The Kasukabe Defense Group at MEDIQA-MAGIC 2025: Clinical Visual Question Answering with Resource-Efficient Multi-modal Learning*

Notebook for the ImageCLEF Lab at CLEF 2025

Khushi Bahadur Desai[1,*,†], Varunkumar S Hiregoudar[1,†], Ishan G Kulkarni[1,†], Ratan Dhane[1,†], Padmashree Desai[1,†], Sujatha C[1,†], Uma Mudenagudi[1,†] and Ramesh Ashok Tabib[1,†]

[1]*KLE Technological University, Hubli, India*

Abstract

Dermatology is among the more visually demanding specialties of clinical medicine, where correct diagnosis is heavily dependent on the evaluation of skin lesion appearance, texture, and site. With the increasing takeup of telemedicine, the demand for strong, automatic systems capable of interpreting dermatoscopic images along with clinical queries is increasingly becoming necessary. To this end, we introduce an efficient and competitive multimodal model for the closed-ended Visual Question Answering task in dermatology. Our model blends DistilBERT to learn textual features and EfficientNet-B0 to learn visual features, balancing performance with computational cost. With the late fusion approach augmented using QID-specific classification heads, our model efficiently learns to accommodate the variability present in the dermatological queries within the dataset. We make large-scale comparisons with larger models such as ClinicalBERT, RoBERTa, Swin Transformers, and prompt-engineered counterparts. Although the top-performing model is 0.54 on macro-F1, our DistilBERT–EfficientNet-B0 baseline scores competitive accuracy with many fewer parameters and faster inference. These results highlight the utility of efficient encoders and modular design for real-world clinical decision support. Our result on the MEDIQA-MAGIC 2025 dataset demonstrates the efficacy of resource-limited models for large-scale, multimodal applications in healthcare.

Keywords

Multi-modal Learning, Visual Question Answering, Dermatology, EfficientNet, DistilBERT, MEDIQA-MAGIC, Telemedicine

1. Introduction

Dermatology is one of the most visually focused areas in clinical medicine, where making an accurate diagnosis often depends on carefully observing the appearance, texture, and location of skin lesions. In recent years, telemedicine has made it easier for patients to get dermatological care remotely by sharing images and descriptions of their skin conditions. However, this shift also brings new challenges: Automated systems now need to understand and analyze both the images and the accompanying clinical questions to provide meaningful answers.

Medical Visual Question Answering (VQA), especially in dermatology, is an exciting and growing research area aimed at tackling these challenges. The goal is to build models that can simultaneously interpret images and clinical questions to support diagnosis or patient care. While powerful architectures like RoBERTa and Swin Transformer have shown strong abilities to capture semantic and visual details, they often overfit when trained on smaller datasets, struggle with uneven question types, and demand heavy computational resources.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

khushibdsai@gmail.com (K. B. Desai); varunkumarsh369@gmail.com (V. S. Hiregoudar); ikishankulkarni16@gmail.com (I. G. Kulkarni); ratandhane748@gmail.com (R. Dhane); padmashri@kletech.ac.in (P. Desai); sujata\protect1_c@kletech.ac.in (S. C); uma@kletech.ac.in (U. Mudenagudi); ramesh\protect1_t@kletech.ac.in (R. A. Tabib)

0009-0005-8953-749X (K. B. Desai); 0009-0000-5935-8080 (V. S. Hiregoudar); 0009-0004-0530-2187 (I. G. Kulkarni); 0009-0001-4907-097X (R. Dhane); 0000-0002-0429-9682 (P. Desai)

To address these issues, we propose a lightweight but effective multi-modal model designed specifically for closed-ended VQA tasks in dermatology. Our approach uses DistilBERT [1] to process the text and EfficientNet-B0 [2] to extract visual features. This combination helps keep the model fast and efficient without compromising too much accuracy. Further, with its dynamic, question-specific classification heads and late fusion strategy, our system is able to accommodate a large number of relevant clinical questions without entailing a significant increase in complexity.

We have extensively evaluated our model on the MEDIQA-MAGIC 2025 DermaVQA-DAS database [3], and the results indicate that it achieves a very good trade-off between performance and computational cost. It performs similarly to larger, more complex models and can be more suitable in real-world clinical settings with low computing power and data quality.

2. Related Work

The survey *Medical Visual Question Answering: A survey* by Zhihong et al. reviewed methods from a large number of published papers in medical VQA [4]. Their study identifies central elements, such as model architecture and feature fusion method. From the many methods, the joint embedding framework [5] was found to be the most extensively used [6, 7, 8]. This architecture combines independent encoders for image and question input, a feature fusion mechanism, and an answering module specific to the task, e.g., Multiple Choice Questions (MCQ) or open questions. Image encoders are usually built on top of proven Convolutional Neural network (CNN) backbones such as ResNet [9] and VGG Net [10], whereas question encoders tend to use language models like Transformers [11] or LSTM [12]. Such encoders are often pre-trained with weights and can be frozen or end-to-end fine-tuned when training VQA models. The answering module is usually realized as either a neural network classifier for MCQs or a generative model for open-ended questions. A popular fusion approach is concatenation that directly combines question and image features. More recent methods have used attention mechanisms to enhance feature fusion. In general, although architecture diversity in the field is quite limited, joint embedding models—generally utilizing VGG Net and EfficientNet-B0 for image encoding and different models like LSTM, Bi-LSTM [13], GRU [14], and transformer-based encoders such as BERT and DistilBERT for text—predominate [15, 16, 17]. But a lightweight natural language encoder such as LSTM could be the best fit to optimize efficiency and resources for medical VQA as the task generally deals with fewer question types [18, 19, 16, 8]. In order to address various kinds of questions like MCQs of varying number of options, open-ended questions and many more. Multiple Classifier Heads are proposed which are question type specific. Each head is tasked with predicting the response to a specific kind of question so that the model can better understand the semantics of various types of questions.

In addition to subtask strategies, other techniques have also been investigated. Global Average Pooling (GAP) [20], for instance, substitutes standard fully connected layers with averaged feature maps, resulting in better generalization and less overfitting. Other developments are Embedding-based Topic Latent Question Semantics modeling to incorporate latent question semantics, Question-Conditioned Reasoning for input query-guided dynamic decision-making, and Image Size Encoders to consider spatial properties of medical images. These methods together emphasize the increasing trend of including domain-sensitive and task-sensitive mechanisms in medical VQA systems.

3. Proposed Methodology

Our framework is based on a modular late-fusion model intended to jointly reason dermatological images and associated clinical questions for closed-ended VQA. Every model input is a dermatoscopic image with a clinical question marked by a Question Identifier (QID). The model aims to predict the right answer index for each QID-image pair.

To transform the text input, we use DistilBERT, a distilled variant of BERT that achieves an optimal trade-off between representational capacity and computational cost. Every clinical question goes through tokenization and then through DistilBERT. The representation of the `[CLS]` token is taken and fed through a linear layer to generate a 256-dimensional embedding capturing the semantic content of the query.

On the visual front, we use EfficientNet-B0—a small convolutional neural network with an excellent performance-to-parameter ratio. We resize each input image to 224×224 pixels and normalize it before passing it through the EfficientNet backbone. The end convolutional features are pooled globally using average pooling to produce a 1280-dimensional vector of an image, which is subsequently projected to 256 dimensions by a linear layer.

The image and text projected embeddings are concatenated to produce a 512-dimensional multi-modal representation. To address the challenge's requirement of multiple different question types being answered, we add a QID-specific set of classification heads. Each head predicts a probability distribution over available answers for a question, allowing the model to tailor its predictions to each question type's semantics.

The model is trained end-to-end with categorical cross-entropy loss, with the correct classification head chosen dynamically depending on the QID of each sample. We train the network with AdamW optimizer and learning rate 2×10^{-5} and measure performance with the macro-F1 score to counteract the class imbalance. Training continues for as many as 25 epochs with early stopping by stagnation in validation F1 score. At inference, each test image is combined with all valid QIDs, and predictions are produced with their corresponding classifier heads. The final predictions are formatted in the appropriate JSON format for submission.The current implementation predicts a single best answer for each question. Extending the model to support multiple-answer predictions could be considered in future work

3.1. Architecture

M1 – Baseline Multi-modal Model: M1 uses the basic architecture outlined above, with DistilBERT as the text encoder and EfficientNet-B0 as the image encoder. This model serves as the basis of our solution and sets a solid baseline that is both efficient and effective. It has a shared classification head for all question types.

M2 – Clinical Text and Larger Image Encoder: In M2, we replace DistilBERT with ClinicalBERT, a language model trained on clinical text, which increases the model's capacity to understand domain-specific jargon. We also improve the visual encoder to EfficientNet-B3, a deeper variant that can detect more subtle visual features. The model retains late fusion and a shared classifier head.

M3 – RoBERTa with Prompt Engineering: To enhance text understanding, M3 substitutes the text encoder with RoBERTa-base, which is renowned for its excellent contextual insight. We incorporate prompt engineering methods by rephrasing and augmenting the input queries in a manner that enables the model to read more pertinent linguistic signals. The image encoder is still EfficientNet-B3, while predictions are derived via a shared classification head.

M4 – Domain-Aware Transformers: Model M4 uses Bio_ClinicalBERT as text encoding, a model that has been pre-trained on biomedical corpora. It uses the Swin Transformer for processing images, which is robust at learning spatial relationships with its hierarchical attention mechanism. The model continues to use the late-fusion scheme and shared head but gets improved domain-specific feature extraction in both modalities.

M5 – Question-Specific Decoding with Weighted Loss: M5 keeps the backbone encoders of M3 (RoBERTa-base and EfficientNet-B3) but adds a collection of specialized classifier heads, one for each QID, to facilitate fine-grained reasoning across question types. To compensate for skewed label distributions, the training loss is compensated with class-weighted cross-entropy to pay closer attention to minority classes.

M6 – Prompt-Augmented and Fine-Tuned Model: Extending M5, this model incorporates contextual cues like patient history as direct input to the input text in order to enhance semantic awareness. Also, the entire model is end-to-end fine-tuned on the VQA dataset. These changes as a whole provide the best performance among all the tested configurations.

4. Training and Validation

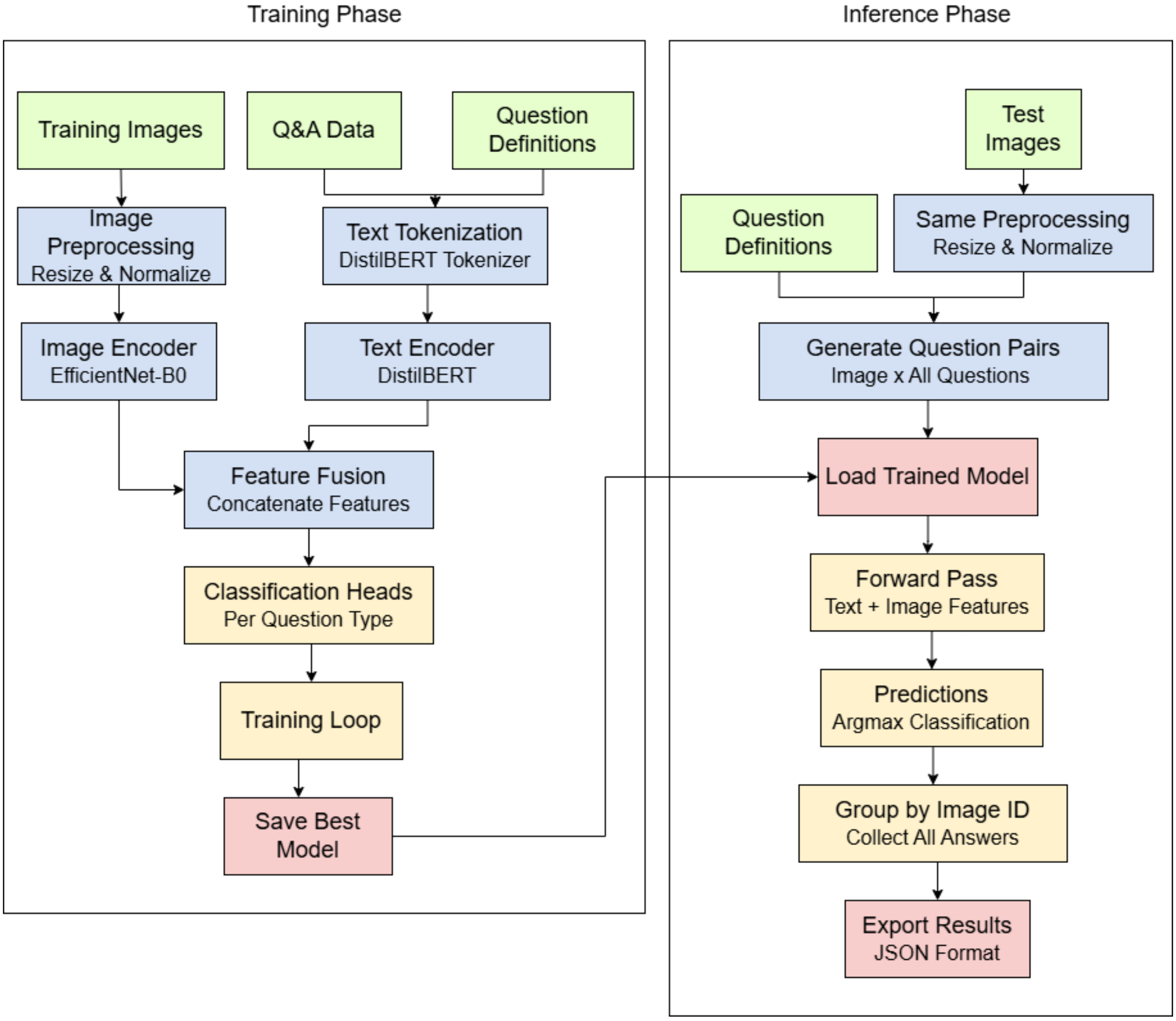

Figure 1: Training and Inference workflow

The proposed framework presents a domain-adapted, efficient, and explainable multi-modal system for closed-ended VQA in dermatology as part of the MEDIQA-MAGIC 2025 competition. The problem consists of interpreting dermoscopic images and answering structured clinical queries based on visual and textual information. This is motivated by the operational limitations common to real-world deployments, including computational resource constraints, unbalanced question distributions, and patient-generated imagery variations.

Figure 1 depicts the end-to-end workflow of our model during both the training and inference steps. Architecturally, we use a late fusion approach, where image and text features are encoded separately and subsequently combined to create a single multi-modal representation employed for classification.

For the text modality, we use DistilBERT, a distilled BERT that preserves much of the linguistic ability of BERT but is faster and less resource-hungry. Clinical queries are tokenized and fed into DistilBERT, from which the `[CLS]` token embedding is taken as a semantic summary. This embedding is linearly projected into a 256-dimensional latent space.

From the visual perspective, we utilize EfficientNet-B0, a small but efficient CNN, which is pretrained on ImageNet. Dermatoscopic images are normalized and resized and then fed into the encoder. Deep visual features are obtained with GAP such that we get a 1280-dimensional vector, which is then projected to 256 dimensions by another projection layer in order to be compatible with the features of text.

The two 256-dimensional vectors are concatenated into a 512-dimensional joint representation. To address the variability of clinical questions—each of them defined by a QID—we create a modular classifier head mechanism. A different classifier is attached to each QID, enabling the model to specialize and generate predictions sensitive to the semantics of each question type. This architecture facilitates multi-task learning and enhances robustness over heterogeneous clinical attributes, like lesion size, color, texture, and anatomical location.

We train the model end-to-end with categorical cross-entropy loss. The correct head for each instance is dynamically chosen based on the corresponding QID. We address class imbalance by measuring model performance with macro F1 score and using early stopping on validation F1 gains. We optimize the model with the AdamW optimizer with learning rate 2×10^{-5}, and train for 25 epochs, keeping track of the best-performing checkpoint.

While inferring, the model goes through each test image by combining it with all relevant questions. Prediction is done based on the respective QID-specific heads, and the output is a structured JSON file that maps every `encounter_id` and `QID` to its predicted answer index, following the MEDIQA-MAGIC submission format.

In short, the suggested framework is computationally efficient in terms of predictive accuracy, suitable for being deployed in clinical or mobile settings. Its extensibility and modularity guarantee that it can be extended to handle new question types and changing dermatology applications with ease.

5. Results and Discussion

Understanding the capabilities and limitations of our model requires a well-defined evaluation set-up grounded in representative datasets. In this section, we first describe the official DermaVQA-DAS dataset prepared for the challenge, followed by an analysis of the model's performance to the challenge-specific evaluation metric.

5.1. Dataset Description

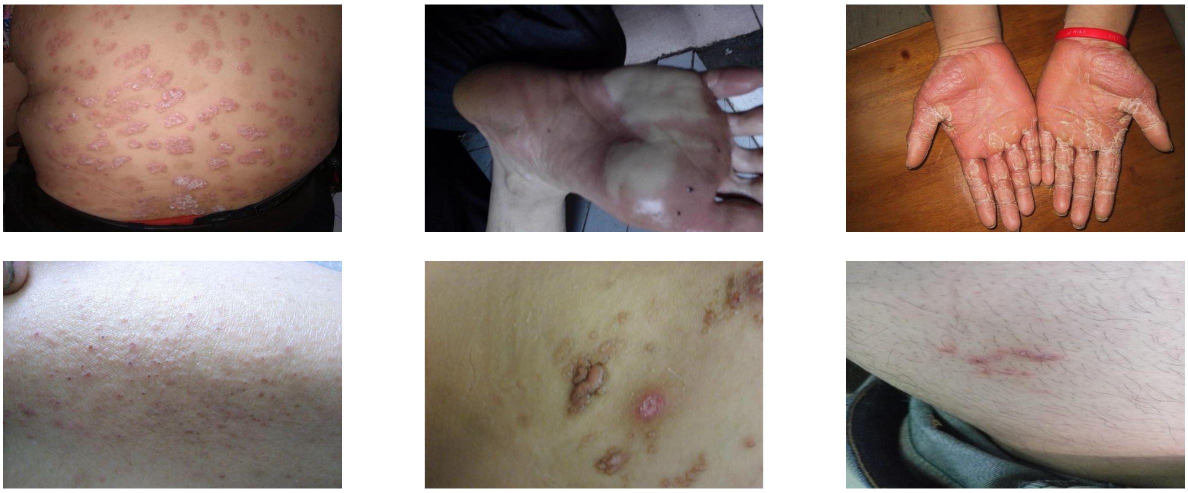

Figure 2: Dataset samples of varying skin diseases

In this study, we use the dataset provided as part of the ImageCLEF 2025 MEDIQA-MAGIC challenge [21], specifically designed for the closed-ended VQA task in dermatology. Figure 2 displays samples from the DermaVQA-DAS dataset [3]. This dataset aims to assess the ability of multi-modal models to interpret and analyze over both clinical images and related structured questions related about skin conditions.

The image corpus is made up of high-resolution dermatological images gathered from real-world clinical settings, as well as patient-generated submissions. Each image corresponds to a unique encounter, identified by an encounter_id embedded in the filename (e.g., IMG_ENC00001_00001.jpg). A wide range of dermatological issues are captured in different parts such as the back, abdomen, palms, and feet. The dataset reflects real-world conditions in lighting, skin tones, and lesion types — including rashes, pigmentations, bumps, and more—making it well-suited for training robust models.

Accompanying each image is a set of closed-ended questions. The questions follow a consistent schema, and each one is linked to a unique identifier QID. For example, a question like "How much of the body is affected?" may offer fixed response options such as "single spot," "limited area," or "widespread." These structured options enable consistent supervision across training samples. The definitions of all questions and their answer choices are provided in a dedicated JSON file (closedquestions_definitions_imageclef2025.json), which our model parses dynamically during both training and inference. Each image is associated with multiple questions, which allows for a multi-task learning setup where the model must produce separate predictions for different aspects of the same case.

5.2. Evaluation Metrics

To evaluate our system performs on the dermatology VQA task, we used the official evaluation script provided by the MEDIQA-MAGIC 2025 challenge organizers. This script is specifically designed to evaluate models that generate closed-ended clinical responses to image-related questions in dermatology. Each question is associated with a unique identifier (e.g., `CQID010-001`), allowing for easy tracking and performance analysis across different question categories.

Unlike traditional classification metrics, this evaluation employs a Jaccard Index-based accuracy measure—commonly known as Intersection over Union (IoU) accuracy. This metric is particularly advantageous in cases where multiple correct answers may exist, as it provides partial credit for overlapping predictions. The IoU-based accuracy for a single prediction is defined as in the equation 1.

$$\text{Accuracy}(x, y) = \frac{|x \cap y|}{\max(|x|, |y|)} \tag{1}$$

where x is the set of predicted labels and y is the set of true labels. For most single-label questions, this metric simplifies to a binary match.

To compute the overall performance score, the script calculates the average IoU-based accuracy across all question-image pairs as shown in the equation 2.

$$\text{Total Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \frac{|x_i \cap y_i|}{\max(|x_i|, |y_i|)} \tag{2}$$

where N is the total number of evaluated samples.

In addition to the overall accuracy, the evaluation tool provides a detailed breakdown by question category—such as lesion size, affected body area, or skin texture. This enables a more fine-grained analysis of model performance across clinically relevant subgroups.

By leveraging this IoU-based metric, the evaluation framework aligns more closely with real-world clinical practice, where ambiguity and partial correctness are common. It offers a more nuanced and forgiving measure of model effectiveness, as opposed to strict all-or-nothing correctness.

5.3. Model comparison

Table 1 shows the performance of six different multi-modal models on the dermatology VQA task using the validation set. Among these, Model M1—built with *DistilBERT* for processing clinical text and *EfficientNet-B0* for image analysis—achieved the highest macro-F1 score. Notably, M1 combines strong accuracy with a relatively lightweight design, making it a practical candidate for clinical environments where computing resources are limited.

Table 1
Comparison of multi-modal architectures on the dermatology VQA task (Validation Set).

Model	Text Encoder	Image Encoder	Features	Accuracy	Macro F1
M1	DistilBERT	EfficientNet-B0	Baseline multi-modal architecture	~0.62	~0.54
M2	ClinicalBERT	EfficientNet-B3	Clinical pretraining, larger image model	~0.62	~0.53
M3	RoBERTa-base	EfficientNet-B3	Better text modeling, prompt-engineered input	~0.60	~0.51
M4	Bio_ClinicalBERT	Swin Transformer	Better spatial attention, domain-specific text encoder	~0.59	~0.49
M5	RoBERTa-base	EfficientNet-B3	QID-specific heads, weighted CE loss	~0.57	~0.47
M6	RoBERTa-base	EfficientNet-B3	Prompted text (Q: + history), fine-tuned end-to-end	~0.56	~0.45

Model M2, which uses *ClinicalBERT* alongside a deeper image encoder (*EfficientNet-B3*), reached similar accuracy but required more computational resources. Although it was more complex, the improvements over M1 were marginal. Models M3 through M6 introduced additional strategies—such as prompt engineering, advanced attention mechanisms, and full fine-tuning—but these did not lead to consistently higher scores.

Interestingly, the most sophisticated setup, Model M6, turned out to be the weakest performer overall. This outcome suggests that adding complexity does not always translate into better results, especially when data availability is limited or question types are unevenly distributed.

To validate our findings, we submitted the top three models for official evaluation on the test set. The results are presented in Table 2, which reflects the final macro-F1 scores provided by the MEDIQA-MAGIC 2025 organizers. Model M1 again achieved the strongest performance, followed closely by M4. These results reinforce the idea that simpler, well-balanced designs can be more reliable in real-world applications than more elaborate alternatives.

Table 2
Official test results as reported by the MEDIQA-MAGIC 2025 evaluation server.

Model	Text Encoder	Image Encoder	Macro F1 (Test)
M1	DistilBERT	EfficientNet-B0	0.537
M4	Bio_ClinicalBERT	Swin Transformer	0.526
M6	RoBERTa-base	EfficientNet-B3	0.464

Overall, these findings indicate that M1's streamlined approach offers a strong balance of performance, speed, and ease of use. In clinical scenarios where reliability and efficiency are priorities, a clear and focused architecture like M1 can often be the most effective solution.

6. Conclusions

In our proposed framework, a light yet effective multi-modal model specifically designed for the closed-ended visual question-answering task in dermatology. Through integrating DistilBERT for processing clinical questions with EfficientNet-B0 for dermatoscopic image analysis, our system finds a robust balance between accuracy and computation time.

Our testing on the MEDIQA-MAGIC 2025 dataset indicated that this reduced design is as effective as more elaborate, resource-intensive models. Due to its simplicity and resilience, the model is especially conducive to real-world deployment—particularly in clinical settings where computing resources might be restricted and image quality can be inconsistent. Its modular architecture, combined with specific classifier heads per question type, allows for flexible adaptation to virtually any dermatologic query.

Declaration on Generative AI

Either:
The author(s) have not employed any Generative AI tools.

Or (by using the activity taxonomy in ceur-ws.org/genai-tax.html):
During the preparation of this work, the author(s) used X-GPT-4 and Gramby in order to: Grammar and spelling check. Further, the author(s) used X-AI-IMG for figure 1 in order to: Generate images. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL: https://arxiv.org/abs/1910.01108. arXiv:1910.01108.

[2] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. URL: https://arxiv.org/abs/1905.11946. arXiv:1905.11946.

[3] W. Yim, Y. Fu, A. Ben Abacha, M. Yetisgen, N. Codella, R. A. Novoa, J. Malvehy, Dermavqa-das: Dermatology assessment schema (das) and datasets for closed-ended question answering and segmentation in patient-generated dermatology images, CoRR (2025).

[4] Z. Lin, D. Zhang, Q. Tao, D. Shi, G. Haffari, Q. Wu, M. He, Z. Ge, Medical visual question answering: A survey, Artificial Intelligence in Medicine 143 (2023) 102611.

[5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, Vqa: Visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.

[6] H. Gong, R. Huang, G. Chen, G. Li, Sysu-hcp at vqa-med 2021: A data-centric model with efficient training methodology for medical visual question answering, in: Conference and Labs of the Evaluation Forum, 2021. URL: https://api.semanticscholar.org/CorpusID:237298665.

[7] H. Gong, G. Chen, S. Liu, Y. Yu, G. Li, Cross-modal self-attention with multi-task pre-training for medical visual question answering, 2021. URL: https://arxiv.org/abs/2105.00136. arXiv:2105.00136.

[8] D. Sharma, S. Purushotham, C. K. Reddy, Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain, Scientific Reports 11 (2021) 19826.

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[10] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[12] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[13] Z. Huang, W. Xu, K. Yu, Bidirectional lstm-crf models for sequence tagging, arXiv preprint arXiv:1508.01991 (2015).

[14] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. URL: https://arxiv.org/abs/1412.3555. arXiv:1412.3555.

[15] W. Zheng, L. Yan, F.-Y. Wang, C. Gou, Learning from the guidance: Knowledge embedded meta-learning for medical visual question answering, in: Neural Information Processing: 27th

International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV 27, Springer, 2020, pp. 194–202.

[16] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, 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.

[17] Q. Xiao, X. Zhou, Y. Xiao, K. Zhao, Yunnan university at vqa-med 2021: Pretrained biobert for medical domain visual question answering., in: CLEF (Working Notes), 2021, pp. 1405–1411.

[18] B. Liu, L.-M. Zhan, X.-M. Wu, Contrastive pre-training and representation distillation for medical visual question answering based on radiology images, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24, Springer, 2021, pp. 210–220.

[19] L.-M. Zhan, B. Liu, L. Fan, J. Chen, X.-M. Wu, Medical visual question answering via conditional reasoning, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2345–2354.

[20] M. Lin, Q. Chen, S. Yan, Network in network, arXiv preprint arXiv:1312.4400 (2013).

[21] W. Yim, A. Ben Abacha, N. Codella, R. A. Novoa, J. Malvehy, Overview of the mediqa-magic task at imageclef 2025: Multimodal and generative telemedicine in dermatology, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Span, 2025.