
21101326

Approval

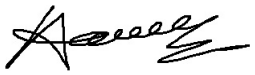
The thesis/project titled "Lightweight VQA Model for Skin Disease Detection" submitted by

1. Raiyan Habib Mahe (21301737)
2. Abtahi Noor (21301304)
3. Ridwan Noor Tasin (21101326)

of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 20, 2025.

Examining Committee:

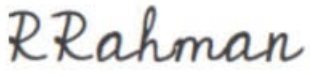
Supervisor:
(Member)



Dr. Amitabha Chakrabarty

Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)



Rafeed Rahman

Senior Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Md Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Chairperson
Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Visual Question Answering (VQA) is an area of artificial intelligence that combines image analysis with natural language understanding to generate context-aware answers to visual queries. Its application in the medical field, particularly in dermatology, holds transformative potential by enabling accessible, interpretable, and efficient diagnostic support. However, there is a lack of structured VQA dataset for skin diseases that can be used for training as well as benchmarking models. Moreover, existing VQA model are all extremely heavy weight and require specialized hardware to train and run. Hence, we developed a custom dataset of 1,038 images annotated by a medical expert for 11 disease classes, with seven question-answer pairs per image. The entire dataset is split into three sections: Train set with 833 images, Test set with 103 images and Validation set with 102 images. In addition, this research proposes a lightweight VQA model pipeline capable of identifying common skin diseases from images and responding to clinically relevant questions related to disease name, severity, causes, diagnostic approach, prevention, contagiousness, and cancer risk. The model uses a modular architecture that integrates a Vision Transformer (ViT) with 86 million parameters for image encoding and MiniLM, a transformer-based text encoder with 22 million parameter, ensuring high accuracy while minimizing computational requirements. We have also trained state-of-the-art vision language models such as Gemma-3, QwenVL-2.5, LLaVA-1.5, and BLIP-2 using our dataset for comparison. We achieved high semantic similarity scores on these models having 93.52%, 94.54%, 93.78%, and 94.47% BertScore on Gemma-3, QwenVL-2.5, LLaVA-1.5, and BLIP-2. Unlike these models, our solution is optimized for performance in low-resource settings. The system provides an intuitive interface for users to interact and receive actionable insights, benefiting both patients and healthcare professionals by reducing diagnostic delays and improving decision-making with accuracy of 94.87% with a total parameter count of only 108 million, rivaling those with billions of parameters. The results demonstrate that lightweight, domain-adapted VQA models can effectively bridge the healthcare accessibility gap through accurate and interpretable AI assistance.

Keywords: disease, detection, Visual Question Answering(VQA) , medical, severity assessment, image input.

Acknowledgement

Firstly, all praise to the Great Almighty Allah for keeping us strong mentally and physically- to complete our thesis titled "Lightweight Visual Question Answering (VQA) Model for Skin Disease Detection" within the deadline without any difficulties.

We are wholeheartedly thankful to our supervisor, Dr. Amitabha Chakrabarty, Professor, Department of Computer Science and Engineering, BRAC University for his invaluable support, supervision, suggestions, and motivation throughout the thesis journey. His unwavering support allowed us to progress our work and ideas flawlessly. We also express gratitude to our co-supervisor, Rafeed Rahman, Senior Lecturer, Department of Computer Science and Engineering, BRAC University and Mr. MD.Fahim-Ul-Islam for their guidance and support which were essential to the accomplishment of our research.

The success of this study is not solely the result of our individual efforts, but rather the outcome of the collective contributions of many individuals. We are especially grateful to the Department of Computer Science and Engineering for their constant support and overall assistance throughout the thesis period.

Lastly, we would like to extend our heartfelt thanks to our parents for their unwavering support and belief in us. Their constant encouragement and understanding have been essential in helping us complete our thesis while pursuing our professional goals.

Table of Contents

List of Figures

List of Tables

Nomenclature

VQA - Visual Question Answering
ViT - Vision Transformer
RL - Reinforcement Learning
FPN - Feature Pyramid Network
MLB - Multimodal Low-rank Bilinear pooling
MUTAN - Multimodal Tensor-based Tucker Decomposition
CLIP - Contrastive Language-Image Pre-training
LLM - Large Language Model
NLM - Neural Language Model
JUST - Name of a model mentioned
SLE - Soft Label Encoder
MA - Mutual Attention
SAU - Self-Attention Unit
TM - Transformer Module
BLEU - Bilingual Evaluation Understudy
SOTA - State-Of-The-Art
DAE - Denoising Auto-Encoder
BAN - Bilinear Attention Networks
PubMedCLIP - CLIP model fine-tuned on PubMed articles
SQuINT - Approach focused on Reasoning questions with Perception sub-questions
VLP - Vision Language Pre-training
GPT - Generative Pre-trained Transformer
CC - Conceptual Captions dataset
COCO - Common Objects in Context dataset
ISIC - International Skin Imaging Collaboration
PH2 - Name of a dataset
MCB - Multimodal Compact Bilinear Pooling
HAM10000 - Human Against Machine dataset
GloVe - Global Vectors for Word Representation
SAN - Stacked Attention Network
NegBio - Negation detection system for radiology reports
MIMIC-CXR - Medical Information Mart for Intensive Care dataset
RadVisDial - Radiology Visual Dialog dataset
GoogLeNet - Google's Inception Neural Network

Chapter 1

Introduction

Over the past years, artificial intelligence (AI), incorporated into the medical industry has induced cutting-edge innovations that allowed making the process of diagnosing patients faster, more precise, and scalable. VQA is a task that requires the integration of computer vision and natural language processing with the aim to enable AI systems to answer open-ended questions about images, demonstrating human-like reasoning capabilities [2]. The special feature of this method of visual data processing is that they are able to analyze and also provide useful responses to questions that are relevant and are asked in natural language. VQA systems are specifically needed where interpretability and context-awareness are required.

There is a significant requirement of early and correct diagnosis of skin disorders in the medical field, especially in the field of dermatology. But the fact is that millions of individuals all over the world (in underdeveloped or rural areas most of all), do not have access to professional dermatologists or even to any diagnostic equipment [3]. This puts an urgent demand on having an intelligent system that can facilitate or assist in medical decision making using images and visual data as well as queries based on symptoms. A customized VQA model in skin disease diagnosis can hold an immense potential to fill this gap. Instead, the user can upload skin images and pose appropriate questions (What is this disease? Is it contagious? What is the recommended treatment?) on such a system and get instant and informative feedback, empowering not only the patients, but healthcare professionals as well.

A medical-grade VQA system is quite a complex venture, although its potential remains high. Medical images are dynamic in nature and can be interpreted based on the subtleness that is viewed in the imagery, so it requires excellency in interpretation [1]. Questions might be extremely different in form and purpose: identification ones and severity assessment ones or treatment planning questions. In addition, the handling of effective VQA models typically requires big annotated datasets and significant computational resources, and they are scarce in the healthcare sector because of patient privacy implications and the requirement of professional curation.

The purpose of the presented research is to address these kinds of problems by proposing a domain-specific lightweight model pipeline of VQA model that can diagnose the usual skin conditions using an image and answer the predefined series of clinically relevant questions. A combination of compact vision transformer (ViT) to extract visual data and transformer based text encoder optimizes our model to have high results and low computational footprint. We also have developed custom the set of annotated set of 1,038 photos in 11 categories about skin diseases where the provided each picture 7 sets of questions and answers about the diagnosis, high degree, the reasons of skin diseases, and the methods of their avoidance.

This research not only advances the application of VQA in the field of dermatology but also contributes to the development of practical, deployable AI tools for real-world clinical environments. Our goal is to bridge the accessibility gap in dermatological care through an intelligent, interpretable, and efficient VQA system that supports informed decision-making and promotes equitable healthcare delivery.

1.1 Motivation

Availability of healthcare is another important challenge, especially in the rural and underserved communities, whereby most people do not have access to specialist dermatologists, diagnostic services and basic health care [3]. Moreover, low levels of knowledge in the area of hygiene and prevention of disease further contribute to the problem of these populations. Easy to reach technology enabled solution to this problem is becoming an urgent necessity to mainly deal with early stage diagnosis. A possible solution is a representation of the form of Visual Question Answering (VQA) systems, which combine analysis of images and understanding of the natural language that allows people to load photos and generate questions, as a result of which they will immediately receive convenient, practical feedback [2]. Nonetheless, some models that are already in place are resource-consuming, and they are not practical to implement in low-resource settings because they either need specialized hardware or internet connectivity.

In this study, we are filling this gap together by designing an efficient, domain-specific VQA model that can be used to diagnose skin diseases in a manner that is both clinically accurate and lightweight such that it can be executed on an off-the-shelf hardware with minimal computing requirements. This intends to deliver real-time diagnostic assistance advantageous to patients and medical workers and also to improve diagnostic performance and precision in settings with poor access to medical professionals and, finally, improve the delivery of equitable healthcare services.

1.2 Research Problem

The concept of creating a quality and effective Visual Question Answering (VQA) model in the context of skin disease diagnosis is a complex research task. Although the potential applications of VQA have proven to be impressive in the sphere of combining visual and textual information on general tasks, the same range of properties is required in a medical field of dermatology. Medical images particularly those referring to skin condition are very diverse in terms of its texture, color, resolution and the way it is viewed upon different persons. This makes it difficult to generalize well using models, particularly on such differences as visual similarity of symptoms which define different diseases.

One of the biggest challenges, in that regard, is that there are only a small number of annotated medical datasets. Before meaningful performance may be observed, VQA models usually need significant numbers of question answer pairs of images. In the healthcare field, however, such data is hard to obtain as a result of strict privacy policies and the necessity of domain experts to obtain specific annotations that reflect actual data. Consequently, most of the current models either have the overfitting effect, lack of generalization, or reliance on non medical pre-training databases that lack the finer details of skin diseases. This inability to coach the domain specific learning limits the model in relation to provision of clinically reliable outputs.

A key issue resides also in the computational cost of a large portion of state of the art VQA models. Such models also tend to need a lot of hardware to train and make inference with, making them difficult to deploy in resource-constrained settings e.g. a rural clinic or mobile health platform. Lightweight models are in urgent demand that are efficient and do not compromise the accuracy of the diagnosis. Besides, the majority of the available models comprise limited diagnostic abilities. Although they can perfectly recognize the names of diseases, they usually cannot evaluate its severity or recommend preventive steps, and take any actions, which are essential in a real-life application in the medical field.

Besides these problems, the problem of consistency in reasoning is also yet to be solved. The model may have the ability to answer a key question in diagnosis properly, but it may not give a proper answer to the subsidiary sub-questions or rationalize its conclusions in a coherent manner. The inconsistency of the reasons negatively affects the reliability of the system and can be a significant threat to clinical decision-making. Accordingly the main research question is to develop a VQA model that would have high diagnostic accuracy using a very little annotated dataset, would also be able to perform multimodal reasoning, would be really interpretable, would have very low computation needs, and would be applicable to the real-life medical setting because of the extensive and consistent diagnostic answers.

1.3 Research Objective

The main objective of this research is to improve the accuracy and efficiency of Visual Question Answering (VQA) systems in the medical domain, specifically for skin disease diagnosis. We aim to achieve this by creating a domain specific dataset and fine-tuning multiple vision language models to perform better on visual dermatology questions. To begin, we created a custom annotated skin disease dataset. This dataset consists of 1,038 images, covering 11 different skin conditions. Each image is paired with 7 question-answer sets, focusing on clinical features such as color, shape, border, severity, and diagnosis. The goal of this dataset is to represent the domain knowledge of skin disease diagnosis in a structured and interpretable way.

The next objective was to fine-tune state of the art vision-language models on this dataset. We selected four pre-trained models: Gemma 3, QwenVL-2.5, BLIP-2, and LLaVA-1.5. These models were fine-tuned using our annotated dataset to adapt them to the task of skin disease VQA. After fine-tuning, we evaluated their performance using semantic similarity metrics. This is because the generated answers are open ended which cannot be directly matched with ground truth from the dataset. Hence we used semantic similarity metrics such as BERTScore and Sentence-BERT score which are able to understand the contextual meaning of text for finding a score. Using these metrics we measured how close the model-generated answers were to the ground truth in terms of their contextual meaning.

Other than making fine-tuning of large models, we also wanted to develop an alternative lightweight model that would be able to do the same work of visual question answering under reduced computing demands. To meet this, we constructed a comparable ViT-based encoder of images having the same dataset. To every image, we also derived valuable characteristics of its question-answer pairs and coded them to be structured data. This is the visual and the diagnostic features representation of the image. In order to model the language understanding component of this lightweight structure, we utilized MiniLM which is a small language model. MiniLM uses a text question provided by a user and compares its semantic match with predefined attribute keywords during the encoding. According to the greatest similarity, the right response is chosen out of the encoding. This solution enables this model to produce quick, sensible answers without the leverage of large-scale generative models.

Overall, the intended impact of the research is to develop a good quality annotated dataset on skin disease VQA, assess how fine-tuned vision-language models perform in a specific domain, and possibly present a lightweight alternative using ViT and miniLM. Along with accuracy, this combination is focused on efficiency and is applicable to real-world medical use, where speed and always-interpretable results are important factors to consider, and where domain alignment must be optimized.

1.4 Research Contributions

The key contributions of this research are as follows.

- In our research, we developed a lightweight Visual Question Answering (VQA) model for the purpose of skin disease diagnosis through the integration of a Vision Transformer (ViT) and a MiniLM-based text encoder in order to optimize accuracy while minimizing computational requirements.
- We prepared a custom domain-specific dataset annotated by medical experts which consists of 1,038 skin disease images covering 11 disease categories with seven structured question-answer pairs per image allowing better training of dermatology-focused VQA models.
- We tried to minimize the challenge of high-parameter models by proposing a modular VQA architecture, allowing separate training for image and text encoders, which helped to reduce model complexity without compromising diagnostic performance and accuracy.
- We compared our model with state-of-the-art (SOTA) VQA models, showing that our lightweight model performs competitively well in terms of both accuracy and efficiency, making it suitable for deployment in low-resource environments.
- We mitigated the issue of data scarcity in medical VQA tasks for skin disease detection by leveraging supervised learning, expanding the model's generalization capabilities with limited annotated data.
- We demonstrated a feasible and easy solution for scalable web deployment in clinical settings with low computational resources.

1.5 Thesis Outline

Chapter 1

This chapter presents Visual Question Answering (VQA) in the medical field—more especially, for the diagnosis of skin conditions. It clarifies the relevance of artificial intelligence in the medical field and how VQA models—which combine natural language processing with image analysis—may improve diagnosis procedures. Particularly in areas with restricted access to healthcare professionals, the chapter offers a general picture of the difficulties in identifying skin diseases. It summarizes the research problem, goals, and contributions, thus clarifying the inspiration behind the development of a lightweight VQA model for skin disease diagnosis.

Chapter 2

Reviewing the body of current research on VQA, this chapter especially addresses its use in the medical field—more especially, dermatology. It looks at several VQA models applied for skin disease detection, stressing advantages, drawbacks, and the dearth of domain-specific data. The chapter investigates the disparity in medical image VQA research and points up chances to create more easily available models for healthcare.

Chapter 3

The custom dataset produced for this study is detailed in this chapter. It covers specifics on the choices of images, annotations, and the method applied to produce question-answer pairs for every skin disease picture. The chapter also addresses preprocessing techniques used on the dataset, data splitting into training, validation, and test sets, and a comparison with other already available VQA datasets.

Chapter 4

Beginning with the design and preparation of the dataset, this chapter offers the approach followed in this research. It shows the results of fune-tuning already-existing state-of- the-art Vision Language Models (VLMs) including Gemma-3, QwenVL-2.5, and BLIP-2 on the custom dataset. The chapter also presents the proposed lightweight VQA model, which combines MiniLM for text understanding and Vision Transformer (ViT) for image encoding. Furthermore contents covered are model evaluation, optimization techniques, and training parameters.

Chapter 5

The application of the models including the training parameters, loss functions, and performance criteria is presented in this chapter. It presents the findings of the custom dataset and compares the lightweight model with the fine-tuned big models. We present a thorough investigation of the accuracy, precision, recall, F1 score, and semantic similarity of the model. Performance visuals and a review of the findings also abound in this chapter.

Chapter 6

This chapter details the method of using a Flask-based web application to distribute the trained VQA model. The chapter addresses the integration of the model into an understandable system for users to interact with and the evolution of the web interface for real-time skin disease diagnosis. Furthermore covered are system limitations and possible uses in resource-limited and offline environments.

Chapter 7

Emphasizing the value of the lightweight VQA model to skin disease diagnosis, the last chapter compiles the main results of the research. It also offers possible directions for next research: broadening the dataset to include rare diseases, model optimization for maximum efficiency, and application of the model in actual clinical settings. Recommendations for additional study to increase the generalizability and functionality of VQA systems in healthcare round out the chapter.

Chapter 2

Literature Review

We will now take a look at some of the relevant existing work done in our area of study:

In paper [13], the authors introduced a new fusion mechanism for skin cancer detection. Skin cancers have come to bear significant morbidity and early detection is imperative to increase survival. Deep learning is an effective method of aiding diagnosis. In this paper a new architecture has been proposed that fuses models to utilize both image and text information. CNN are effective in understanding image and heavily used in computer vision. Modern CNN models such as VGGNet, ResNet, GoogLeNet are very effective in learning the input image. RNN are useful for extracting text information. In the proposed model, two pretrained models were used to obtain text and image features together at the same time. In the image input black, Faster R-CNN is used which uses ResNet 101 as its backbone. For the text input block, first the pairwise meta data was padded with 10 words then a pretrained GloVe weights were used to transform the text data with embedding and fed to 1 layer LSTM.

Paper [5] is about RadVisDial, a dataset for VQA tasks in radiology which is derived from MIMIC-CXR. Every image, I is accompanied by a caption, C and a history, H, which is 10 question-answer pairs. The objective is that given an I, C and an empty H with a follow-up question q, it will generate answers a where q,a belongs to H. The Radiology reports contain a part called Medical Condition section which is a single sentence description of the patient's medical history. NegBio is used to extract sections within the report and treat this as the caption, C. To evaluate this dataset, SAN(Stacked Attention Network is being used). 512 hidden layers were used in the LSTM when training. The model also has a question representation that is based on LSTM and a stack of two image attention layers, the first of which is modified to include a term for LSTM. As a result, the image attention weights become question and history guided.

The paper [19] explores the different models used for Medical VQA tasks. Usually CNN was used to obtain image features and RNN was used to obtain textual features. The VQA-Med 2019 dataset is used. VGG-16 and LSTM are chosen for image and text extraction as benchmark model and all other models are tested against it. The models that are being evaluated are ResNet-50 and ResNet- 152 for

image encoder, BERT for text encoder and Stacked Attention Network for Feature Fusion. In order to increase the size of dataset, they used image normalization and augmentation. Results show that VGG-16 for image, BERT or BioBert for text and Concatenation for fusion yielded the best accuracy of 60%. Complex networks like ResNet-50 or ResNet-152 show a higher degree of overfitting when working with small dataset.

The paper [9] discusses the state of the VQA in medical field showcasing ImageCLEF-2021. One of its task was to produce answers based on the input, which is Visual Question Answering. Another of its task was to produce questions from the input image, which is Visual Question Generation. The dataset for VQA was MedPix with several filters applied on few instances where only image was used for diagnosis. For model used on the VQG task, VQA-RAD dataset was used to train VQGR model and a T-5 based model trained on SQuAD and MACRO datasets. The team which achieved hurst accuracy on the VQA task was SYSU-HCP from China with 38.2% accuracy and 41.6% BLEU score. For the VQG task, team Chabbiimen got the highest average BLEU score (with 10 runs) of 38.3%.

The paper [11] discusses a new type of model called Multimodal Medical BERT (MMBERT). As previous models used different modes for image and text and trained them separately, large multimodal datasets were not utilized properly. Hence, could not properly learn text and image representation. In the model, instead of using the transformer to perform single self-attention, it is used to perform multiple self-attention in parallel and concatenates the output. The image feature is extracted from different convolution layers using the ResNet-152. It has 4 BERT layers and 12 attention heads. Masked vision-language modeling is used for pretraining. Only medical keywords were masked.

Paper [14] discusses a VQA method called TraP-VQA. It is a transformer based method for question-answering in the pathology field. Here, low-level image features are embedded with contextual information related to the image's domain. To extract question features, BioELMo is trained with 10 million abstract texts from PubMed and has a success rate of 64.82%. For image feature extraction, several models are being used. The best performing model was pre-trained ResNet-50. The extracted features are then sent to the transformer layer. The encoder of the transformed fused the vision and text features while the decoder upsamples the encoded features. The finished model was then experimented with the PathVQA dataset. The proposed model, with pre-trained ResNet for image extraction and BioELMo for text extraction worked best with 64.82% overall success rate with both open-ended and close-ended questions.

Both CNN and ViT can be a suitable approach for extracting image features for our model. The authors of this paper [12] introduced MedFuseNet which used CNN for extracting features from image. They implemented the model using attention-based deep learning which is multimodal and designed for Visual Question Answering. It addressed challenges like limited training data and privacy concerns, while also aiming to maximize learning with minimal complexity. Special importance was given to the interpretability of the results as its reliability can be increased. There are

four parts of their model. They are extracting image feature, question feature, feature fusion with attention mechanism, and categorizing answers. CNN was used to extract feature from image. Suitable models like BERT and XLNet were used to extract features from questions. Fusion techniques like Multimodal Compact Bilinear Pooling (MCB) were used to optimize interactions between modalities. Finally a LSTM based decoder with heuristic was used to improve model performance in generating answers. The datasets MED-VQA (2019) and PathVQA were used in the experiments. They have achieved superior results compared to some of the best VQA approaches.

Another Paper [28] introduced an approach to classify lesions using a transformer module and self-attention unit. CNN was modified to be used with it. Their model has got 97.48% and 96.87% of accuracy on multitissue classification among the datasets ISIC-2019 and PH2. The approach involved a multi-stage process where the global receptive field was used to extract global features and a CNN branch for local feature extraction. Bidirectional feature fusion enhanced recognition rates, and a fusion unit evaluated categorization vectors to fine-tune the model. The fusion unit employed a bi-directional fusion structure between the CNN and Transformer branches for efficiency and accuracy.

The main focus of this paper [7] was to address the challenge of Visual Question Answering (VQA) tasks that need to interpret complex perceptual data. Visual Question Answering (VQA) models often give the right answer to the main question but give wrong answers to sub-questions related to the main question. For example, when asked if a banana is ripe, the model correctly answers 'yes'. However, if we then ask if that banana is 'green' or 'yellow', it gives the wrong answer by mentioning 'green'. This type of inconsistency reveals the weakness of a model at properly interpreting a given image. For this reason, the authors have put special focus on the distinction between perception questions and reasoning questions. Current VQA datasets mix these question types without differentiation, leading to potential inconsistencies in model evaluation. The paper proposes a new dataset split focused only on Reasoning questions, supported by a sub-dataset of Perception sub-questions to ensure that the models answer Reasoning questions consistently and accurately. The authors named this novel approach as 'SQuINT'. The results show that SQuINT improves model performance without sacrificing accuracy. This way the reliability of the VQA models can be increased when they show consistency when answering the sub-questions.

This research [8] introduced unified Vision Language Pre-training (Unified VLP) after looking at the effectiveness of models like BERT and GPT. Existing methods often rely on separate encoder-decoder models or task-specific designs, which can be inefficient. This model used a unified multi-layer transformer on both encoding and decoding. Large dataset of captioned images was used to pre-train the model. BERT was used as its backbone. Their trained model did better decisively in almost all of the categories when compared to the baseline model. This approach can be useful as it gets rid of the need to use different pre-trained models for a particular task.

Traditional Visual Question Answering (VQA) methods need extensive labeled data, which is hard to obtain for various reasons such as privacy concerns. For example, the VQA-RAD dataset was manually constructed, but it includes only 315 images, insufficient for training powerful deep learning models effectively. To address these challenges, they [6] came up with two key observations. First is that a large number of unlabeled medical images are available, which if used to train an unsupervised model, could provide weights more suitable for medical VQA than those pretrained on general images like those in ImageNet. The VQA-RAD dataset is small but can be used to apply meta learning after getting class labels from it. The initial weights to extract image features were set by utilizing the MAML and CDAE, which improves accuracy with minimal labeled data. End to end fine tuning approach was used with the help of medical data. They used BAN for their attention mechanism. The final model did better than the baseline model by 16.3% on open ended questions and 8.6% on close ended questions. The process introduced by them can help a model to better adapt to the training data.

This paper [10] investigated the effectiveness of CLIP for MedVQA. Here the authors introduced PubMedCLIP. This is a visual encoder and is pre-trained. It is a version of CLIP fine-tuned using PubMed articles that contain thousands of image–caption pairs. The experiments were conducted using two well-known datasets, VQA-RAD and the SLAKE. Then, they Utilized two MedVQA methods which are Conditional Reasoning with Question answering and Mixture of Enhanced Visual Features. CLIP and PubMedCLIP used as visual encoder has improved the performance on both of the methods. Their model showed the best results with ResNet-50 backend which improved the accuracy of MEVF up to 6% and QCR up to 3%. Another take away from this paper is that CNN-based models like ResNet did better in identifying local features and performed better than Vision Transformers in case of VQA-RAD dataset. While Vision Transformer based models performed better with the SLAKE dataset where the questions required a comprehensive understanding of the entire image and capturing long-range dependencies.

The authors have conducted a comprehensive survey in this paper [21], on medical Visual Question Answering (VQA), highlighting its unique challenges and possibilities. The paper examines current datasets, emphasizing the value of different question categories in better simulating real-world situations in healthcare. They also emphasize the importance of incorporating more medical data and utilizing additional expertise to improve model performance and interpretation. The study examines the application of Large Language Models (LLMs) and points out some issues like biasness and misinformation.Despite mentioned challenges, the authors are suggesting collaboration with healthcare specialists and applying data from training for reducing risk factors.Besides, the survey also focuses on the probable advantages of adopting medical VQA technology into healthcare practices which will enable the healthcare specialists for better and effective communication and thus resulting in better decision making process. Furthermore, the study shows directions for potential research, giving importance to the application of real-world technologies, verification of data and analysis of biasness.In this way, the aim is to create useful dependable VQA technology for the medical field.

This research article [18] describes an innovative approach in skin disease classification that uses neural network. That takes advantage of both picture data and clinical metadata. The authors present an innovative multimodal Transformer architecture that includes distinct encoders for images and information, as well as a decoder to properly combine these modalities.. For image encoding, they employ Vision Transformer (ViT) models known for their skill in visual tasks, selecting the appropriate ViT model based on dataset size and performance requirements. A newly constructed Soft Label Encoder (SLE) facilitates metadata encoding by embedding descriptive labels into soft label vectors. It enables the network to acquire label correlations more efficiently than current techniques.The Mutual Attention (MA) block situated in the decoder is added to effectively mix metadata and images , containing Multi-head Cross Attention blocks along with residual links to ensure proper balanced representation of data from every single source. Extensive research on both private and standard datasets of the skin disease shows that the model which was proposed overperforms previous technologies. The metrics include accuracy, specificity, sensitivity, F1 score while AUC demonstrates its usefulness.Ablation examination encourages the contributions of the SLE and MA components rather than model performance.The research points out the need of adopting clinical information by the help of ViT models and adding new components to improve the accuracy of the diagnostic.Lastly, the researchers states that their multimodal design of the Transformer is a substantial improvement for identification of skin diseases. They stated the potential directions for future research and increasing accuracy of the identification using their proposed technique.

The following paper [23] shows a technique to improve the VQA (Visual Question Answering) System by proposing a feature extraction by combining both image and text.In case of representation of image, the mentioned method is improved based on a pre-trained residual network (ResNet-152). Features from various convolutional layers like conv3, conv4, conv5 are acquired and extracted, then fused by the help of a method which is inspired from Feature Pyramid Network (FPN).For representation of text, a multiscale approach is taken for feature extraction at word,phrase and sentence levels by the help of CNN (Convolutional Neural Networks) and skip-thought models.Then both the extractions are fused using bilinear pooling methods like MUTAN and MLB which helps in reducing the dimensionality of the mixed features.Experimental reports of the VQA dataset illustrates that mixing multiscale feature extraction with fusion mechanism improves the accuracy of both image and text information in the task of VQA.This study contributes to the advancement of the system of VQA by improving the task of feature extraction and fusion mechanism for both textual and visual modalities.

The research article [24] reviews various methodologies to analyze skin lesions and cancer detection, highlighting the significance of early diagnosis as mortality rate is high when it comes to skin cancer. It discusses traditional machine learning classifiers like SVM and KNN, which have shown varying degrees of success. Particularly, deep learning models like CNNs in achieving high classification accuracy. Even after having remarkable results, there exists some shortcomings like noise, low contrast and diversity of dataset.The study remarks that future research must include these challenges and find out more advanced techniques like quantum computing to im-

prove further detection and accuracy of classification.The survey underscores the potential of diagnosis system which is helped by computer systems to improve detection of cancer detection.

The paper [4] represents an overall summary of the VQA-MED task that was held at ImageCLEF 2019, which mainly focused on answering medical questions based on radiology images divided into four categories namely- Modality,Plane, Organ System and Abnormality. The new dataset made by them had 4200 images and 15,292 pairs of question and answer. It included 3200 images and 12,792 pairs of question-answer allocated for training and 500 images for validation and finally testing 500 QA pairs. These test sets went through manual validation of doctors. Application of deep learning techniques were used like transfer learning , multitasking learning along with attention mechanisms. The best team achieved 64.4% BLEU score and overall accuracy of 62.4% which indicated better performance on classification tasks rather than open-ended abnormality questions. While an improvement over 2018, challenges remain, motivating future editions to consider more complex questions involving contextual information and domain-specific inference to better support clinical applications.

This research paper [16] proposes a VQA model in medical imagery using a transformer encoder-decoder architecture. It has separate encoders for encoding the input image using a ViT using a text encoder based on BERT. The encoded visual and text extractions are concatenated and passed into a decoder which is multi-modal to generate the result in an autoregressive manner. The model is trained end-to-end on VQA datasets -PathVQA and VQA-RAD. The results from the proposed model performed better than existing models, resulting in maximum accuracy on questions which are both open and close ended. It obtains accuracies of 84.99% and 72.97% on VQA-RAD, and success rate of 62.37% and 83.86% on PathVQA respectively for closed and open ended. Qualitative examples with attention visualizations provide insights into the model's reasoning process. The key novelty lies in adapting transformer architectures for the medical VQA task by encoding visual and textual modalities separately before fusing them for answer generation in an end-to-end trainable framework.

In their paper [20], the authors present BLIP-2 as one of the VLP methods that use the frozen image encoders in combination with large language models for the sake of having an optimal computational cost and avoiding the leakage of knowledge retained by the models. Q-Former is the key system that provides a concise transformer that goes through two-stage guidance in order to enhance modality connection performance efficiently. On its first stage of training, the Q-Former becomes capable of extracting visual features associated with text by taking frozen image encoders as inputs. The system gets enhanced generative-learning powers as it uses a frozen LLM. Using BLIP-2 the researchers succeeded to obtain the best results on vision-language tasks such as visual question answering, image captioning, and image-text retrieval while having decreased trainable parameter counts as compared to previous methods. The performance of BLIP-2 outperformed Flamingo 80B by 15.6 points in zero-shot VQAv2 evaluation while only having 54 times fewer computational parameters. In the framework of the zero-shot image-to-text generation,

the model performs very efficient natural language instruction compliance.

The authors [25] introduced MedBLIP as a new vision-language pretraining (VLP) model proposed for medical computer-aided diagnosis based on electronic health records and 3D medical imagery. The most important part of MedBLIP is MedQ-Former since it resolves a dimensional inconsistency between three-dimensional views of medicine and vision encoders, as well as integrates visual and textual components. MedBLIP shows excellent zero-shot classification and VQA for medical images with the tests being run on over 30,000 images of Alzheimer's disease, and with high performance on unseen datasets. The model works effectively in delivering the diagnostic reasoning, making it more useful in the practical clinical setting. The lightweight system performs efficiently on minimal computing resources that represent the system's versatility beyond the use in the diagnosis of Alzheimer's disease.

This survey [34] gives an extensive account of the developments of the Visual Question Answering (VQA) from its early phases to current progresses. The paper reviews various VQA models, from old CNN-based architectural designs to the more recent transformer-based models such as ViLBERT, VisualBERT, and BLIP. The paper also discusses a set of fusion strategies meant for combining vision and language comprehension of VQA systems. It focuses on the use of the critical datasets such as MSCOCO, VQAv2, and Visual Genome for training and assessment. The challenges in the field are the bias in data, difficulty to generate robust models that generalize well on other domains, and interpretability challenges. The paper recommends future research in enhancing the zero-shot learning, model generalization, and real-world applicability; in particular, for medical imaging, robotics, and the assistive technologies. Improving the evaluation metrics and the problem of the data imbalance are also marked among the most significant aspects of the future development.

The technical report [30] of Qwen 2.5-VL by its authors encapsulates an advanced vision-language model that gives better object localization with better document parsing and long videos handling. This model receives images of different sizes in the absolute time encoding processes which will enable close observation of temporal occurrences during a lengthy video perusal. Natural dynamic-resolution Vision Transformers when coupled with optimized window attention generates reduces the hardware demands required on the platform. The results of evaluation of Qwen 2.5-VL meet the GPT4o and Claude 3.5 Sonnet standards and the model has an advantage in understanding diagrams, as well as processing documents. The application itself includes several size options that offer great ability over different devices without losing the high-level language abilities drawn from its large language model.

In their work [22] "Visual Instruction Tuning" the authors introduce LLaVA as a combined model that integrates vision encoder technology with elements of large language model (LLM) for end-to-end training to support a general vision-language understanding. The paper is the first to expand the language-only instruction tuning methods with multimodal instruction-following data generated by GPT-4. As the process of fine-tuning LLaVA, synthetic data is essential and enables the authors to develop new processes of assessment for practical visual tasks. Changes

applied to LLaVA allowed it to achieve 85.1% of GPT-4's performance levels on unrecognized images and instructions in terms of the synthetic data. Combined with GPT4, LLaVA makes 92.53% of success, which is a new benchmark for Science QA challenges. Such results demonstrate that the models tuned to instruction can achieve high-performance rates in the visual reasoning tasks if it is used for following instructions.

A research [26] on medical adaptation of Large Language Models and Vision-Language Models tries to ascertain their current achievements. The research analyzes the impact of techniques of domain-adaptive pre-training upon general large language models and vision-language models in case of their use for medical purposes. Medically adapted models, which currently exist, do not showcase a better level of performance at the baseline general-domain level than what they are in medical question-answering tasks that require zero to a few shots. The paper provides through careful statistical explanation and structured experiments that the current advantages of DAPT approaches are inconsequential as the common general-domain models are nowadays showing immense medical reasoning capabilities. The integrated assessment requires improved medical model adoptions along with the direct evaluation techniques and best prompting systems to achieve valid medical pretrained benefits.

Gemma appears to be an open-source library of models that comes from Google's Gemini technology to provide the best performance in various hardware platforms [29]. The 2 and 7 billion parameter models perform at the tier of the best for numerous language processing and reasoning functions, and outperforms models of the same size or slightly more in over 90% of benchmark tests. Gemma utilizes multi-query attention with rotary positional embeddings and GeGLU activations in order to improve its mathematical, scientific, code completion and logical reasoning properties. Gemma is pre-trained and tuned according to human feedback reinforcement learning in order to exhibit human preferred alignment performance that yields better safety and helpfulness performance. The model is validated by evaluation results that show that Gemma is 61.2% better than Mistral according to human judges in instruction-following situation tests thus showing its practical viability for safe deployment.

According to a paper [16], there is a description of a transformer-based encoder-decoder system for medical visual question answering which belongs to "Vision Language Model for Visual Question Answering in Medical Imagery". The model incorporates ViT as its visual processing unit for close medical images, whereby it uses a textual transformer encoder that operates for question embedding. The system generates reliable text answers through the process of auto-decoding after it combines multimodal representations that were sequentially concatenated. The PathVQA validation test along with VQA-RAD validates the fact that the model yields excellent scores while answering both open-ended and closed-ended questions correctly. There is a well-defined matching mechanism at this model because the BLEU score evaluation metrics are utilised to gauge how accurate the prediction is against truth for improved clinical diagnostic outcomes.

The explanations of Gemma 3 by the authors [41] are described through "Gemma

3 Technical Report" as its extended visual and multilingual model of the Gemma series that employs the extended 128K token context while maintaining multiple programmable scales from 1 to 27 billion parameters. Based on the authors' suggestion, the memory requirements on extended context are reduced using alternating local and global attention layers. Gemma 3 employs SigLIP vision encoder to do flexible high resolution image understanding by its adaptive Pan and Scan. The post-training methods along with the distillation methods help augment the mathematical and multilingual and the code-writing and instruction-following skills of Gemma 3 considerably. The Gemma 3-27B-IT achieves performance levels similar to that of the Gemini-1.5-Pro at a time when there are significant improvements made in chat abilities and reasoning from where it was in the Gemma 2.

This paper [32] provides an extended review of the shift between the discriminative to generative models in the field of Medical Visual Question Answering (MedVQA). It shows the development of MedVQA systems whereby generative models that are being driven by large language models (LLMs) and multimodal large language models (MLLMs) are changing the capacity to manage complex unanswered clinical questions. The survey talks about different generative architecture, such as encoder-decoder structure and attention models, and sophisticated pre training approaches, including vision language pretraining (VLP), and instruction tuning. The incorporation has made MedVQA perform better in contextually rich and clinically relevant responses, surpassing the discriminative methods when dealing with complex medical questions. Besides, the paper discusses the difficulties caused by data scarcity and model hallucinations and the emerging techniques in form of data augmentation and parameter efficient fine tuning (PEFT) of dealing with them. Also, by looking at generative vs discriminative models, it highlights the role of multimodal fusion techniques and future directions that will let you enhance the clinical reasoning, interpretability, and scalability of the real-world healthcare application.

Visual Large Language Models (VLLMs) are covered in the survey [36] which combine techniques from both vision and language to manage a broad range of tasks. It describes how VLLMs have changed from being simple vision-language models to ones that now include improvement through instruction and advanced logical thinking abilities. Vision applications are grouped as vision-to-text (for tasks such as image captioning and visual question answering), vision-to-action (for things like autonomous driving and robotics) and text-to-vision (included in image/video production, 3D modeling). Key datasets mentioned in the paper are COCO, LAION, Med-VQA and ScienceQA which are used for training and evaluation. Although VLLMs excel in a variety of tasks, they have difficulties in being computationally efficient, understandable, free from bias and with spatial reasoning. Efforts will be made in the future to reduce the number of tokens, create parameters in relaxing models, improve explainability factors, build privacy and security and evolve reasoning. The work also includes thinking about ethics, focusing on labor concerns and bias. All in all, the paper gives a detailed summary of the present status, drawbacks and possible new areas to study for VLLMs in general as well as for different fields.

In the paper [42] a wide survey about the topic of Large Language Models (LLMs) applied to medical images analysis is presented, with a deep research about their

taxonomy, their application, their trends for the future. It provides a new "x-stage tuning" paradigm that consists of zero-stage, one-stage and multi-stage tuning methods for the fine-tuning of LLMs on medical image assignments. Zero-stage tuning uses pre-trained models without fine-tuning, which solves the problem of the lack of data, but is not necessarily ideal. One-stage fine-tuning of models tunes models on domain-specific medical image data, enhancing the performance of tasks. This is furthered by multi-stage tuning which includes more domain-agnostic data across stages, increasing model flexibility and effectiveness. The challenges and the future research directions concerning the challenges in this field addressed in the survey involve the issues of the datasets, multilingual models, knowledge updating and model generalization. Eventually, the paper seeks to give a clear picture and further progress in the application of LLMs for the analysis of medical images.

In this paper [38], an in-depth review is presented about the evolution of Visual Question Answering (VQA) systems inclusive of the history of the field and advancements leading to the modern technologies. The paper starts by touching the beginning methods of image-captioning and continues to the more sophisticated, the state-of-the-art models that incorporate the understanding of vision and language. The authors study different architectures that have been used in the VQA, giving a prioritization to the shift from old-fashioned techniques (on the basis of convolutional neural networks (CNNs)) to contemporary solutions in which transformers and attention mechanisms are used. These developments have gone a long way in enhancing the accuracy and robustness of VQA systems, especially when it comes to treating open-ended questions about images. The survey also touches upon the difficulties that the VQA community experiences, such as the problem of the quality of the dataset, the problem of generalization of a model, and the problem of interpretability of the VQA predictions. The paper presents a deep exposition of the significance of multimodal learning, clarifying the role of joint vision-language models capable of analysing and understanding concurrent visuals and texts. The authors also set out research directions for future works towards developing VQA systems to generalize to new, unseen data, and, at the same time, be able to reason about complex multi-step problems in the visual domain.

This paper [39] explores the zero-shot visual question answering (VQA) use in multimodal large language models (MLLMs) to interpret 12-lead ECG images. The assessment of how effectively these models may work on performing medical image interpretation tasks without further fine-tuning alone based on the zero-shot learning capabilities is also conducted through the study. It also evaluates the shortcomings and strengths of the MLLM approach on complex medical images data. The paper proposes a zero-shot evaluation framework, according to which the MLLMs process ECG images and respond to the questions related to their interpretation of the visual and textual material. Although the models demonstrate a lot of potential, the study notices that there is still a lot of performance difference between MLLMs and the human experts, and the models have to put a lot of effort into the ability to correctly interpret some details found on the ECG images. Based on the evaluation results, it may be concluded that even though zero-shot models can perform a basic set of tasks, they tend to give unsuitable interpretations of fine medical properties that require additional training. The study ends up by highlighting a need

for domain-specific fine-tuning to enhance medical reasoning in MLLMs, especially for situations where highly technical and subtle data like ECG interpretation is involved. The authors also propose that the next tasks should be aimed at improving the models' performance in processing complex clinical data and with the help of continuous learning being able to resolve new challenges arising in the quickly developing field of medicine imaging.

In this paper [40], a benchmark called MTVQA is presented, and it is oriented toward the multilingual text-centric visual question answering (TEC-VQA). The benchmark is the first one that offers high-quality annotated data by humans in nine languages, such as Arabic, Korean, Japanese, Thai, Vietnamese, Russian, French, German, and Italian. MTVQA deals with images exhibiting various visual-textual elements that are documents, menus, signage, and maps, and seeks to solve the visual-textual misalignment problem that emerges when machine translation is implemented in multilingual VQA tasks. The challenges of the existing models in the multilingual text-centric scenes, particularly with visual text in images are emphasized by the authors. They also analyse various state-of-the-art multimodal large language models (MLLMs) on the MTVQA dataset namely Qwen2-VL, GPT-4o, GPT-4V and Claude3, despite significant improvements, there is still a wide scope of improvement.The paper concludes with the necessity to carry out further research in mult The authors wish the dataset to be a useful instrument for further studies in this sphere.

The paper [35] presents Med-R1, a reinforcement learning (RL)-based model aiming at improving the generalizability of medical reasoning in vision-language models (VLMs). The overall purpose of Med-R1, in turn, is the addressing of challenges associated with making a decision in medical AI, primarily in areas where visual thinking and medical knowledge are critical for accurate diagnosis and decision-making. Med-R1 draws on reinforcement learning for iterative enhancement of its reasoning skills, so that the model can generalize across multiple medical projects: image diagnosis, medical report creation, or clinical decision support. Using the system to train the model to learn from its engagement with medical data and with feedback received from medical experts, the system can formulate a significant number of medical scenarios. The paper highlights the potential of RL in enhancing contextual knowledge and enhancing the accuracy of reasoning as compared to the traditional methods of solving and presents a performance comparison of Med-R1 to conventional methods to demonstrate superiority of Med-R1 in handling complex medical tasks. Finally, the study concludes with the future issues and how RL can further bring forth advancement in medical AI through enhancing model understandability and practicality along the clinical field.

Lesion-Aware Transformer (LAT) model for scoring the severity of skin diseases put forward in this paper [33] combines principles of clinical decision making with transformer based neural networks. LAT model aims at using lesion segmentation, along with feature extraction for increasing the accuracy in severity assessments for skin diseases. It uses dermatology datasets like HAM10000 and ISIC for training and testing the model, reporting 92.5% accuracy in classifying skin lesions. Limitations include difficulties in generalizing to scarce or not common skin diseases and

the difficulty of promoting model robustness for a set of diverse patients. The paper proposes integrating multimodal data (patients' history and clinical symptoms, for instance) to improve the reliability of the model. The future work includes the enhancements in generalization, creation of various datasets, explanation of model usage for clinical adoption.

In the paper [37] the authors present a framework of MedVLM-R1, an RL-based model, which is supposed to improve the medical reasoning of the VLMs. The RL is incorporated into the methodology to advance the capacity of a given model to perform various medical tasks including the diagnosis of diseases, generation of medical reports and captioning of images. It makes use of medical datasets such as ChestX-ray, NIH Chest X-ray, and MIMIC-CXR in training and evaluating the model. In comparison with the traditional methods, MedVLM-R1 is more accurate and general in medical reasoning tasks.Limitations include troubles with sparsity of data as well as the absence of clinical context in the training data that impair the performance on complex medical questions. The paper proposes data augmentation, and involvement of real-world clinical data, reducing model adaptability and improving its accuracy. Further research should also focus on improving explainability, interpretability and considering problems of data bias in medical AI systems, as well as improving the reasoning ability of the system in various medical contexts.

This review [31] discusses the exploitation of Visual Question Answering (VQA) in robotic surgery, with the emphasis on the ways in which the VQA systems can enhance real-time decision-making in surgery through processing the visual data (e.g., surgical images or videos) together with natural language input. Some of the techniques discussed in the paper include CNNs, RNNs, and transformers for processing images and understanding languages in terms of robotic surgeries. The review highlights major bottlenecks in robotic surgery systems, from issues of accuracy, real-time processing, lack of massive data for training. Some solutions are data augmentation, creation of domain-specific datasets, and enhancement of instantaneous feedback systems. Future research should consider multimodal learning, instant decision making and better interpretability of models that will increase transparency and reliability of high-risk surgical operations.

This paper [27] explains how to set up effective in-context sequences for the case of Visual Question Answering (VQA) using Large Vision-Language Models (LVLMs), demonstrating the necessity of emphasizing the role of ICL for the purpose of improving performance in VQA. The authors study diverse demonstration configuration strategies, including retrieval of related images, questions, and answers, so as to produce optimal in-context sequences for VQA tasks. The study employs Open-Flamingo models and compares performance of the models on popular VQA datasets: VQAv2, VizWiz, and OK-VQA. The implications show that Task Recognition (TR), which aims for recognizing the format of the task, input distribution, and label space is much more important than its opposite number – Task Learning (TL) that deals with the learning of the relationship between inputs and outputs. Three main properties of LVLMs are unraveled by the study. poor TL capabilities, shortcut effect, as well as the partial compatibility of the vision and language modules. Notably, the authors note that it is possible to randomize triplets (im-

ages, questions, answers) does not lead to the significant reduction of performance, which emphasizes the leading role of TR. According to the results, the 4-shot and 8-shot configuration provide the best trade-off between accuracy and efficiency, with Open-Flamingo v2 achieving 53.49% accuracy on OK-VQA using 8-shot demonstrations. The research shows that the demonstration configuration techniques can always enhance VQA accuracy, such as retrieving similar question-answer pairs and manipulating the sequence order. Limitations are the requirements of better understanding of the internal mechanisms of LVLMs, especially their level of sensitivity to the order and presentation arrangement of demonstrations. According to the paper, future work should be dedicated to improving demonstration retrieval techniques, adding the instruction-based prompts to enhance performance, and perfecting the configuration strategies. Also, extension of research into multi-modal systems and enhancement of the generalizability of models to cover VQA and multimodal learning are the future aspects to focus on.

In this paper [43], there is a suggestion of a fine-grained adaptive visual prompt that is used in the improvement of generative models for medical visual question answering (VQA) tasks. Using dynamic adaptive visual prompts that vary according to the contents of an image, the model comes up with more accurate and relevant answers to medical questions. The approach uses MedVQA, ChestX-ray, and VQA-Med datasets for the evaluation of the model performance. The Results give an F1-score of 90.8%, outperforming the traditional ones in terms of accuracy and relevance of the answers. The limitations of the paper include the cases of handling rare or complex medical conditions and problems with reliability of models only upon visual features. Possible ways out would be the use of synthetic data and data augmentation for rare conditions, as well as pre-training on more diverse sets of medical data. Future research should be aimed at the enhancement of the model's capacity to process unseen/novel medical conditions and customisation of balance between precision and recall for medical answers' generation.

Summary of the findings that we reviewed are mentioned in the Table 2.1. It summarized the models used in each paper, the datasets used, and the accuracy of the models.

Ref	Model	Dataset	Accuracy
[29]	Gemma Open Models: Lightweight transformers + SigLIP vision encoder	General NLP, vision-language	MMLU 7B 5-shot: 64.3%, strong reasoning & coding
[37]	MedVLM-R1: RL-based model to improve reasoning	MRI, CT, X-ray VQA datasets	Accuracy improved from 55.11% to 78.22% with RL
[4]	Image encoder: VGG-16, ResNet-50/152; Text encoder: BERT, BioBERT, LSTM baseline	VQA-Med 2019	60% accuracy
[12]	VQGR, T-5 based Model	MedPix (VQA), VQA-RAD, SQuAD, MACRO	38.2% accuracy (VQA), BLEU 41.6%, VQG BLEU 38.3%
[23]	MM-BERT with ResNet-152 image extractor	Various medical datasets	Modality 83%, Plane 86.4%, Organ 76.8%
[5]	DAE, MAML, CDAE	VQA-RAD	16.3% open questions; +8.6% descriptive
[33]	Lesion-Aware Transformer (LAT)	HAM10000, ISIC datasets	92.5% lesion severity accuracy
[8]	Transformer-based; ResNet-50 (image), BioELMo (text)	PathVQA	64.82% success rate
[24]	Vision-language model: CLIP (image), BERT-base (text)	VQA-RAD, PathVQA	BLEU 75.41% (VQA-RAD), 67.05% (PathVQA)
[16]	Transformer encoder-decoder w/ ViT (image), BERT-like text encoder	VQA-RAD, PathVQA	77.27% (VQA-RAD), 70.54% (PathVQA)
[40]	MTVQA: Multilingual text-centric VQA benchmark	9 languages, 6,778 images	Best MLLM 32.2%; human 79.7%
[14]	Unified VLP with BERT backbone	Conceptual Captions, Flickr30k, VQA 2.0, COCO Captions	10% improvement on CIDEr, Flickr30k
[12]	SOTA / Pythia base model	Custom dataset	7% improved consistency
[27]	In-Context Learning for LVLMs in VQA	VQAv2, VizWiz, OK-VQA	Open-Flamingo v2 8-shot: 53.49% accuracy
[28]	Pre-trained CLIP fine-tuned with PubMed articles	VQA-RAD, SLAKE	3% QCR, +6% MEVF
[25]	MedBLIP: BioMedLM + LoRA finetuning + lightweight encoder	ADNI, NACC, OASIS, AIBL	78.7% (ADNI), outperforms ViT-G 72.2%
[7]	Faster R-CNN w/ ResNet101 (image), GloVe (text), adaptive fusion	HAM10000	98.28% accuracy

Ref	Model	Dataset	Accuracy
[13]	MedFuseNet: CNN (image), BERT/XLNet (text), MCB fusion, LSTM decoder	MED-VQA 2019, PathVQA	Classification 84%, Yes/No 63.6%, Answer generation 60.5%
[22]	LLaVA (CLIP vision encoder + Vicuna LLM)	Multimodal instruction tasks	85.1% GPT-4 relative on instructions; 92.53% Science QA
[9]	Various CNNs (VGG16, ResNet-152), LSTM, BERT, attention	PathVQA, VQA-Med	MedFuseNet best: 93.8%
[21]	Image: VGGNet, ResNet; Text: BERT (Transformer)	VQA-Med 2019	62.4% accuracy
[16]	Transformer encoder-decoder (ViT + text encoder)	VQA-RAD, PathVQA	Closed: 84.99% VQA-RAD, 83.86% PathVQA; Open: 72.97%, 62.37%
[35]	Med-R1: Reinforcement Learning for reasoning	ChestX-ray, NIH Chest X-ray, MIMIC-CXR	69.91% avg accuracy; +29.94% over base Qwen2-VL-2B
[11]	ResNet-152 + multi-scale features; text: skip-thought + GRU	VQA dataset (123k images)	70.73% accuracy
[10]	Transformer w/ SAU + modified CNN	ISIC-2019, PH2	97.48% and 96.87% accuracy
[30]	Qwen2.5-VL: Dynamic-resolution ViT + window attention	MMMU, MathVista, MMBench-EN, VQA	MMMUval 70.2%, MathVista 74.8%, MMBench-EN 88.6%
[18]	Stacked Attention Network (SAN) with LSTM dialog history	RadVisDial (chest X-ray visual dialog)	Macro F1 score: 0.243 (early baseline)

Table 2.1: Summary of Related Works

2.1 Research Gap

Despite significant advancements in both VQA and medical image analysis, several research gaps remain unaddressed. This limits the practical application of VQA systems in the medical field. Our proposed model pipeline aims to bring an improvement on some of these lackings such as-

Unavailability of data : In the VQA sector, properly annotated data is scarce. More so in the case of medical data for skin disease VQA tasks. Medical VQA requires both images and properly curated question answer pairs for each image. This can be a problem in the medical domain as getting patients' images and data to create a proper dataset can be problematic.

Resource heavy models : In general, VQA models are made with the intent of using it for general usage. For getting a higher accuracy for a broad range of tasks, VQA models have hundreds of millions even billions of parameters. This makes them impossible to use without sophisticated gpu's. This becomes a problem when we need to use these models locally in handheld devices. These devices need the internet and only act as an interface between the model running somewhere else and the user. Our aim is to make a lightweight, specialized model that can be operated on a wide range of devices without needing huge hardware resources.

Modular workflow : In typical VQA models, the architecture makes it so that any change in the image or text can cause the model to be trained entirely from scratch for both image and text part. This can be a problem with models having heavy resource requirements. There is a lack of research on implementing modular training methods where image and text training can be separate. Such a modular approach can greatly reduce hardware requirements for updating existing models with new data.

2.2 Chapter Summary

In this chapter, we mainly focused on reviewing and investigating current VQA models and their applications in the medical field along with issues such as limited annotated medical datasets and high computational expenses. The fusion of CNNs and RNNs for image and text processing in VQA systems is discussed together with several vision-language models (VLMs). The review also emphasizes the need of lightweight, domain-specific VQA models in dermatology and points out areas of lacking knowledge in current research that our thesis intends to fill in.

Chapter 3

Dataset

3.1 Description of the Custom Annotated Dataset

Existing Visual Question Answering (VQA) datasets such as VQA-RAD and Path-VQA are primarily designed for radiology and pathology images, focusing on internal organ diagnostics and medical imaging reports. While these datasets have advanced medical VQA research, they do not include dermatological images or skin-specific visual features. On the other hand, existing skin disease datasets like ISIC provide labeled images for classification or segmentation but lack question-answer pairs needed for VQA tasks. They also do not include detailed textual annotations covering attributes like color, border, cause, and severity. Therefore, there is a clear gap in VQA datasets tailored for skin disease analysis. Our proposed dataset addresses this gap by combining high quality skin images with structured clinical question answer pairs, enabling the development and evaluation of interpretable, domain-specific VQA models for dermatology.

Our custom annotated dataset is tailored specifically for the classification and question answering tasks related to skin diseases. The images in the main dataset were obtained at the International Skin Imaging Collaboration (ISIC) repository that offers an extensive amount of high-resolution pictures of dermatology. Also, we collected images from the Skin-Disease-Dataset compiled by Subir Biswas (2023), hosted on Kaggle [17]. From these sources, we selected a total of 1,038 images, which were then meticulously annotated by a certified medical professional to ensure the accuracy and clinical relevance. The dataset is made with a folder containing all the unique skin disease images. In addition, information such as the image file name, question, answer, image class, and question type are stored in the json format in separate file. Our dataset does not contain any private or personal information about the patients and ensures privacy.

The dataset covers 11 distinct skin disease classes, covering a wide spectrum of dermatological conditions with varying severity and medical implications. These classes include:

1. Melanoma
2. Chicken Pox
3. Impetigo
4. Nail Fungus (Onychomycosis)
5. Ringworm
6. Shingles
7. Vascular Lesion
8. Melanocytic Nevus
9. Benign Keratosis
10. Squamous Cell Carcinoma
11. Basal Cell Carcinoma

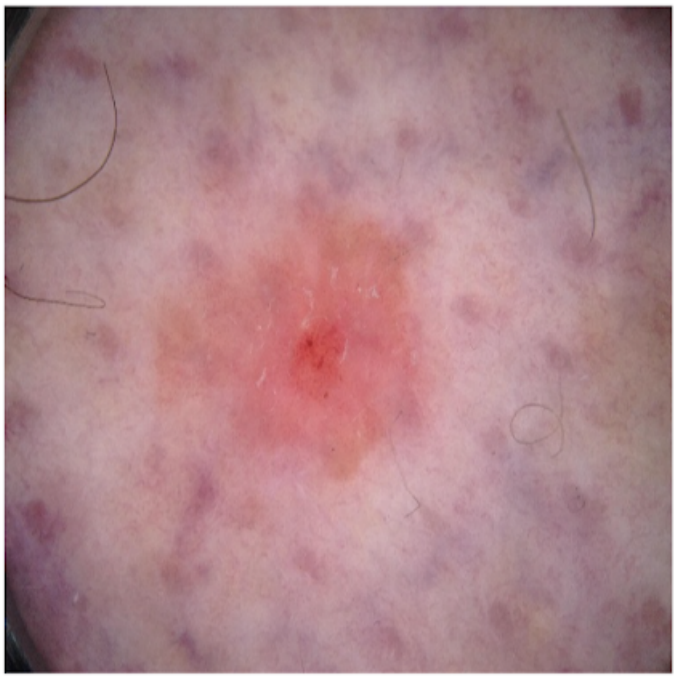
Each image in the dataset is associated with seven structured question-answer pairs, designed to simulate real diagnostic inquiries that could be posed by either patients or medical practitioners. Each question covers a distinct attribute about the skin disease image. Example questions from the dataset:

1. What is the name of the disease?
2. Is it cancerous or non-cancerous?
3. What are the visible features in this image which indicate the disease?
4. What is the severity of this skin condition?
5. What are the preventive measures for this disease?
6. What diagnostic tests are needed?
7. What is the cause of this disease?

Sample from the custom dataset:

Question: What diagnostic tests are recommended?
Answer: Skin biopsy and histopathological examination.

Figure 3.1: Sample-1 from our custom dataset



Question: What is the cause of this disease?

Answer: Melanoma is primarily caused by damage to DNA in skin cells due to UV radiation exposure

Figure 3.2: Sample-2 from our custom dataset


```
{
    "qid": "melanoma4",
    "image_name": "ISIC_7623407.jpg",
    "question": "What are the preventive measures for this disease?",
    "answer": "Use sunscreen, avoid excessive UV exposure",
    "question_type": "open",
    "class": "melanoma",
    "image_source": "https://www.isic-archive.com/"
},
```


Figure 3.3: json format sample from our custom dataset

From the given Figure 3.1 and Figure 3.2 we can see sample images paired with question-answer pairs. Also in Figure 3.3 we see the json format used to store information including question-id, image name, question, answer, question type, class and image source. The questions are made in context of the image and the answer is based on the context of both image and the question. This multi-question annotation format allows the dataset to serve not only as a classification benchmark but also as a foundation for training models in visual question answering (VQA), where the model must interpret the image and respond to a variety of medically relevant queries. The inclusion of both objective and subjective questions enhances the model's reasoning capability, pushing it to infer, classify, and recommend based on visual and contextual cues. The comprehensive annotation and domain specificity of this dataset make it a critical component of our research, supporting both training

and evaluation of the proposed VQA model pipeline in the context of skin disease analysis.

3.2 Dataset Preprocessing

At first, we make sure that there are no duplicate images or missing images. Next, we check whether any images are missing from the json file. Afterwards, the dataset is divided into three subsets to facilitate training, testing, and validation in a supervised learning setup. Specifically, 833 images were allocated for training, 103 for testing, and 102 for validation. Hence, about 80% of the dataset is kept for training, 10% for validation during training, and 10% for testing accuracy after training. Also, it is made sure that there is an equal percentage of images per class in each split. In addition we also made sure to maintain all seven question-answer pairs per image in the splits.

	No. of Images	**No. of Q/A pair**
Train	833	5831
Test	103	721
Validation	102	714
Total	**1038**	**7266**

Table 3.1: Train, Test and Validation set distributions of the dataset

3.3 Description of the Benchmark Dataset

To evaluate our proposed VQA model, the Medical Images with Q&A: A Multi-Domain Dataset [15] was used as benchmark dataset. The dataset has images in various fields related to medicine such as radiology, pathology, dermatology and ophthalmology. In the case of skin diseases it is a sub dataset, namely with 737 images and 1,237 pairings of questions and answers, with respect to the large cross-domain dataset. Although it offers an effective benchmark, the diversity of questions is a little lacking in the data set, as there are not many variations in the nature of the questions pertaining to dermatology.

For example, the dataset includes question-answers like-

Q1: "Is it benign or malignant imaged lesion?"
Response: "The imaged lesion is non cancerous."

Q2: "What is the diagnosis of skin condition in the picture?"
Response: "The skin condition in the picture is a melanoma."

This dataset covers multiple domains in medical field, which is why the annotations and question sets it contains are not as specialized toward dermatology compared to our custom annotated dataset. Also, we were required to encode the benchmark dataset to a format that is compatible with our proposed model. This encoding

procedure enabled us to test the model across more questions of a medical nature, but the diversity of questions were not as comprehensive compared to our custom dataset.

3.4 Comparison of VQA Datasets

In Table 3.2 we can see the comparison of different relevant VQA datasets. There are decent sized VQA datasets like PathVQA and VQA-RAD for Pathology and Radiology. However, the Dermatology dataset lack versatility in terms of question-answer pairs per image. Our dataset having seven question-answer pairs per image ensures more information extraction from the images. Also, this will allow the models to better generalize and understand the context of each images on which they are trained with.

Corpus	**Images**	**QA**	**Image Type**	**Query Types**
PathVQA	4,289	32,632	Pathology	open-ended
VQA-RAD	315	3,515	Radiology	open-ended
DermaVQA	3,434	1,488	Dermatology	open-ended
Medical Images with Q&A: A Multi-Domain Dataset	749	1,818	Dermatology	open-ended
Our Dataset	1,038	7,266	Dermatology	open-ended, close-ended

Table 3.2: Comparison of existing VQA datasets with our custom dataset

3.5 Chapter Summary

With 1,038 annotated skin disease images spanning 11 disease classes, this chapter offers a thorough overview of the custom dataset produced for our research. Focusing on clinical elements including disease identification, severity, cause, and treatment choices, each image is matched with seven question-answer sets. Publicly available images were collected and then annotated by an experienced medical expert. Diversity in disease types and severity was ensured to make the dataset more versatile. Stratified sampling is used for balanced representation across classes, the dataset is split out into training (833 images), testing (103 images), and validation (102 images) in the ratio 80:10:10.

Chapter 4

Methodology

4.1 Workplan

This chapter shows the overall workflow that was followed. The work plan of this research was divided into several structured phases. Each phase contributed to the successful implementation of the overall objective. The implementation steps followed a logical order, beginning with dataset preparation and ending with evaluation of model performance.

1. Dataset Construction

The first step involved the creation of a domain-specific skin disease dataset suitable for Visual Question Answering (VQA). A total of 1,038 skin disease images were collected from the ISIC archive. These images covered 11 skin disease classes, including common and clinically relevant conditions. Each image was annotated with 7 question-answer pairs, focusing on various visual and diagnostic features like disease name, shape, color, border, texture, cause, and severity. Annotation was performed by a medical professional to ensure reliability and accuracy.

2. Fine-tuning Vision-Language Models

We selected four pre-trained vision-language models: Gemma 3, QwenVL-2.5, BLIP-2, and LLaVA-1.5. Each model was fine-tuned using our dataset to align their understanding with the skin disease domain. The models were trained using standard training loops with loss tracking and checkpoint saving. For each model, we ensured consistent input formatting and question types to allow fair comparison.

3. Model Evaluation Using Semantic Metrics

After fine-tuning, we evaluated the performance of each model. Instead of exact answer matching, we used semantic similarity metrics to compare model responses with the ground truth answers. We used BERTScore and Sentence-BERT score to measure how close the generated answers were to the reference answers in meaning. This method gave a more accurate evaluation of open-ended VQA tasks, especially in medical contexts.

4. Development of Lightweight Alternate Model

In parallel, we developed an alternate lightweight model to support efficient VQA. A Vision Transformer (ViT) was trained to encode image attributes into a fixed structure. This encoding was built from the image's question-answer pairs and stored as attribute-value mappings. For language processing, we used a MiniLM model to interpret the user's input question. The system calculates the semantic similarity between the input question and predefined attribute keywords. The most similar keyword is selected, and its corresponding value from the image encoding is returned as the final answer.

5. Comparison and Analysis

Finally, we compared the performance of the fine-tuned vision-language models and the lightweight model. We analyzed their accuracy, semantic similarity, and efficiency. Also, we trained our lightweight model using a benchmark dataset. This comparison helped us understand the trade-off between performance and model complexity. Visualizations such as loss curves and score graphs were used for better interpretation of results.

4.2 Finetuning existing VLM's

We have finetuned four state-of-the-art vision language models using our skin disease VQA dataset. These models are Gemma-3, QwenVL-2.5, LLaVA-1.5, and BLIP-2. The original size of these model are too large to fit in our consumer grade GPU's. Hence, we had to use techniques like quantization and parameter-efficient-finetuning in order to finetune the models with low resources. We have used 4 bit quantization along with Low-Rank Adaptation (LORA) to significantly reduce the required VRAM for training. Also, we used the standard hyperparameters that are used to finetuning these models. The training procedure involved loading the model weights from huggingface and then finetuning it using our custom VQA dataset. These models were trained on the RTX3060 12 GB graphics card.

Gemma-3 4B Model

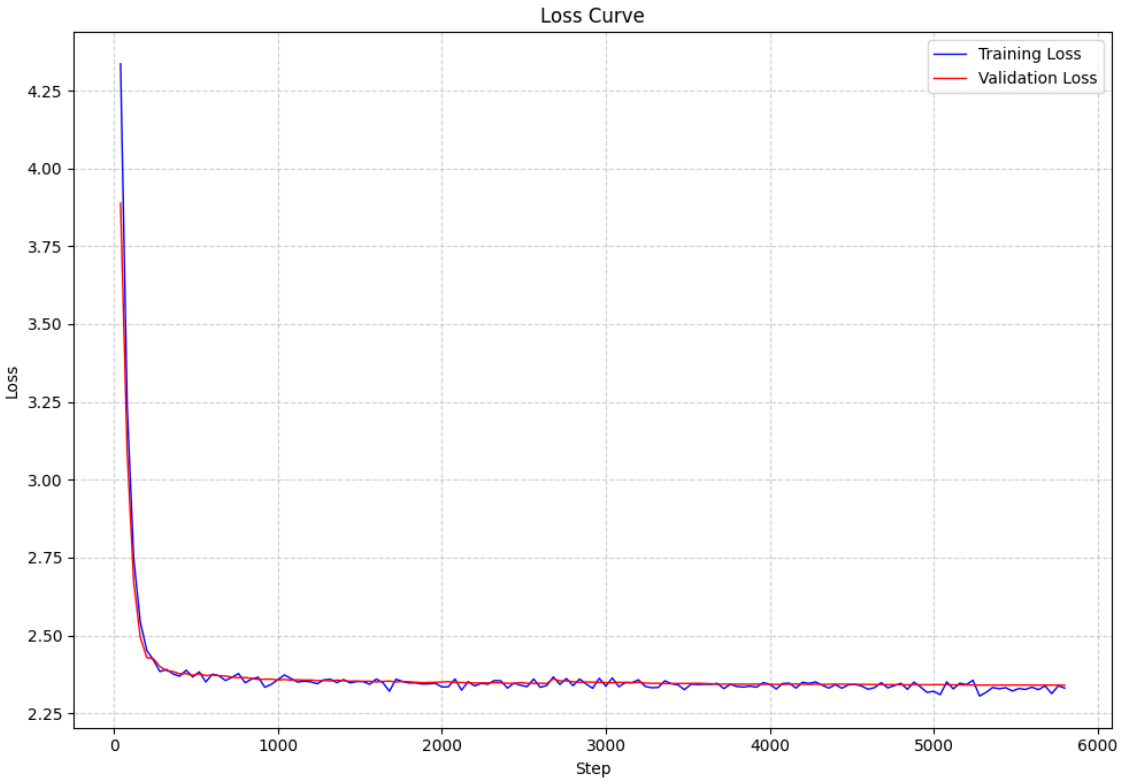


Figure 4.1: Gemma 3 Loss Curve.

Benchmark semantic similarity scores

- **Bert Score:** 93.52%
- **Sentence Bert Cosine Similarity:** 77.73%

The loss curve for the Gemma-3 4B model in Figure 4.1 demonstrates a rapid decline in both training and validation loss during the initial training steps, stabilizing after approximately 1000 steps. This indicates that the model quickly learned core patterns in the dataset and maintained stable performance without significant overfitting or underfitting, as the training and validation losses remained closely aligned throughout. The benchmark semantic similarity scores further validate the model's performance, with a high BERTScore of 93.52% and a Sentence-BERT cosine similarity of 77.73%. This indicates strong alignment between the model's predicted answers and the ground truth responses.

Qwen-2.5-VL 3B Model

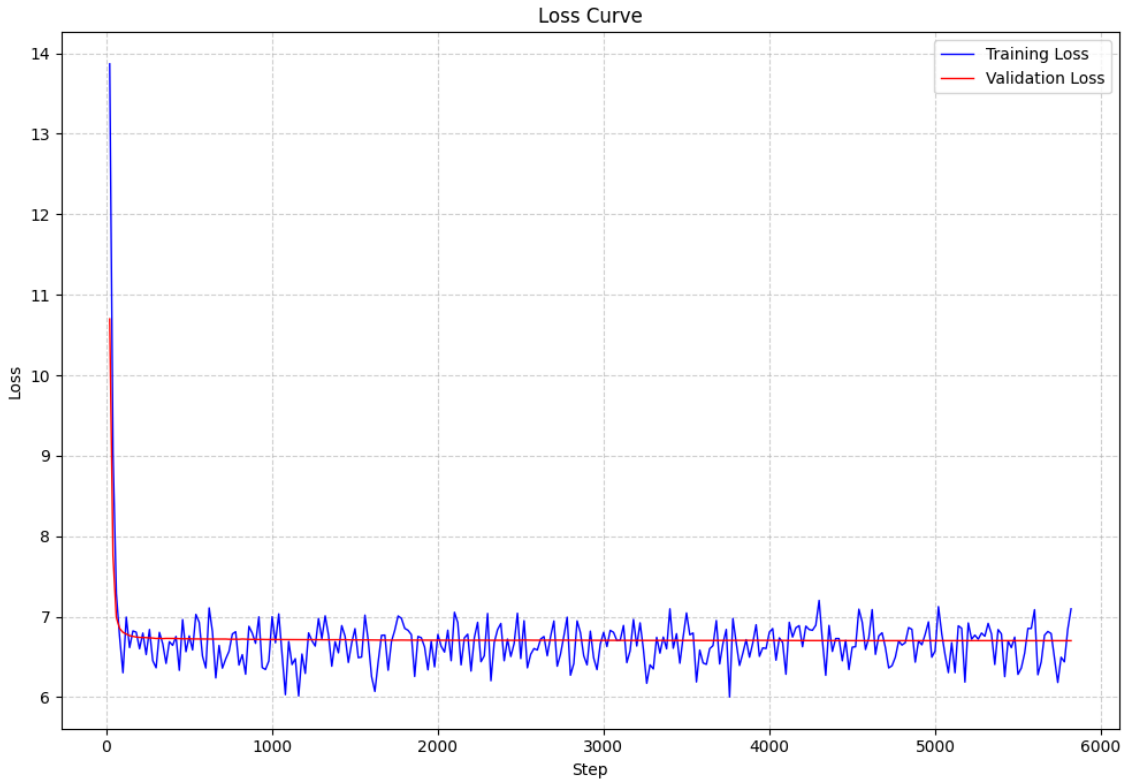


Figure 4.2: Qwen-2.5-VL Loss Curve.

Semantic similarity scores for Qwen-2.5-VL 3B Model

- **Bert Score:** 94.54%
- **Sentence Bert Cosine Similarity:** 81.21%

The loss curve for the Qwen-2.5-VL 3B model in Figure 4.2 shows a steep initial drop in both training and validation loss, followed by a relatively stable but noisy plateau around step 1000 onward. The considerable variation in training loss implies that there is a certain amount of fluctuation. performance instability or sensitivity in training but the validation loss is stable as well as is closely correlated with the training loss, so there is no significant overfitting. Despite the noise, the model gets good benchmark results with a BERTScore of 94.54% and a Sentence-BERT cosine similarity of 81.21 percent. This implies that it is efficient recaptures the semantic value of ground truth answers with a large degree of accuracy.

LLaVA-1.5 7B Model

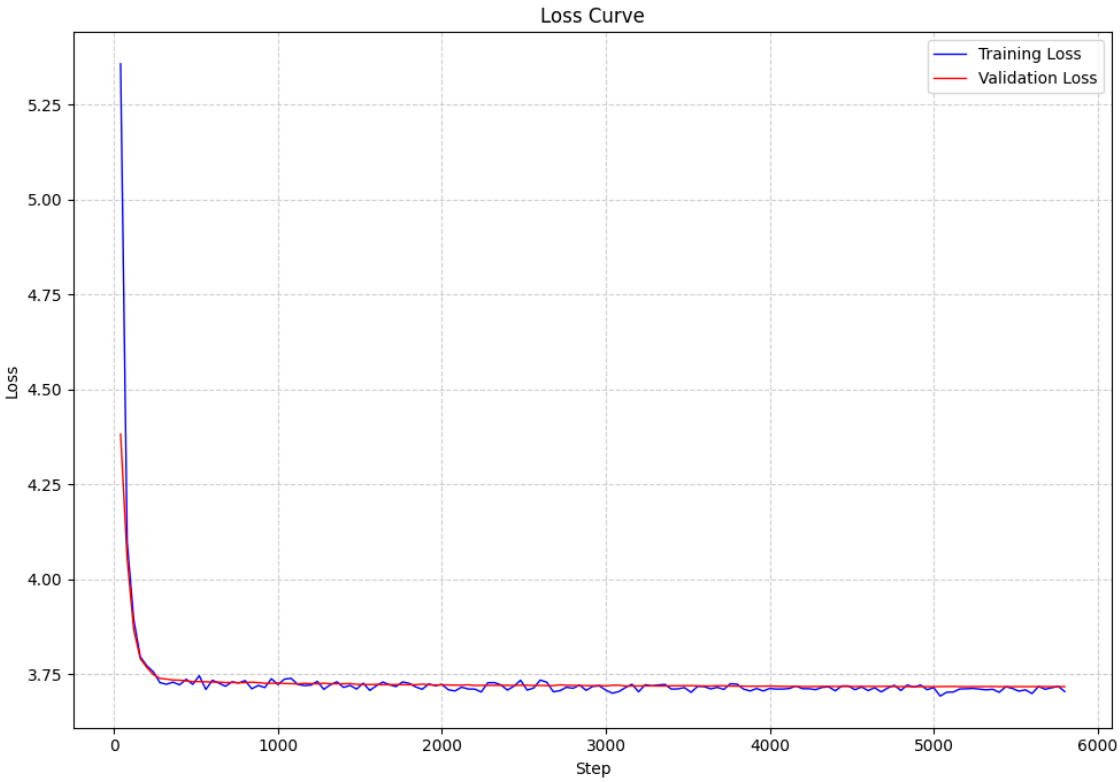


Figure 4.3: LLaVA-1.5 Loss Curve.

Semantic similarity scores for LLaVA-1.5 7B Model

- **Bert Score:** 93.78%
- **Sentence Bert Cosine Similarity:** 77.80%

In Figure 4.3, the LLaVA-1.5 7B loss curve of model i.e. there is a sharp decline in both there is an initial training and validation loss then it will stabilizes and converges after step 1000. The validation and the training loss are almost matched across the board and therefore representing a sufficiently generalized model which is not over-fitted. There is a high score on benchmark semantic similarity of 93.78% BERTScore and a Sentence-BERT cosine similarity of 77.80 percent. This implies that the model well preserves and recaptures semantic meaning of the ground truth answers in great fidelity.

Blip-2 2.7B Model

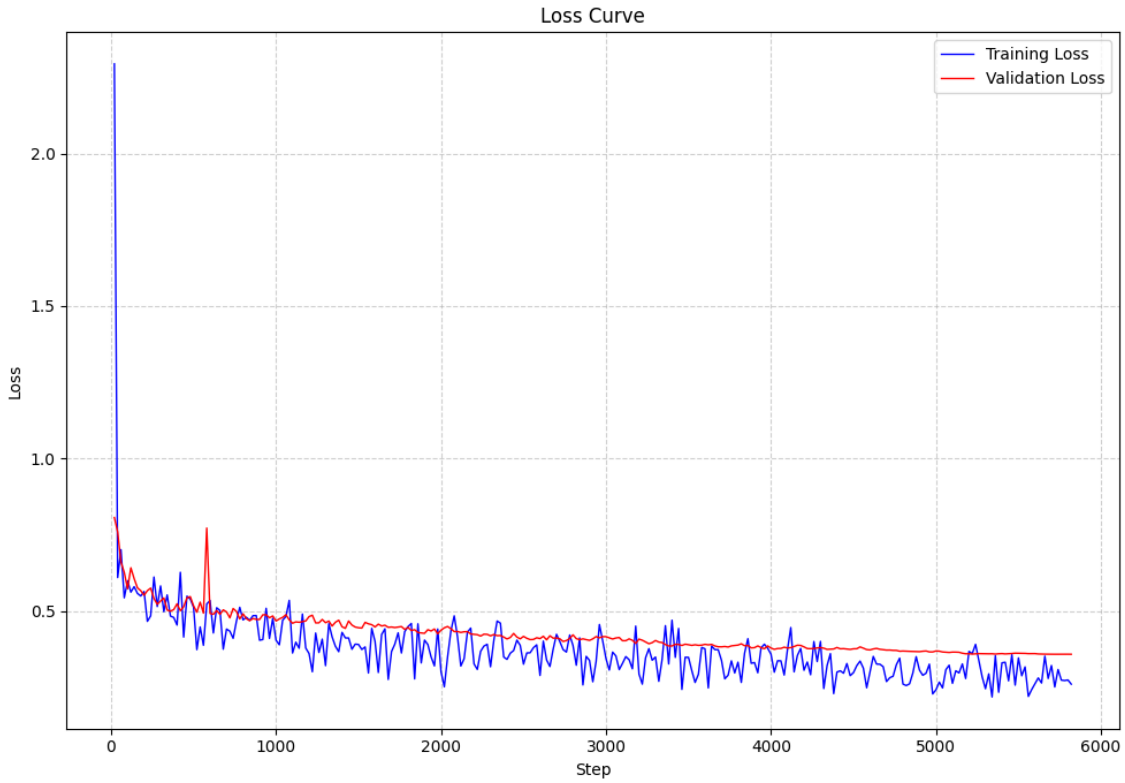


Figure 4.4: Blip-2 Loss Curve.

Semantic similarity scores for Blip-2 2.7B Model

- **Bert Score:** 94.47%
- **Sentence Bert Cosine Similarity:** 81.41%

In Figure 4.4, the Blip-2 2.7B This is evidenced in the loss curve of the model as the plot of the train drops top at a fast rate. Initial loss consisting of validation loss and then peaked to be stable yet very fluctuating behavior with time. Whereas the training loss is still steadily decreasing, the validation losses tend to stabilize at about a little above the training losses ranging good performance without much overfitting. The guideline semantic the high-similarity scores belong to all models, with the BERTScore of 94.47 percent and Sentence-BERT cosine similarity of 81.41 percent. This is an indication of the good model capacity to produce semantically correct responses much near to the base truth.

4.3 Hyperparameters used for VLM's

Hyperparameter	**Value**
LoRA Rank (r)	64
LoRA Alpha	64
LoRA Weight Initialization	Gaussian

Table 4.1: LoRA Hyperparameters Used in Fine-Tuning

Hyperparameter	**Value**
Number of Epochs	1
Train Batch Size per Device	1
Eval Batch Size per Device	1
Gradient Checkpointing	Enabled
Learning Rate	2×10^{-5}
Logging Steps	20
Evaluation Steps	20
Evaluation Strategy	Steps
Max Gradient Norm	1
Warmup Steps	0
Optimizer	Paged AdamW 32-bit

Table 4.2: Training Hyperparameters for Fine-Tuning

Here, the hyperparameter tunings are shown in Tables 4.1 and 4.2 define the fine-tuning configuration used for VLMs in this research study. Table 4.1 outlines the LoRA (Low-Rank Adaptation) parameters, where a rank of 64 and a scaling factor (alpha) of 64 were used, and weights were initialized using a Gaussian distribution. These settings control how the LoRA modules adapt pretrained model weights efficiently with fewer trainable parameters. Table 4.2 lists the training hyperparameters, which include one training epoch, a batch size of 1 for both training and evaluation, and enabled gradient checkpointing to reduce memory usage. The learning rate was set to 2×10^{-5}, with evaluations and logging performed every 20 steps. The model was trained using the "Paged AdamW 32-bit" optimizer, with gradient clipping applied at a maximum norm of 1 and no warmup steps. These configurations were chosen to balance performance with limited computational resources while ensuring stable optimization during fine-tuning.

4.4 Proposed Model

The core objective of our research was to develop a lightweight yet highly effective Visual Question Answering (VQA) model tailored for diagnosing skin diseases from images and accompanying symptom-based queries. While many state-of-the-art models in the field such as LLaVA-7B, BLIP-2, Qwen, and Gemma have demonstrated strong performance, they often come with massive parameter counts and significant computational demands. This makes them unsuitable for deployment in resource-constrained environments, such as rural clinics, mobile devices, or edge computing systems.

To address this limitation, we designed and implemented a compact VQA architecture that significantly reduces model size and computational requirements without sacrificing performance. Our model architecture consists of two key components:

1. Image Encoder

ViT

For the visual input, we utilized the base Vision Transformer (ViT) model which has 86 million parameters. This variant strikes a strong balance between representation power and model efficiency. Despite its lightweight nature compared to heavy ViT backbones or convolutional ensemble models, our chosen ViT encoder achieved high accuracy in learning discriminative features across the 11 disease classes.

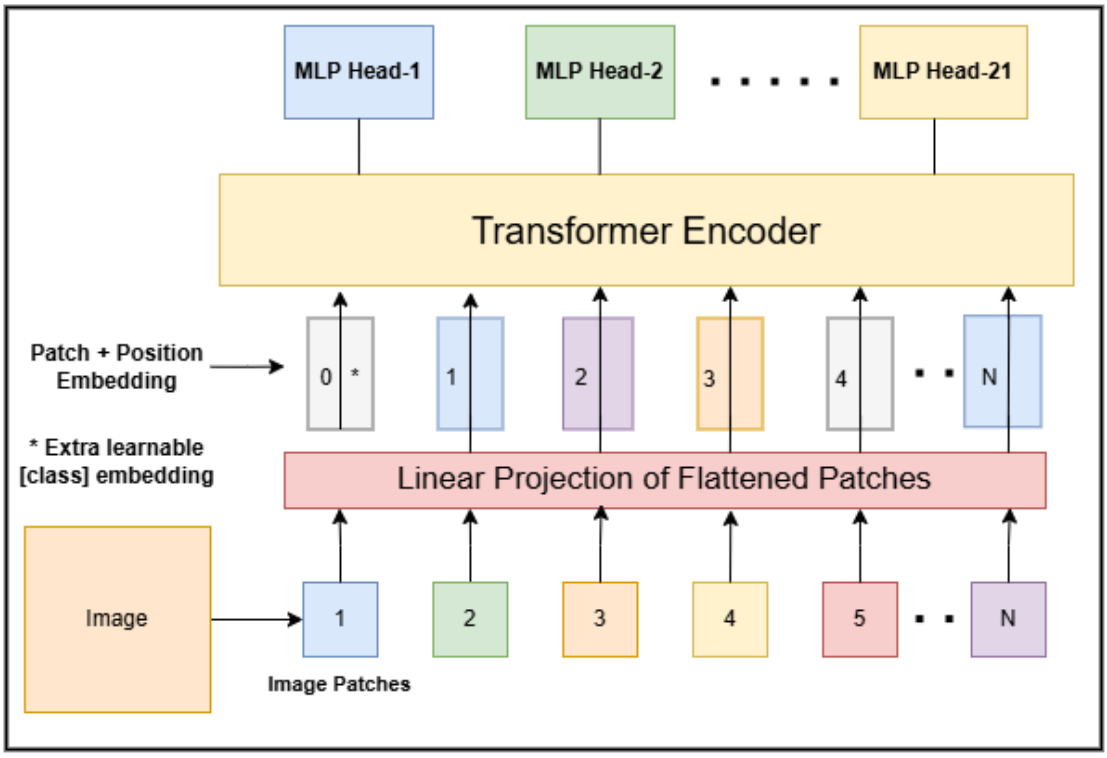


Figure 4.5: The Vision Transformer (ViT) architecture

The Vision Transformer (ViT) architecture, as shown in the Figure 4.5, divides images into fixed-size patches to process them. Every patch is flattened and linearly projected into embeddings; spatial information is retained using positional encodings. A standard Transformer encoder is fed this series of patch embeddings together with an additional learnable class token. Alternating layers of multi-head self-attention and MLP (multi-layer perceptron) blocks make up the encoder, which enables the model to detect local and global dependencies over the image. The class token's output at the last layer is then fed via a classification head (MLP) to forecast the class of the image.

Position embeddings are included to the patch embeddings in ViT to encode spatial connections between patches. ViT differs from CNNs in not depending on convolutional inductive biases like locality or translation equivariance. Rather, the model learns these interactions from the data. Particularly when pre-trained on vast datasets, the architecture helps ViT to perform competitive image classification, so surpassing conventional CNNs on tasks like ImageNet classification.

ResNet (Residual network)

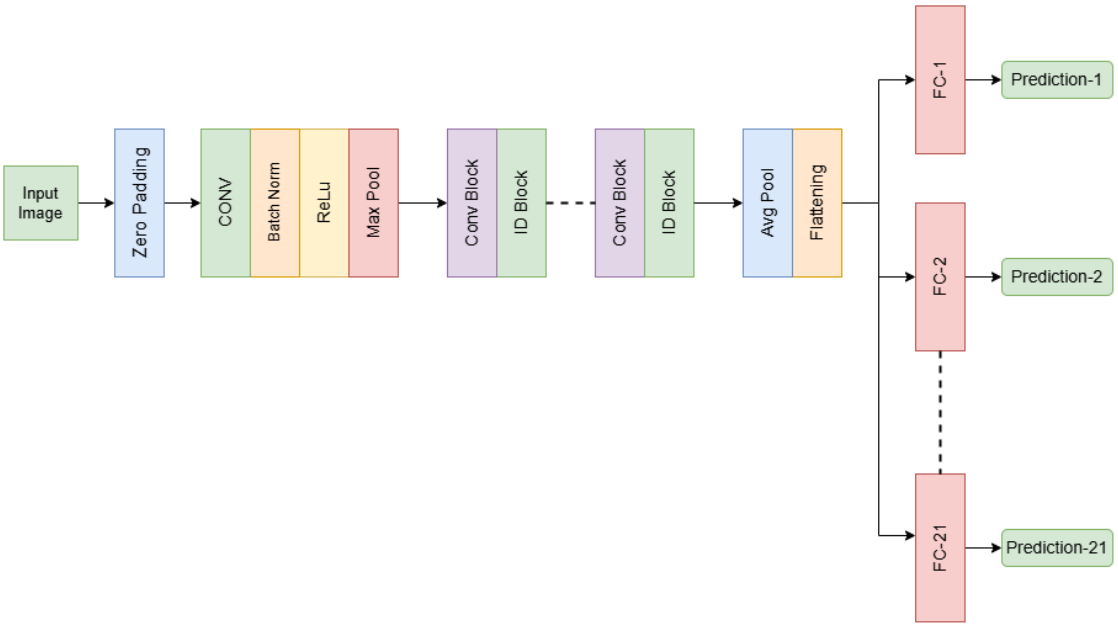


Figure 4.6: ResNet Model Architecture

The notion of Residual Networks or ResNet is the one that changed the preparation of the deep neural networks whereby the concept of residual learning, the problems that very deep networks created, as well as mostly the problem of the vanishing gradient, were solved. Traditionally, as networks got deep, the backpropagation gradients got too small, and learning with deep networks became complicated. To overcome such limitation, ResNet introduces shortcut links: it bypasses some of the layers and adds to their output the input of the given layer. The concept enabled the network to be learnt residuals i.e. The change between input and the expected output as opposed to the total transformation. The advantage of ResNet networks

is that very deep networks (including models up to several hundred layers, or even several thousand) can be trained very efficiently, just by making a point of learning the residuals.

Another significant fact about ResNet is that says shortcut connections to the identity i.e. input is added directly to the output of the sequence of layers and in turn, enables the gradients to freely flow in the network. Through this short cut route, the model has been in a position to perform excellently even though the number of layers has been maintained high. When the input and the output dimension are incompatible ResNet uses the projection shortcuts: the ResNet shortcuts channels between adjacent residual blocks are situate in a transform of the dimension consisting of a learned transformation (e.g., convolutional layer) that changes the dimension and an addition of a transform to itself. This architecture allows the ResNet models to be scalable and at the same time very precise. This has made ResNet an application used widely on programs that require deep networks, such as image recognition and object identification and semantic segmentation programs, since it is more optimized irrelevant of the training of more complex deep networks involving a lot of layers.

Image Name	*Predicted Classes*
ISIC_6753148.jpg	[0, 1, 2, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0]
_3_3279.jpg	[3, 0, 1, 3, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0]

Figure 4.7: Sample Prediction from Image Model.

From Figure 4.7, we can see some sample prediction from the image models. The first 5 position denotes name, cancerous, severity, cause and contagiousness of the disease, respectively. As these can be answered in a single word, 1 Linear Layer was used per attribute. The next 12 positions are used to determine the necessary diagnostic test recommended for that disease. As this cannot be answered in a single word, a series of linear layers were used. From our dataset, we indentified 12 unique diagnostic tests. Those were then binary encoded(0 for not necessary, 1 for necessary) and get a string of 0's and 1's that can properly recommend tests for that particular disease. The last 4 positions are for recommended prevention methods. Similarly, 4 linear layers were used to identify prevention methods.There are a total of 4 unique prevention methods. Those were also binary encoded and we get a string of length 4 that recommends prevention methods.

2. Text Encoder

MiniLM

The textual component comprising structured medical questions was processed using MiniLM- a transformer-based text encoder with 22 million parameters. This encoder was optimized for handling domain-specific medical queries with high semantic clarity and minimal latency during inference.

The MiniLM compresses big pre-trained models such as BERT using deep self-attention distillation. Using several Transformer blocks, this process moves knowledge from the last Transformer layer to a smaller Student model. MiniLM aims to enable the student to learn efficiently by distilling the self-attention module (queries, keys, and values) from the teacher's last layer to the student, thus reducing their size.

MiniLM presents two primary methods for distillation: value-relation transfer, which distills the scaled dot-product values between attention heads, and attention distribution transfer, whereby the teacher's attention maps are transferred to the student using KL-Divergence. MiniLM is quite effective since these techniques let the student model keep important knowledge from the teacher even with a lower number of layers and flexible hidden dimensions.

MiniLM differs from past models such as DistillBERT and TinyBERT in that it does not demand exact layer mapping or same layer counts between teacher and student models. MiniLM is a strong tool for task-agnostic model compression since it concentrates on the last Transformer layer and adds a teacher assistant, achieving notable compression without compromising performance.

LSTM

LSTM networks are a modification of Recurrent Neural Network (RNN), predicted to succeed over classical RNNs in their inability to learn long-term dependencies in sequential data. In classic RNN information travels through the network across time but because of the problem of vanishing gradient it becomes inefficient to make the network learn long-term dependencies. LSTM solves this problem by adopting peculiar structure that consists of gates used to control the information transmission within the network and make it capable of storing the information during long intervals.

The fundamentals included in an LSTM network encompass the forget gate, the input gate, and the output gate that help in the regulation of the information found in the memory cell. The forget gate decides what of the past information should be forgotten in the memory so that the network is able to concentrate on the pertinent past information. The input gate controls what new data needs to be put into the memory cell and allow the network to learn new data in every time step. Output gate has the role to control the output of the hidden state of the network, which is to produce prediction or to generate a sequence. The gates cooperate to enable LSTM networks to store relevant information across long sequences and are therefore suited

to time-series forecasting tasks as well as speech recognition and language modeling.

LSTMs have especially been successful in Using tasks in which order matters and long-term context is significant (i.e. machine translation, text generation and sentiment analysis). LSTMs learn how to concentrate on the appropriate patterns with time as they selectively update and forget information and make predictions on the basis of acquired knowledge. State of a cell is viewed as a conveyor belt through which useful information is transported between time steps and the gates implement the process of addition or removal of information to this state. This structure makes LSTM network one of the best architectures to work with long sequences of information storage.

Question	***Ques. Type***	***Output***
What is the name for this disease?	q1	“name”
What is the reason for this disease?	q2	“cause”
What is the severity of this disease?	q3	“severity”
What diagnostic tests are recommended?	q4	“diagnosis”
What is this disease called?	q1	“name”
How serious is this?	q3	“severity”
Is it cancerous or benign?	q5	“cancerous”
What tests are recommended for this disease?	q4	“diagnosis”

Figure 4.8: Sample Prediction from Text Model.

From Figure 4.8, we can visualize how outputs from our text model will look. When given a question, the model will extract what the question means and output a single word from a predefined set that best matches the question. The words in this predefined set must match those of the map dictionary that maps the outputs of image and text models together. As we can see from the figure, questions like ”What is the severity of this disease?” and ”How severe is this?” have the same meaning. As a result, the text model gives the same output in both cases.

3. Combined Model

The combined model offers a fraction of the size of traditional VQA models while retaining robust multimodal reasoning capabilities. We performed categorical encoding on the annotated textual data (e.g., disease names, severity levels, diagnosis types) to ensure compatibility and consistency with the model's output Linear Layers. This encoding strategy also simplified the integration of question-answer templates into the model's tokenization process and helped streamline the learning process during supervised fine-tuning. While an image is inputed, only the ViT model becomes active and extract all the necessary information from the image. It gives a 1D vector output which can be thought of as a downsized logit vector. Afterwards, when a question is provided, the text model extract the meaning or keyword from the question.

What distinguishes our approach is not only the compactness of the model but also its scalability and deployability. Our model is designed to be deployed on standard hardware without GPU acceleration, making it accessible to a broader range of clinical and field applications. Despite the reduced model size, the results were impressive: the system delivered competitive accuracy on both close-ended and open-ended question types, demonstrating that lightweight models when trained with domain-specific optimization can rival or even surpass larger counterparts in medical VQA performance.

Furthermore, as the text and image encoder works independently, there is no dimensionality mismatch error. With traditional VAQ models, the output vector dimension of the image encoder has to match the output vector dimension of the text encoder. This can often result in dimension mismatch and the vector with the lower dimension has to be expanded to match that of the higher dimension vector. This can prevent the use of a higher end model with a low cost small scale model as the dimension ultimately has to match the higher one. Our model avoid this by making them working separately.

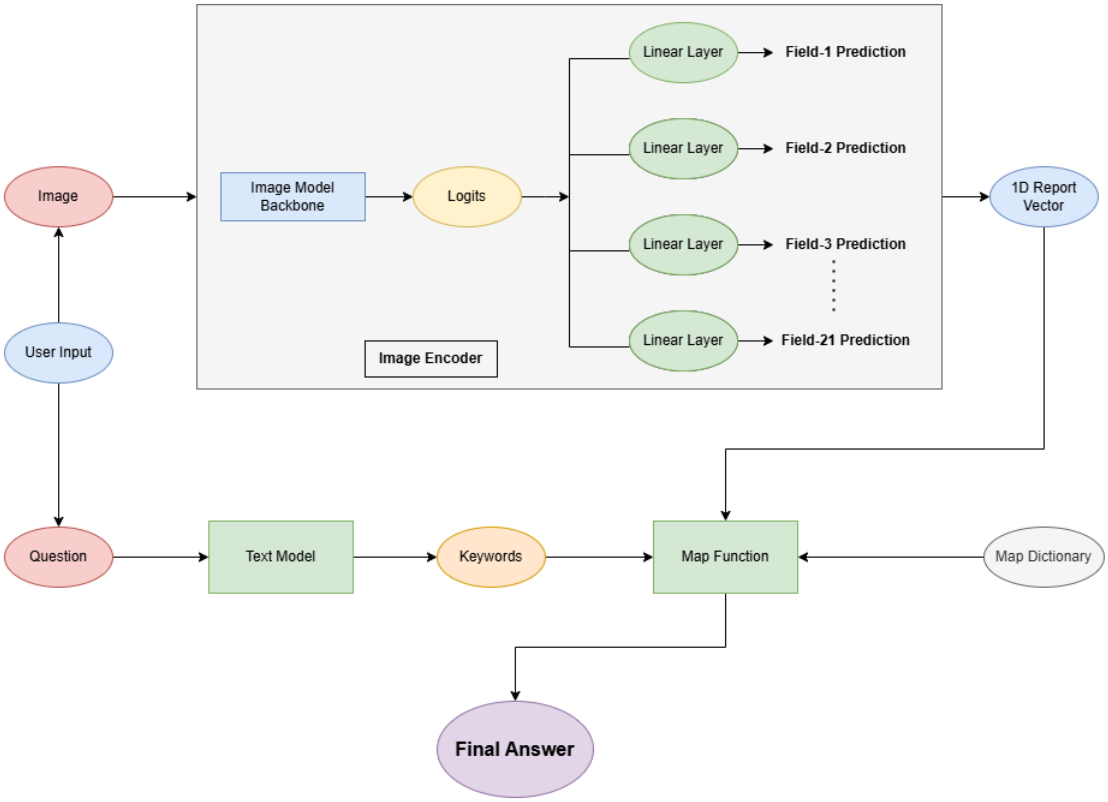


Figure 4.9: Flow chart of our proposed Modular Hybrid model

For the image section we chose ViT and ResNet. Both are relatively lightweight models, one being transformer based and the other a sequential model respectively, and yields good results for image classification. The Figure 4.9 shows the key change we made to the ViT model. For this model, we only keep the part until we get the logits. Then we pass this logits vector to multiple linear layers and each layer gives an integer as output. Each head or a series of heads give an integer value that represents the answer to a single corresponding question. As we have both open ended and close ended questions, for close ended questions, we used a single linear layer. But for open ended questions, a series of linear layers are used, the number of layers depends on the question.

As we have 5 close ended questions, 5 layers were used for them. As for closed ended questions, the diagnosis question used 12 layers as there are 12 unique diagnoses in our dataset. And for the prevention methods, 4 layers are used as we have 4 unique prevention methods mentioned in the dataset.

Figure 4.9 ilustrates the whole model pipeline from user input to the final output. While an image is inputted, only the ViT model becomes active and extracts all the necessary information from the image. It gives a 1D vector output which can be thought of as a downsized logit vector. This vector contains all the necessary information of the relevant image in the form of an 1D array containing integers values. Each position or a section of the array represents the answer of a particular question.

Afterwards, when a question is provided, the text model extracts the meaning or

keyword from the question. The keyword matches a particular key of the map dictionary. The map is a dictionary that maps a number to an answer. Value of each key is also a dictionary.

["name" : 0 : "melanoma", 1 : "chicken pox", 2 : "impetigo", 3 : "nail fungus/Onychomycosis", 4 : "ringworm", 5 : "shingles", 6 : "vascular lesion"]

This is the name key from the map dictionary. As we can see, the name key also has a dictionary as value. Each of the keys in this dictionary represents a disease. When we input a question like "What is the name of the disease?", the text model extracts the meaning/keyword from the question which in this case is "name". We then look at what is the value for disease name in the 1D array which we got from the ViT model. We then match the value with the dictionary of the "name" key and get the disease name.

In conclusion, our proposed model fulfills the primary aim of this research: to develop a lightweight, low-computation VQA system capable of delivering high-accuracy, interpretable responses for skin disease diagnosis. This contribution makes a strong case for resource-efficient AI in healthcare, opening new avenues for real-world applications where computational resources are scarce but diagnostic precision is essential.

4.5 Chapter Summary

The approach used to create the VQA model is described in this chapter together with the custom dataset-based fine-tuning of state-of- the-art (SOTA) Vision-language models (VLMs) including Gemma-3, QwenVL-2.5, BLIP-2, and LLaVA-1.5. To produce a light-weight, domain-adapted model, the proposed model combines MiniLM for text encoding with a Vision Transformer (ViT) for image encoding. This chapter also covers the hyperparameter tuning, training these models using an annotated dataset, and optimization techniques used to strike a compromise between accuracy and computational economy. The idea was to design a model fit for low-resource, real-world medical environments.

Chapter 5

Implementation and Results

5.1 Training parameters

The Table 5.2 shows all the configurations of ViT Backbone. In Table 5.3 , Table 5.4 and Table 5.5, we can see the results of hyperparameter tuning with ViT, Resnet and LSTM model. After the hyperparameter tuning, we selected the best combination of parameters for our final training of the model.

Table 5.1: ViT Backbone Parameters.

Config. Key	**Description**
Epoch Size	The total number of loops the model will run.
Batch Size	Number of images the model will look at a time.
Patch Size	Each image is split into patches of size n×n.
Image Size	Input image size.
attention_probs_dropout_prob	Dropout rate applied to attention probabilities.
hidden_dropout_prob	Dropout for hidden layers.
layer_norm_eps	Epsilon for layer normalization.
Optimizer	Algorithm that adjusts the model's weights
Activation Function	Activation function in the feed-forward layers.
Learning Rate	Steps size when updating model's weights during training..
Weight Decay	Regularization technique to prevent overfitting, adds a penalty to the loss function for having large weights.

Table 5.2: ResNet Backbone Parameters.

Config. Key	Description
Epoch Size	The total number of loops the model will run.
Mean and Standard	Used to normalize the image.
Batch Size	Number of images the model will look at a time.
Image Size	Input image size.
Optimizer	Algorithm that adjusts the model's weights
Learning Rate	Steps size when updating model's weights during training..
Weight Decay	Regularization technique to prevent overfitting, adds a penalty to the loss function for having large weights.

5.2 Hyperparameter Tuning

Epoch Size	Batch Size	Patch Size	Img Size	Hidden Dropout Prob	attention_probs_dropout_prob	Act. Func	Layer Norm Epsilon	Opt.	Lr rate	Weight decay	Acc
25	2	16	224	0.0	0.0	gelu	1e-12	Adam	2e-5	N/A	94.77%
25	4	32	224	0.1	0.1	relu	2e-5	AdamW	2e-5	.005	94.78%
25	16	32	224	0.0	0.0	relu	2e-6	AdamW	2e-4	0.005	94.36%
10	18	32	224	0.0	0.0	relu	2e-9	RMSprop	0.001	N/A	70.18%
20	4	32	224	0.0	0.0	gelu	2e-6	AdamW	2e-6	0.01	94.87%

Table 5.3: Hyperparameter tuning results for ViT with accuracy scores

Epoch Size	Mean	Standard	Img. Size	Batch Size	Learning Rate	Weight Decay	Optimizer	Accuracy
25	[0.485, 0.456, 0.406]	[0.229, 0.224, 0.225]	(224, 224)	8	2e-5	1e-4	AdamW	94.50%
25	[0.5, 0.45, 0.4]	[0.2, 0.2, 0.2]	(224, 224)	16	1e-3	1e-4	AdamW	88.58%
25	[0.485, 0.456, 0.406]	[0.229, 0.224, 0.225]	(224,224)	16	1e-3	1e-4	AdamW	89.09%
10	[0.485, 0.456, 0.406]	[0.229, 0.224, 0.225]	(224,224)	8	2e-5	N/A	Adam	94.73%

Table 5.4: Hyperparameter tuning results for Resnet with accuracy scores

Epoch Size	Batch Size	num_words	Activation	Optimizer	Learning Rate	Validation Split	Accuracy
10	16	3000	gelu	Adam	0.001	0.2	100%
10	16	1000	sigmoid	Adam	0.01	0.1	100%
15	32	5000	sigmoid	AdamW	0.01	0.2	100%

Table 5.5: Hyperparameter tuning results for LSTM with accuracy scores

5.3 Final Model Accuracies

The Table 5.6 shows the accuracies obtained in each of the separate heads used in our ViT model. The Test set from our dataset was used when finding these accuracies. Overall, we have achieved high accuracy in all of the heads with the lowest one having 79.61% accuracy.

Table 5.6: ViT Accuracy for Test Set

L.L	Acc.
name	90.29%
cancerous	95.15%
severity	89.32%
cause	98.06%
contagious	100.00%
D0:Biopsy	79.61%
D1:Dermatoscopy	91.26%
D2:Imaging	83.50%
D3:full skin examination	100.00%
D4:Blood test	96.12%
D5:DFA Test	100.00%
D6:Clinical evaluation	82.52%
D7:Serological tests	100.00%
D8:PCR test	97.09%
D9:Antigen detection test	100.00%
D10:KOH Test	94.17%
D11:Culture Test	97.09%
P0:Vaccine	100.00%
P1:Avoid UV	100.00%
P2:Maintain hygiene	97.09%
P3:Anitfungal Powder	99.03%

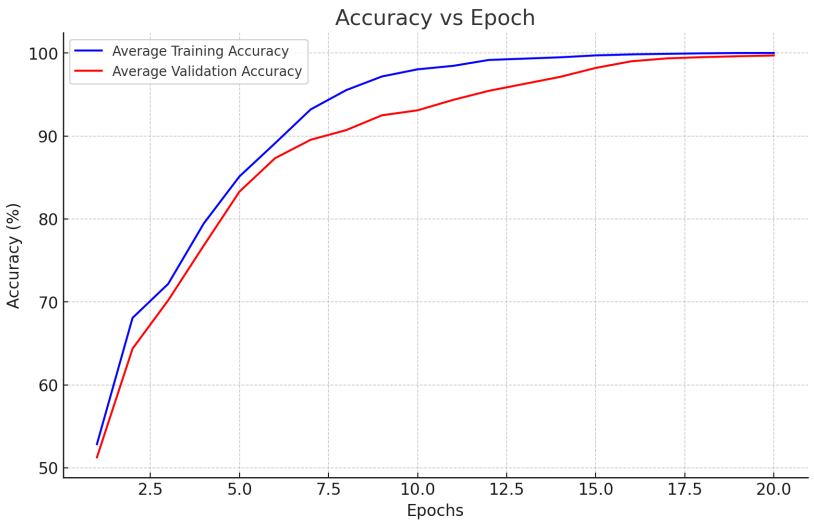


Figure 5.1: Curve of Average Training and Validation Accuracy vs Epoch for ViT

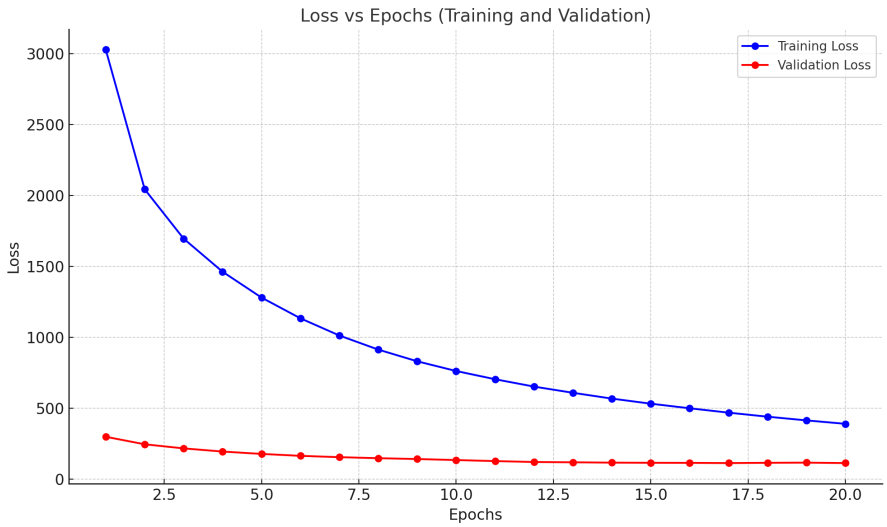


Figure 5.2: Curve of Training and Validation Loss vs Epoch for ViT

The accuracy and loss curves presented in Figures 5.1 and 5.2 demonstrate a healthy learning pattern over 20 epochs. In the accuracy plot, both training and validation accuracy increase steadily, with the training accuracy approaching 100% and the validation accuracy closely following behind, indicating strong learning and generalization capabilities without signs of overfitting. The corresponding loss curve further supports this, as both training and validation loss decrease consistently throughout the epochs. While the training loss drops more sharply, the validation loss also steadily declines, suggesting that the model is effectively minimizing error on both seen and unseen data. Overall, the curves indicate stable training dynamics and good generalization performance. Finally, in Figure 5.3 we can see the overall accuracy, precision, recall and F1 scores.

Average Head Accuracy for Test Set: 94.87%

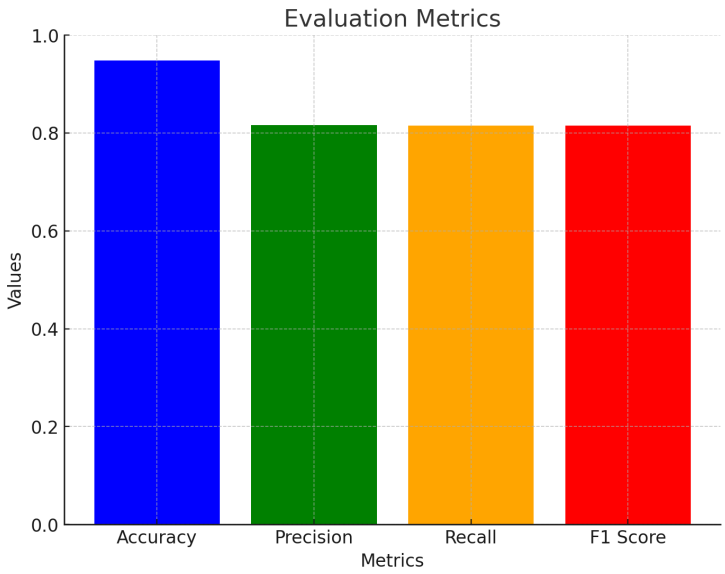


Figure 5.3: Bar Graph for ViT of Accuracy, Precision, Recall, F1

The Table 5.7 shows the accuracies obtained in each of the separate heads used in the ResNet model. The Test set from our dataset was used when finding these accuracies. Overall, for this model as well we have achieved high accuracy in all of the heads with the lowest one having 77.67% accuracy. Then in the Figures 5.4 and 5.5 we see the training and validation accuracy per epoch, and training and validation loss per epoch.

Table 5.7: ResNet Accuracy for Test Set

L.L	**Acc.**
Name	95.15%
Cancerous	98.06%
Severity	90.29%
Cause	98.06%
Contagious	100.00%
D0:Biopsy	77.67%
D1:Dermatoscopy	88.35%
D2:Imaging	81.55%
D3:Full skin examination	100.00%
D4:Blood test	96.12%
D5:DFA Test	100.00%
D6:Clinical evaluation	81.55%
D7:Serological tests	99.03%
D8:PCR test	95.15%
D9:Antigen detection test	100.00%
D10:KOH Test	94.17%
D11:Culture Test	97.09%
P0:Vaccine	100.00%
P1:Avoid UV	100.00%
P2:Maintain hygiene	98.06%
P3:Anitfungal Powder	99.03%

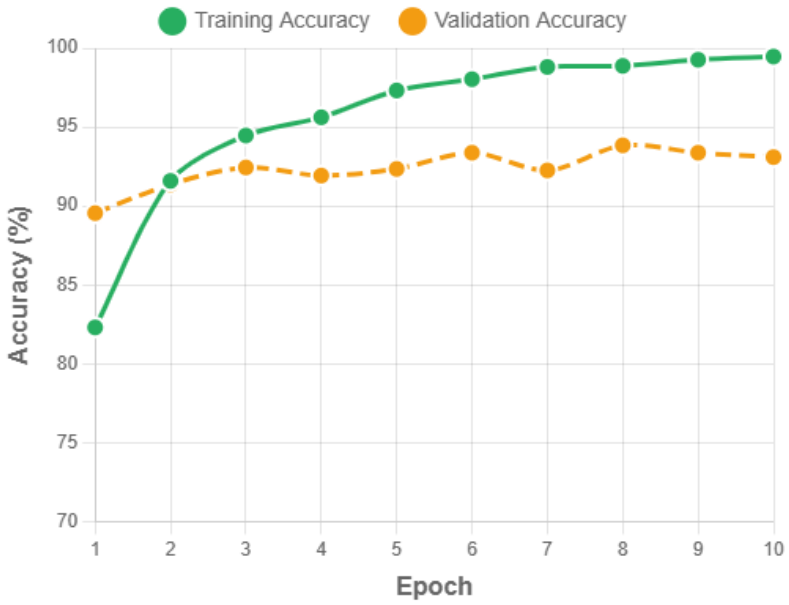


Figure 5.4: Curve of Average Training and Validation Accuracy vs Epoch for ResNet

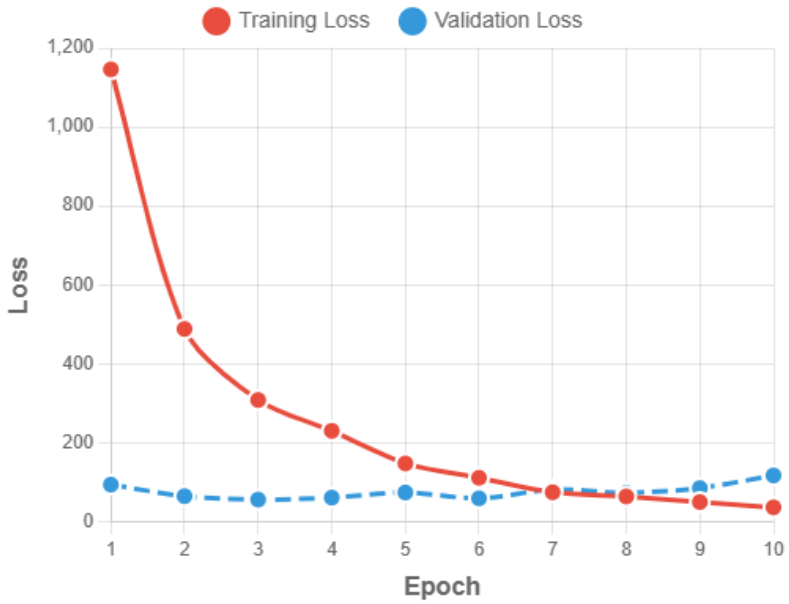


Figure 5.5: Curve of Training and Validation Loss vs Epoch for ResNet

From the Figure 5.4, In this figure we see that the training accuracy grows going epoch by epoch and then gradually leveled off as it goes near 100%. This implies that the model is picking up accurately and learning on the training data and can easily take up the task of classifying the data on high accuracy.This increase in validation accuracy is however very high at the start but is slower than the training accuracy and stays at approximately 90 percent. The difference between the training and validation accuracy shows the model is a bit overfitting to the training data because it shows a bit better accuracy on the training set as compared to the unseen

validation set. Nevertheless, the overall accuracy of validation is quite high, so it is possible to say that the model generalizes quite well in general.

From Figure 5.5, The training loss graph shows a sharp decrease during the initial phase, which is normal in early training process of the model because it is quickly learning how to make as few mistakes as possible. Upon entering the declining phase, the shape of loss curve begins to flatten, indicating that the model has achieved stable point that improvements in reducing the loss is not advancing as fast. This validation loss takes a similar path but is significantly reduced as compared to the training loss and it is good news that the model is also working well on the validation set. The gradual decrease in the loss of validation implies that the model is not overfitting on the training data but rather generalizing on validation data well. On the whole, these loss trends indicate healthy and stable learning in which the model keeps improving on the training and validation sets, and there is no considerable overfitting. Finally, in Figure 5.6 we can see the overall accuracy, precision, recall and F1 scores.

Average Head Accuracy for Test Set: 94.73%

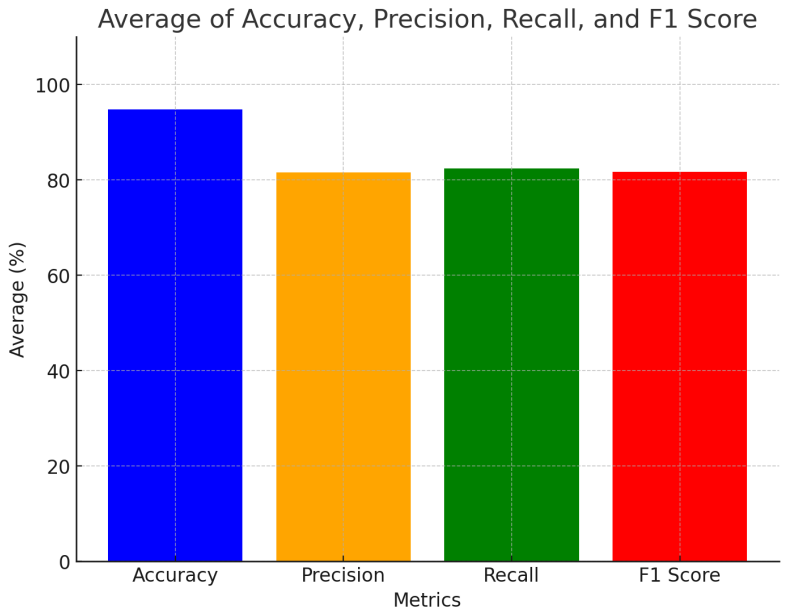


Figure 5.6: Bar Chart for ResNet of Accuracy, Precision, Recall, F1

5.4 Comparison Between ViT, ResNet and SOTA Models

Model Name	Parameters	BERT Score	Cosine Similarity
Gemma-3	4B	93.52%	77.73%
Qwen-2.5-VL	3B	94.54%	81.21%
LLaVA-1.5	7B	93.78%	77.80%
Blip-2	2.7B	94.47%	81.41%

Table 5.8: Final Results of SOTA models

Model Name	Parameters	Precision	Recall	F1 Score	Accuracy
ViT	86.43M	.816	.815	.815	94.87%
ResNet	23.62M	.815	.824	.816	94.73%

Table 5.9: Final Results of ViT and ResNet

The Table 5.8 shows the name, parameter count, BERT score and cosine similarity score of 4 different SOTA(State Of The Art) Models, all with several billion parameters. Gemma-3, which has 4 billion parameters, has a BERT score of 93.52% and a cosine similarity score of 77.73%. Qwen-2.5-VL model has 3 billion parameters and it's BERT score and cosine similarity score are 94.54% and 81.21%. LLaVa-1.5 and Blip-2 have a parameter count of 7 billion and 2.7 billion respectively but Blip gives a slight better performance, with BERT score and cosine similarity score of 94.47% and 81.41% than 93.78% and 77.8% of LLaVA.

As for our models, we have chosen the combination of ViT + MiniLM. As we can see from Table 5.9, the accuracy of ViT is slightly higher than ResNet. As we are working in the medical domain, we are prioritizing accuracy. So we have chosen the combination of ViT + MiniLM. Even Though both MiniLM and LSTM has the same accuracy, the combination ViT + LSTM would be theoretically the best considering the low parameter of LSTM, it could not properly identify a question if it differed even slightly from the training dataset. As a result, it is unsuitable for practical application where we can have spelling error or other differences.

5.5 Applying our Proposed Model on Benchmark Dataset

We have also trained our model using a different benchmark dataset [15] for comparison. The training with benchmark dataset was done using the best hyperparameters that we obtained from the hyperparameter tuning with the primary dataset. There are total 737 images in the Benchmark dataset among which 589 images, 74 images, and 74 images which used as training,validation and testing respectively.

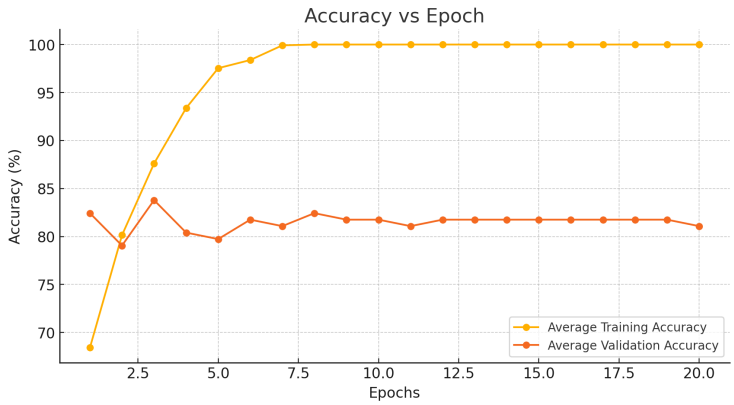


Figure 5.7: Curve of Average Training and Validation Accuracy vs Epoch on Benchmark Dataset

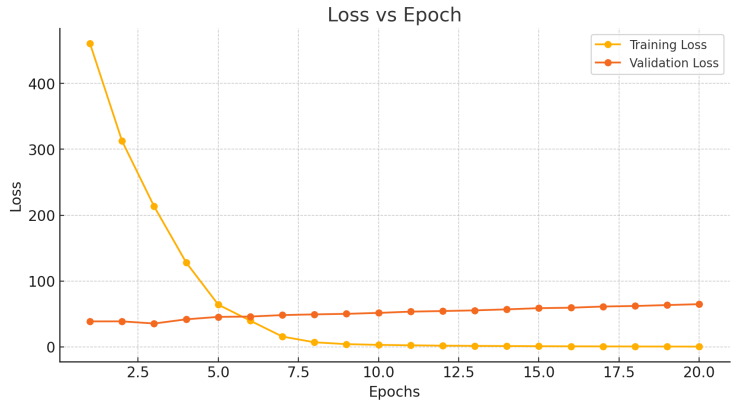


Figure 5.8: Curve of Training and Validation Loss vs Epoch on Benchmark Dataset.

In Figure 5.7, the training and validation curve indicate that the model quickly achieves high training accuracy, reaching nearly 100% by epoch 6 and maintains it afterwards. However, the validation accuracy plateaus around 80% and shows slight fluctuations, suggesting that while the model fits the training data extremely well, its generalization to unseen data is limited. This is further supported by the loss curve in Figure 5.8, where training loss decreases sharply and stabilizes near zero, but validation loss remains significantly higher and gradually increases over epochs. This divergence between training and validation performance indicates overfitting, where the model memorizes training data rather than learning patterns that generalize well to new inputs. This overfitting due to the limitation of the benchmark

dataset which lacks enough data for the model to generalize properly. However, in comparison the model generalized well with our primary dataset.

In Figure 5.10 we see the comparison of both primary and benchmark dataset. The accuracy from the benchmark dataset is 70.27% and primary dataset is 94.87%. The model obtained lower accuracy with benchmark dataset due to it being a smaller dataset and having lower question-answer pairs per image. This made it harder for the model to generalize.

Dataset	**Images**	**Q/A pair**	**Accuracy**
Custom Dataset	1038	7266	94.87%
Benchmark Dataset	737	1237	70.27%

Table 5.10: Dataset comparison

5.6 Chapter Summary

The results of training the proposed lightweight model and the fine-tuned VLMs are presented in this chapter. It ranks the models according to accuracy, precision, recall, F1 score, and semantic similarity measures including BERTScore. With an amazing 94.87% accuracy, the proposed ViT+MiniLM model exceeded other models in terms of efficiency, hence fit for deployment in low-resource settings. The results show that low computational needs allow a lightweight model to achieve great performance. Additionally we included a thorough examination of the training and validation loss curves.

Chapter 6

Website Deployment

To enhance accessibility and facilitate practical usage, we developed a user-friendly web interface that allows real-time interaction with our Visual Question Answering (VQA) model for skin disease diagnosis. The interface was implemented using Flask, a lightweight Python web framework, enabling seamless local deployment on standard computing environments without requiring cloud infrastructure or GPU support. This component plays a crucial role in demonstrating the model's usability in real-world diagnostic settings, especially in resource-constrained clinics or offline environments.

The backend is powered by our custom-trained VQA model stored as vit.pth, which was developed using a Vision Transformer (ViT) architecture. Uniquely, the model is designed with 21 parallel output heads (nn.ModuleList), where each head is responsible for answering a specific clinical question or attribute related to the skin image. For example, head 0 predicts the disease class from 11 possible categories (e.g., melanoma, chicken pox, etc.), head 2 assesses severity, and head 1 determines whether the condition is cancerous. Each output head is a fully connected layer trained independently during the model's supervised training phase

The frontend is a minimalistic form that is easy to navigate and consists of typical web input elements: it is a form where the user can attach a dermatological picture or enter a free-text medical question. When the image is submitted, it is preprocessed with the torchvision.transforms module of PyTorch and fed through the ViT encoder. Depending on the keyword identification of the input question, the equivalent head output is then selected. As an example, when the query contains such words as name or disease, the system directs the query to head 0 and gives a specific disease label based on the predefined list of disease categories. In the same way, in cancer related query, head 4 would be applied, binary output is taken to be either the classification of cancerous or non-cancerous.

To add more to semantic meaning we added a MiniLM-based textEncoder that assists in parsing the question input and can later scalably be directed towards an embedding-based retrieval or matching. This architecture guarantees that the system will be modular and may be expanded to deal with more complex forms of questions with richer textual bases.

The model and the application passed the test on a local machine using usual hardware (CPU-only environment) and reported an excellent performance when running and close-to-zero inference. The time of rendering answers to the image-question pair is less than two seconds, which ensures the great efficiency of the system to serve in the area of clinical triage, telemedicine, or offline health screening kiosk application.

This use connects a gap between academic realization of models and the application of these models in a practical diagnosis. VQA system is made completely runnable in a local environment with a minimum dependency, which makes it extremely accessible by the patients and medical professionals. The deployment strategy also protects the privacy of data since no image or question can escape the local machine and thus fulfills the mandate of clinical confidentiality.

The workflow of the interface implementation is shown in Figure-6.1 while Figure-6.2 shows step by step Flask Website Implementation.

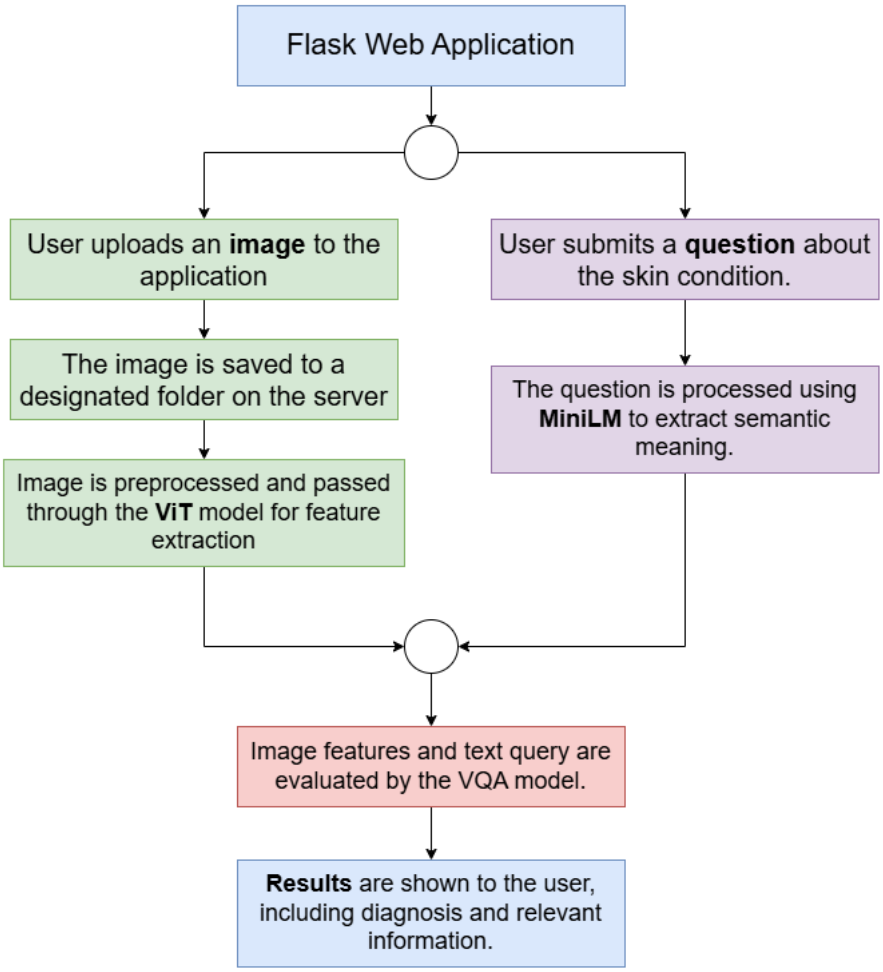


Figure 6.1: Workflow of the Flask Web Interface

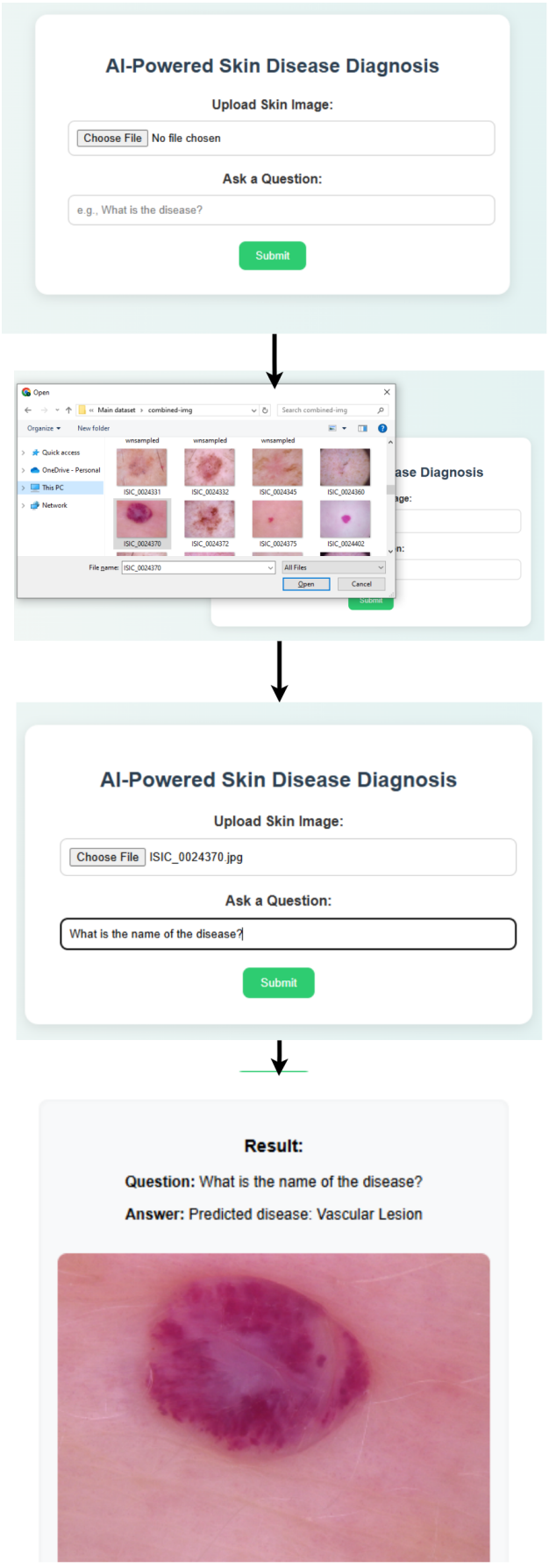


Figure 6.2: Web Deployment of the model using Flask

6.1 Chapter Summary

This chapter addresses the implementation of the trained VQA model via a Flask-based web application enabling users to upload skin disease images and get diagnostic findings in real-time. The chapter emphasizes the shortcomings of the present deployment, such performance problems with bigger images and the necessity of more optimization in environments limited in resources. Notwithstanding these difficulties, the web-based system presents an easy-to-use interface that improves patient and healthcare professional accessibility especially in remote locations.

Chapter 7

Conclusion

7.1 Conclusion

This research introduces a lightweight and efficient Visual Question Answering (VQA) model architecture designed for accurate diagnosis of common skin diseases through image and symptom-based queries. The model combines an 86 million parameter Vision Transformer (ViT) for image encoding with a 22 million parameter transformer-based text encoder-MiniLM, offering strong diagnostic performance with reduced computational demands. Unlike large-scale VQA models such as BLIP-2 or LLaVA-7B, our compact architecture is optimized for standard computing environments, enabling deployment in low-resource clinical settings that can provide reliable primary diagnosis. We developed a custom annotated dataset comprising 1,038 images across 11 disease classes, each paired with seven medically relevant question-answer sets. Using categorical encoding and hyperparameter optimization, the model achieved competitive accuracy on both close-ended and open-ended queries. Its user-friendly interface supports real-time interaction, making it practical for patients. Overall, the proposed model demonstrates a balanced trade-off between accuracy and efficiency, making it a promising solution for accessible, interpretable, and scalable AI-assisted skin disease diagnostics. The model's effectiveness is further validated by its ability to achieve a high accuracy of 94.87% on our custom-annotated dataset—despite having significantly fewer parameters than existing state-of-the-art models.

7.2 Limitations

As we know, no research is without limitations and this research is not different from those either. Although our suggested VQA model is designed to diagnose skin diseases has fantastic results and efficiency, one may observe several areas in which it may be improved to increase its functionality. The limitations are as follows:

1. Encoding process is one such key point as it is the process necessary to pre-process dataset in order to train. Encoding is an obligatory stage, but at the same time, it can be a time-consuming one. Training could be expediated further by optimization of the encoding process when considering bigger and more varied data sets.

2. The model is perfect to deal with close-ended question, although it might need further refinements to be able to better respond to open-ended queries, and particularly in case the questions are less structured or more complex. In the present form, the architecture of the model comprises 21 way output heads that support different tasks, making it more flexible. However, it also has a level of complexity. Furthermore, increase in open-ended answers can exponentially increase its final linear layer count, making it highly complex

3. Finally, the current scope of the model is restricted to the data that it was trained on. Although this will provide accurate and appropriate responses to the considered types,it is unable to answer anything outside its training scope. But this can be mitigated by extending the capacity of the model to process questions beyond the scope defined.

7.3 Future Work

While the current VQA model for skin disease diagnosis demonstrates promising results and practical applicability, several areas remain open for further exploration and enhancement. These directions aim to improve the system's generalizability, efficiency, and usability in real-world clinical environments. We have outlined the key future work directions below:

1. Expansion to Diverse and Larger Datasets

Our current model was trained and evaluated on a dataset containing 1,038 annotated images across 11 skin disease classes. In the future, the model can be extended to work with larger and more diverse datasets, including rare skin conditions and images across varied age groups, ethnicities, and environmental contexts. This would significantly enhance the model's robustness and diagnostic generalization in real-world settings.

2. Model Pruning for Efficiency

Although our proposed architecture is lightweight relative to state-of-the-art multimodal models, there is room for further optimization. In upcoming iterations, we plan to implement model pruning and compression strategies to reduce parameter count, accelerate inference, and minimize memory usage and model's complexity without compromising diagnostic accuracy. This will make the system even more suitable for edge deployment.

3. Mobile and Edge Deployment

One of the practical goals of this research is to make AI-assisted skin disease diagnosis accessible in low-resource environments. Future work will focus on deploying the model on mobile devices or embedded platforms, allowing real-time, natively run skin disease analysis for patients. This will be complemented by usability studies and surveys to evaluate the system's effectiveness and user experience in field conditions.

4. Handling Multilingual Inputs and Conversational Interaction

To expand the system's reach across diverse populations, we aim to extend support for multilingual queries and explanations, enabling users to interact with the system in their native languages. Additionally, incorporating conversational capabilities will allow more dynamic, follow-up questioning and improved user engagement.

5. Reinforcement Learning and Human-in-the-Loop Feedback

Future versions of the system will explore reinforcement learning frameworks where feedback from users—both patients and medical experts—can be used to fine-tune the model iteratively. This will enhance the model's adaptability and reasoning consistency in real-world diagnostic scenarios.

Bibliography

[1] E. A. Krupinski, "Current perspectives in medical image perception," *Attention, Perception, & Psychophysics*, vol. 72, no. 5, pp. 1205–1217, 2010.

[2] S. Antol, A. Agrawal, J. Lu, *et al.*, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Introduced VQA as a multimodal AI task, 2015.

[3] R. Gaffney and B. Rao, "Global teledermatology," *Glob. Dermatol.*, vol. 2, pp. 209–214, 2015.

[4] A. B. Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller, "Vqa-med: Overview of the medical visual question answering task at imageclef 2019.," *CLEF (working notes)*, vol. 2, no. 6, 2019.

[5] O. Kovaleva, C. Shivade, S. Kashyap, *et al.*, "Visual dialog for radiology: Data curation and firststeps.," in *ViGIL@ NeurIPS*, 2019.

[6] B. D. Nguyen, T.-T. Do, B. X. Nguyen, T. Do, E. Tjiputra, and Q. D. Tran, "Overcoming data limitation in medical visual question answering," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, Springer, 2019, pp. 522–530.

[7] R. R. Selvaraju, P. Tendulkar, D. Parikh, *et al.*, "Squinting at vqa models: Introspecting vqa models with sub-questions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 003–10 011.

[8] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 13 041–13 049.

[9] A. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, and H. Müller, "Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain," in *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*, 21-24 September 2021, 2021.

[10] S. Eslami, G. de Melo, and C. Meinel, "Does clip benefit visual question answering in the medical domain as much as it does in the general domain?" *arXiv preprint arXiv:2112.13906*, 2021.

[11] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, and C. Jawahar, "Mmbert: Multimodal bert pretraining for improved medical vqa," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2021, pp. 1033–1036.

[12] D. Sharma, S. Purushotham, and C. K. Reddy, "Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain," *Scientific Reports*, vol. 11, no. 1, p. 19 826, 2021.

[13] L. Zhou and Y. Luo, "Deep features fusion with mutual attention transformer for skin lesion diagnosis," in *2021 IEEE international conference on image processing (ICIP)*, IEEE, 2021, pp. 3797–3801.

[14] U. Naseem, M. Khushi, and J. Kim, "Vision-language transformer for interpretable pathology visual question answering," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 4, pp. 1681–1690, 2022.

[15] S. Ahmed, *Medical images with q&a: A multi-domain dataset*, https://www.kaggle.com/datasets/ahmedsta/qa-vlm-med/data, Accessed: 2025-06-18, 2023.

[16] Y. Bazi, M. M. A. Rahhal, L. Bashmal, and M. Zuair, "Vision–language model for visual question answering in medical imagery," *Bioengineering*, vol. 10, no. 3, p. 380, 2023.

[17] S. Biswas, *Skin-disease-dataset*, https://doi.org/10.34740/KAGGLE/DSV/6695743, Accessed: 2025-06-18, 2023.

[18] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, and D. Yang, "A multimodal transformer to fuse images and metadata for skin disease classification," *The Visual Computer*, vol. 39, no. 7, pp. 2781–2793, 2023.

[19] L. Canepa, S. Singh, and A. Sowmya, "Visual question answering in the medical domain," in *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, 2023, pp. 379–386.

[20] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *PMLR*, pp. 19 730–19 742, 2023. [Online]. Available: https://proceedings.mlr.press/v202/li23q.

[21] Z. Lin, D. Zhang, Q. Tao, *et al.*, "Medical visual question answering: A survey," *Artificial Intelligence in Medicine*, p. 102 611, 2023.

[22] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 34 892–34 916, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

[23] S. Lu, Y. Ding, M. Liu, Z. Yin, L. Yin, and W. Zheng, "Multiscale feature extraction and fusion of image and text in vqa," *International Journal of Computational Intelligence Systems*, vol. 16, no. 1, p. 54, 2023.

[24] M. Zafar, M. I. Sharif, M. I. Sharif, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, "Skin lesion analysis and cancer detection based on machine/deep learning techniques: A comprehensive survey," *Life*, vol. 13, no. 1, p. 146, 2023.

[25] Q. Chen and Y. Hong, "Medblip: Bootstrapping language-image pre-training from 3d medical images and texts," *Thecvf.com*, vol. 2404–2420, 2024. [Online]. Available: https://openaccess.thecvf.com/content/ACCV2024/html/Chen_MedBLIP_Bootstrapping_Language-Image_Pre-training_from_3D_Medical_Images_and_Texts_ACCV_2024_paper.html.

[26] D. P. Jeong, S. Garg, Z. C. Lipton, and M. Oberst, *Medical adaptation of large language and vision-language models: Are we making progress?* ArXiv.org, 2024. [Online]. Available: https://arxiv.org/abs/2411.04118.

[27] L. Li, J. Peng, H. Chen, C. Gao, and X. Yang, "How to configure good in-context sequence for visual question answering," *Thecvf.com*, vol. 26710–26720, 2024. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Li_How_to_Configure_Good_In-Context_Sequence_for_Visual_Question_Answering_CVPR_2024_paper.html.

[28] K. Rezaee and H. G. Zadeh, "Self-attention transformer unit-based deep learning framework for skin lesions classification in smart healthcare," *Discover Applied Sciences*, vol. 6, no. 1, p. 3, 2024.

[29] G. Team, T. Mesnard, C. Hardin, *et al.*, *Gemma: Open models based on gemini research and technology*, ArXiv.org, 2024. [Online]. Available: https://arxiv.org/abs/2403.08295.

[30] S. Bai, K. Chen, X. Liu, *et al.*, *Qwen2.5-vl technical report*, ArXiv.org, 2025. [Online]. Available: https://arxiv.org/abs/2502.13923.

[31] D. Ding, T. Yao, R. Luo, and X. Sun, "Visual question answering in robotic surgery: A comprehensive review," *IEEE Access*, vol. 1-1, 2025. [Online]. Available: https://doi.org/10.1109/access.2024.3525145.

[32] W. Dong, S. Shen, Y. Han, T. Tan, J. Wu, and H. Xu, "Generative models in medical visual question answering: A survey," *Applied Sciences*, vol. 15(6), pp. 2983–2983, 2025. [Online]. Available: https://doi.org/10.3390/app15062983.

[33] K. Huang, K. Sun, J. Li, *et al.*, "Intelligent strategy for severity scoring of skin diseases based on clinical decision-making thinking with lesion-aware transformer," *Artificial Intelligence Review*, vol. 58(4), 2025. [Online]. Available: https://doi.org/10.1007/s10462-024-11083-9.

[34] N. D. Huynh, M. R. Bouadjenek, S. Aryal, I. Razzak, and H. Hacid, *Visual question answering: From early developments to recent advances – a survey*, ArXiv.org, 2025. [Online]. Available: https://arxiv.org/abs/2501.03939.

[35] Y. Lai, J. Zhong, M. Li, S. Zhao, and X. Yang, *Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models*, ArXiv.org, 2025. [Online]. Available: https://arxiv.org/abs/2503.13939.

[36] Y. Li, Z. Lai, W. Bao, *et al.*, *Visual large language models for generalized and specialized applications*, ArXiv.org, 2025. [Online]. Available: https://arxiv.org/abs/2501.02765.

[37] J. Pan, C. Liu, J. Wu, *et al.*, *Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning*, ArXiv.org, 2025. [Online]. Available: https://arxiv.org/abs/2502.19634.

[38] A. Pandey, D. Bodo, A. Phukan, and A. Ekbal, *The quest for visual understanding: A journey through the evolution of visual question answering*, ArXiv.org, 2025. [Online]. Available: https://arxiv.org/abs/2501.07109.

[39] T. Seki, Y. Kawazoe, H. Ito, Y. Akagi, T. Takiguchi, and K. Ohe, "Assessing the performance of zero-shot visual question answering in multimodal large language models for 12-lead ecg image interpretation," *Frontiers in Cardiovascular Medicine*, vol. 12, 2025. [Online]. Available: https://doi.org/10.3389/fcvm.2025.1458289.

[40] J. Tang, Q. Liu, Y. Ye, *et al.*, *Mtvqa: Benchmarking multilingual text-centric visual question answering*, Openreview.net, 2025. [Online]. Available: https://openreview.net/forum?id=q6pm9CObJn.

[41] G. Team, A. Kamath, J. Ferret, *et al.*, *Gemma 3 technical report*, ArXiv.org, 2025. [Online]. Available: https://arxiv.org/abs/2503.19786.

[42] P. Wang, W. Lu, C. Lu, R. Zhou, M. Li, and L. Qin, "Large language model for medical images: A survey of taxonomy, systematic review, and future trends," *Big Data Mining and Analytics*, vol. 8(2), pp. 496–517, 2025. [Online]. Available: https://doi.org/10.26599/bdma.2024.9020090.

[43] T. Yu, Z. Tong, J. Yu, and K. Zhang, "Fine-grained adaptive visual prompt for generative medical visual question answering," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39(9), pp. 9662–9670, 2025. [Online]. Available: https://doi.org/10.1609/aaai.v39i9.33047.