Overview of the MEDIQA-MAGIC Task at ImageCLEF 2025: Multimodal And Generative TelemedICine in Dermatology

Notebook for the IMAGECLEF Lab at CLEF 2025

Wen-wai Yim[1], Asma Ben Abacha[1], Noel Codella[1], Roberto Andres Novoa[2] and Josep Malvehy[3]

[1]Microsoft Health AI, USA
[2]Stanford University, USA
[3]Hospital Clinic of Barcelona, Spain

Abstract

The second edition for the MEDIQA-MAGIC[1] task builds on last year's challenges[2, 3] using an expanded multimodal dermatology dataset. Participants receive clinical narratives with related images and must complete two subtasks: (1) segmenting areas showing dermatological issues, and (2) answering closed-ended clinical questions based on the provided context. Test sets are annotated by at least three annotators. Questions and options are available in both English and Chinese. Six teams competed across both subtasks. The best-performing system for segmenting dermatological consumer health images scored 0.646 Jaccard, 0.785 Dice. For dermatological closed-ended QA, the best system achieved 0.76 accuracy.

Keywords

Visual Question Answering, Segmentation, Dermatology

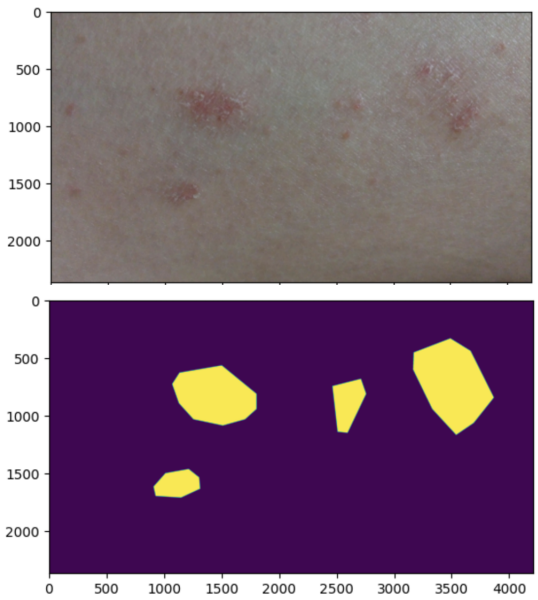
1. Introduction

Improvements in generalized artificial intelligence (AI) models, e.g., ChatGPT and DeepSeek, and their accessibility to consumers have made them powerful tools for question answering and general knowledge discovery. Their application and accuracy as medical assistive tools is crticial as two trends emerge: (a) healthcare systems' increasing adoption of AI into the electronic medical record and healthcare operations, and (b) patients' strengthened empowerment over health information seeking behavior through the internet.

In the first MEDIQA-MAGIC task in 2024 [3], we introduced the problem of consumer health multimodal visual question answering. Participants were given consumer health queries (e.g., "I've had this rash for two weeks what should I do?"), along with a patient-provided image (e.g., photo taken from a mobile device), and tasked to generate free text responses. The task is congruent to asynchronous clinical questions that can be posed to doctors through email or chats in real healthcare settings – a care delivery method shown to be increasing in adoption to lower costs [4]. Given the well-documented rate of physician burnout [5], such a technology can be applied to assist physician efficiency by pre-generating draft responses.

In the second edition of the MEDIQA-MAGIC task at ImageCLEF 2025 [6], we build upon last year's dataset, DermaVQA, [7], and its associated challenges [2, 3], extending them with a focus on closed-ended multimodal dermatology question answering [8]. In this edition, participants were asked to identify areas of interest in an image based on the patient's query (e.g., "the rash on an arm"), as well as to answer structured closed-ended questions (e.g., "is there single or multiple lesions"). These are critical subtasks that can be used to improve end-to-end free text response generation, the subject of the original 2024 challenge.

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



Query Title: Please take a look, is this eczema or dermatitis?

Query Content:
Would expert here help take a look. What is the red spot here? Mainly grow on the back of the body, buttock, and the outer side of the thigh. Is there any way to treat it....

Images: [Image01,Image02,Image03]

Where is the site affected? (1) : back
Where is the site affected? (2) : buttock
Where is the site affected? (3) : thigh

What is the skin lesion texture?: rough

How many skin lesions are there?: Multiple

What is the color of the skin lesion?: Pink

Figure 1: In this task, the original consumer health query and an accompanying image are given. This year's task is to create the relevant segmentation of the problem as well as provide an answer to a multiple-choice question.

2. Task Description and Dataset

Similar to the previous edition, participants were given a clinical narrative context along with accompanying images. The task was divided into two relevant sub-parts: (i) segmentation of dermatological problem regions, and (ii) providing answers to closed-ended questions. The questions, answers, and answer options were given in both English and Chinese.

In the first subtask, given each image and the clinical history, participants are tasked with generating segmentations of the regions of interest for the described dermatological problem. The expected outputs are binary image files with the same size as the original image.

In the second subtask, participants were given a patient dermatological query, its accompanying images, as well as a closed-ended question with accompanying choices – the task is to select the correct answer to each closed question.

The dataset was created by using real consumer health users' queries and images; the question schema was created in collaboration with two certified dermatologists. In total, closed question schema - a comprehensive list of clinically relevant, patient-facing questions for dermatological assessments included a total of 137 questions. More details of this can be found in our corresponding dataset paper [8]. For the challenge, we tested for a total of 27 questions, which were the most common and could be answered using both text and images. These corresponded to nine overall questions when related questions are grouped (e.g., "anatomic region for affected area 1", "anatomic region for affected area 2"). The answers were labeled by at least three annotators: two medical scribe annotators, and one biomedical informatics graduate student. Questions and answers were translated into Chinese by a native Chinese speaker. Further details can be found in the DermaVQA-DAS dataset paper [8]. Congruent with the MEDIQA-M3G edition [3], there was a total of 300, 56, and 100 instances for training, validation, and test splits, respectively. Each query had on average three images.

3. Evaluation Methodology

To leverage multiple gold standard masks for segmentation, we used the majority vote per pixel as the gold standard for microscore calculations of the Jaccard and Dice indices for ranking. The mean of the

per-instance max and mean for all test instances were also reported.

Because the same dermatological problem may have multiple sites, there may be related questions (e.g., "what is the size of the affected area for location 1", "what is the size of the affected area for location 2"). In these cases, the answers to the related questions are collated together. Partial credit was given when there are partial matches to gold. The evaluation code can be found here: github.com/wyim/ImageCLEF-MAGIC-2025.

Table 1
Participating Teams in the MEDIQA-MAGIC 2025 Challenge

Team	Institution	Affiliation
DS@GT[9]	United States	Georgia Institute of Technology
H3N1[10]	Vietnam	University of Information Technology
Kasukabe Defense Group[11]	India	KLE technological university
Anastasia[12]	Vietnam	Universiy of Information Technology
IReL, IIT(BHU)[13]	India	Indian Institute of Technology(BHU)
KLE1 [14]	India	KLE Technological University
Oggy	Vietnam	University of Information Technology

4. Results

Fifty-three teams registered for the event. A total of 56 valid completed runs were submitted by six teams. Table 1 provides a list of participating teams and affiliations. This year's primary participants came from academic institutions in the United States, Vietnam, and India.

Table 2 shows results for the segmentation task. Despite being calculated differently, the Jaccard and Dice metrics yielded identical rankings. Table 3 shows results for the closed-ended question answering task.

Table 2
Performance of the participating teams in the MEDIQA 2025 Subtask 1 on segmentation generation for dermatological problems. Scores from duplicate submissions were excluded.

	meanofmax		meanofmean		majorityvote	
team_name	jaccard	dice	jaccard	dice	jaccard	dice
Anastasia	0.677	0.783	0.591	0.705	0.646	0.785
Anastasia	0.631	0.742	0.550	0.666	0.611	0.759
IRLab@IITBHU	0.655	0.765	0.569	0.686	0.588	0.741
KLE1	0.638	0.751	0.554	0.671	0.541	0.702
H3N1	0.636	0.743	0.547	0.659	0.514	0.679
Anastasia	0.521	0.633	0.411	0.525	0.321	0.485
Anastasia	0.523	0.635	0.411	0.525	0.313	0.477
Kasukabe Defense Group	0.162	0.224	0.135	0.191	0.187	0.315

In the *segmentation subtask*, all four teams took a fine-tuning approach with differences in the exact models employed (e.g., TransUNet, ViT-B, CLIP). The Anastasia team enriched the dataset by performing image transformation techniques (e.g., rotations, contrast adjustments) and were able to achieve top performance after including data with all transformations. The IReL, IIT(BHU) team was the only team that attempted to incorporate textual features. Their strategy used CLIP to embed both text and visual features then afterwards fed the combined feature vector into a binary classification to predict the mask. The remaining teams fine-tuned previously trained skin lesion segmentation models; the H3N1 team used the DermoSegDiff [15] model, whereas the KLE1 team fine-tuned a Multi-Scale Feature Fusion Network model [16]. Though these models were trained for skin lesions, it is likely that further fine-tuning was required to completely adapt the model to this new dataset.

Table 3
Performance of the participating teams in the MEDIQA 2025 Subtask 2 on closed-ended question answering. Scores from duplicate submissions were excluded.

team	CQID010	CQID011	CQID012	CQID015	CQID020	CQID025	CQID034	CQID035	CQID036	ALL
H3N1	0.7	0.89	0.77	0.91	0.69	0.97	0.45	0.86	0.58	0.76
H3N1	0.64	0.89	0.76	0.87	0.71	0.96	0.47	0.85	0.6	0.75
H3N1	0.67	0.74	0.72	0.93	0.69	0.98	0.49	0.87	0.62	0.75
H3N1	0.64	0.88	0.76	0.85	0.73	0.9	0.46	0.86	0.54	0.74
DS@GT MEDIQA-MAGIC	0.53	0.87	0.66	0.81	0.56	0.89	0.6	0.81	0.65	0.71
DS@GT MEDIQA-MAGIC	0.51	0.84	0.7	0.85	0.56	0.87	0.55	0.81	0.67	0.71
DS@GT MEDIQA-MAGIC	0.47	0.86	0.69	0.85	0.56	0.84	0.51	0.82	0.64	0.69
DS@GT MEDIQA-MAGIC	0.44	0.84	0.69	0.78	0.55	0.86	0.48	0.79	0.65	0.68
DS@GT MEDIQA-MAGIC	0.49	0.82	0.63	0.74	0.56	0.79	0.51	0.75	0.59	0.65
KLE1	0.51	0.63	0.75	0.57	0.63	0.56	0.39	0.74	0.35	0.57
KLE1	0.47	0.62	0.7	0.58	0.62	0.56	0.36	0.76	0.3	0.55
Kasukabe Defense Group	0.44	0.66	0.75	0.28	0.66	0.44	0.52	0.77	0.3	0.54
Kasukabe Defense Group	0.4	0.61	0.73	0.29	0.65	0.44	0.52	0.76	0.33	0.53
Kasukabe Defense Group	0.49	0.49	0.67	0.32	0.48	0.41	0.01	0.76	0.55	0.46
DS@GT MEDIQA-MAGIC	0.31	0.38	0.53	0.31	0.31	0.42	0.01	0.72	0.37	0.37
Oggy	0.08	0.26	0.45	0.3	0.02	0.35	0.02	0.03	0.48	0.22
IReL, IIT(BHU)	0	0.44	0.48	0.17	0.44	0	0.02	0	0	0.17

In the *closed-ended question-answering subtask*, the top two performing teams H3N1 and DSGT employed multi-step architectures, including both fine-tuned models and LLM APIs and ensembling methods. The former divided the task into four parts: (1) preprocessing, (2) information enrichment via image captioning, (3) fine-tuning and external API calls, and (4) ensembling models from the previous step. The latter similarly had several layers: (1) LLM fine-tuning with different models e.g., Qwen and LLAMA, (2) reasoning layer over output of (1) using Gemini, and (3) an agent layer that additionally has a RAG to reference the LanceDB dermatology corpus. In contrast, the remaining groups had similar approaches, which utilized encoders for the images and text. After fusing the text and image features, the resulting vector was passed to a classification layer.

5. Discussion

In the segmentation task, the most successful system was able to use data augmentation generated through image transformation techniques (e.g., color contrast changes). This is promising, as other teams did experiment with skin lesion segmentation specific models; however, they were not able to achieve as high results – suggesting more data would be required to adapt those models. The use of textual inputs was only tested by one group, suggesting that this is an area for future exploration.

Given the unique opportunity allowed by multiple gold annotations and the variety of system outputs, we investigated the effect of using multiple gold references in different scoring schemas on final system rankings. To achieve this, we took the full sample of the test set and ordered the samples randomly, incrementally adding more data until the full test set was covered. Changes in rank were calculated by taking the L1 norm of the difference between the rank at each step and the final ranking. We experimented with three sets of gold standards in which each instance is randomly drawn from one of the gold standards, one gold standard using majority vote by pixel, a gold standard created by taking the intersect and union of all annotators, and a gold standard generated by the STAPLE algorithm which generates an estimated ground truth derived from existing gold standards [17]. We calculated Jaccard and Dice, at a corpus level (e.g., all areas for intersections and unions are added from all instances before calculation), shown in Figure 2. Even at the second-to-last sample point, the rankings from intersect did not agree with other calculations. Interestingly, convergence was observed for rand1 and rand3 by around 200 samples. However, rand2 did not show similar qualities, suggesting this method remains sensitive to anomalies inherent in taking both random samples from gold and from choosing instances. The STAPLE algorithm showed the fastest convergence to its final ranking, suggesting that it is a robust approximation to truth.

We additionally calculated macro evaluations at the instance level (e.g., Jaccard is calculated for

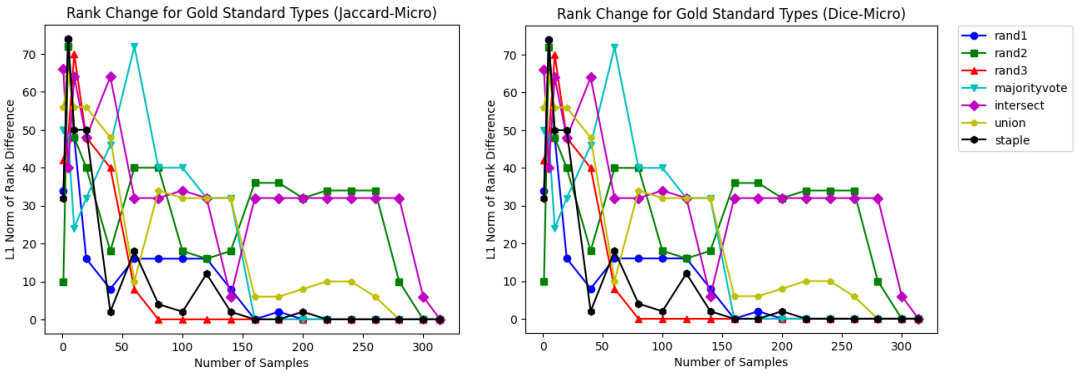


Figure 2: Rank Changes of Number of Samples (Micro)

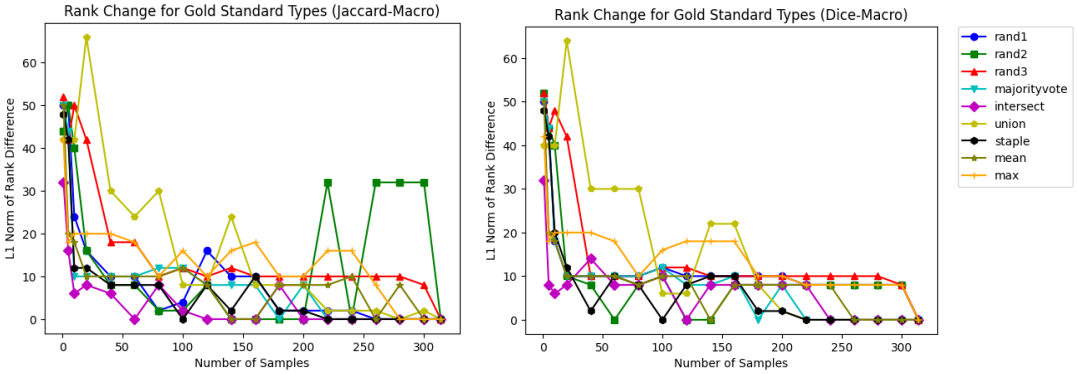


Figure 3: Rank Changes of Number of Samples (Macro)

Table 4
Average Standard Deviation. For reference to the final rankings, the majority vote provides the final Jaccard/Dice scores per each submission. The standard deviations (std) indicate how far each submission's score deviates from the gold standard mean, averaged over all instances.

system	majorityvote		std	
team_name	jaccard	dice	jaccard	dice
Anastasia	0.646	0.785	3.576	4.421
Anastasia	0.646	0.785	3.576	4.421
Anastasia	0.646	0.785	3.576	4.421
Anastasia	0.611	0.759	3.754	4.648
IReL, IIT(BHU)	0.588	0.741	4.131	5.243
KLE1	0.541	0.702	3.84	4.912
KLE1	0.541	0.702	3.84	4.912
KLE1	0.541	0.702	3.84	4.912
KLE1	0.541	0.702	3.84	4.912
H3N1	0.514	0.679	4.245	5.559
H3N1	0.514	0.679	4.245	5.559
H3N1	0.514	0.679	4.245	5.559
H3N1	0.514	0.679	4.245	5.559
Anastasia	0.321	0.485	9.051	12.846
Anastasia	0.313	0.477	9.146	12.971
Kasukabe Defense Group	0.187	0.315	18.882	32.131

each image, then averaged across the dataset). For macro evaluations, we could additionally assess the

effect of taking the mean or maximum Jaccard or dice among all of the gold standard masks available per-instance. As shown in Figure 3. Unlike in micro scoring, intersect converges much faster and union is more prone to fluctuation. Majority vote, on the other hand, exhibits some fluctuations but converges at a similar sample point as with micro scoring. This discrepancy can be attributed to allowing large differences in one or two instances affecting the entire score for micro - whereas in macro calculations, large differences in one instance will not affect more than the weight of one sample. Finally, it is interesting to observe macro-scoring for Jaccard and dice using per instance mean and max values. For both cases, we see that their rankings are often different from that of the other calculations. Here, again the STAPLE algorithm showed the fastest convergence to its final ranking.

Given the multiple gold annotations, it is possible to compute a mean and standard deviation of gold mask instances, wherein the STAPLE-algorithm computed mask is taken as gold. To quantify how often systems' differences are comparable to gold standard differences, we calculate the standard deviations away from gold masks' averages and take the average over all instances. Table 4 provides the average standard deviations for jaccard and dice for all submissions. In general, the systems with the best final scores will have lower standard deviations. The best systems by team Anastasia were at 3.6 standard deviations from the average mean gold mask score; meanwhile the worst-performing submissions were at 9 and 19 standard deviations.

Table 5
Average Differences In Identifying Non-Empty Duplicate Questions. For reference to the final rankings, the accuracy gives the final score per submission. The difference (diff) is the computed average number of answer differences between the gold and the system.

system	accuracy	diff		
		CQID011	CQID012	CQID020
H3N1	0.758	-0.05	-0.1	0.3
H3N1	0.751	-0.04	-0.07	-0.36
H3N1	0.745	-0.61	-0.39	-0.46
H3N1	0.745	-0.61	-0.39	-0.46
H3N1	0.736	-0.1	-0.08	-0.14
DS@GT MEDIQA-MAGIC	0.71	0.38	0.7	1.7
DS@GT MEDIQA-MAGIC	0.706	0.47	0.49	1.31
DS@GT MEDIQA-MAGIC	0.692	0.33	0.42	1.27
DS@GT MEDIQA-MAGIC	0.675	0.54	0.5	1.45
DS@GT MEDIQA-MAGIC	0.653	0.27	0.65	1.17
KLE1	0.57	-0.32	-0.11	-0.37
KLE1	0.553	-0.19	-0.06	-0.34
KLE1	0.553	-0.19	-0.06	-0.34
Kasukabe Defense Group	0.537	-0.18	-0.1	-0.37
Kasukabe Defense Group	0.526	-0.24	-0.12	-0.37
Kasukabe Defense Group	0.464	-0.39	-0.12	-0.37
DS@GT MEDIQA-MAGIC	0.374	3.48	1.87	4.65
Oggy	0.222	-1.39	-1.12	-1.37

In the closed QA task, three out of nine overall questions had duplicate questions allowing for multiple site locations (e.g., "1 where is the affected area", "1 where is the affected area", "2 what label best describes the affected area", "2 what label best describes the affected area", "1 what label best describes the affected area"). For each submission, we calculate the mean differences in the number of unique answers per overall questions. For example, the reference may have QUESTION: 1 where is the affected area, ANSWER: ARM, QUESTION: 2 where is the affected area, ANSWER: LEG, QUESTION: 3 where is the affected area, ANSWER: N/A. Whereas the system may have ANSWER: LEG, QUESTION: 2 where is the affected area, ANSWER: N/A, QUESTION: 3 where is the affected area, ANSWER: N/A. In this case, the difference would be 2. A value less than 1 indicates the system gives a smaller number of answers than the reference on average; a value close to 0 indicates a close agreement on average. Table 5 shows the results. We see that on average the H3N1 team most often provided less answers than

reference; whereas the second top team would provide more answers than reference. Most systems tended to either provide less or more on average consistently across all 3 questions. It is worth noting, the highest performing system had the closest difference (near 0) for all question categories, suggesting that their handling of multiple related questions helped their overall performance. Given the spread of over- and under- answering across submissions, generating the correct number of answers in itself may be challenging.

For the closed QA task, we found the best systems included multiple models fine-tuned for the task as well as some ensembling and aggregation. The use of multi-modal large language models were critically more successful than the suite of fine-tuned multimodal approaches which relied on a shared embedding representation then fine-tuned for the classification task. This could be because the current dataset is relatively small thus the importance of the large language models' access to external information became a determining factor.

6. Conclusion

In this challenge, participants benchmarked the consumer health dermatological image segmentation task as well as the closed VQA task. In general, the performances were promising with segmentation performances at 3.6 standard deviations from gold annotators. Meanwhile, closed QA achieved an accuracy of 0.76.

The best performing segmentation systems took fine-tuning approaches along with data augmentation methods. Here, only one team explored using textual clinical history as input, suggesting that this area can be further explored. In closed VQA, the best performing teams applied multiple models and ensembling methods. Successful applications may need to adapt such steps for proper pre- and post-processing.

Here, we report the benchmarks for our segmentation and closed VQA. Exploring the impact of these subtasks on an end-to-end free text response generation would be an interesting direction for future studies. Future work includes expanding the dataset to capture more dermatological cases and demographics. Furthermore, these technologies should be incorporated into real-world clinical workflows and measured by their ability to increase workflow efficiency.

Declaration on Generative AI

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

References

[1] 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, Spain, 2025.

[2] W. wai Yim, A. B. Abacha, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, Overview of the mediqa-magic task at imageclef 2024: Multimodal and generative telemedicine in dermatology, in: Conference and Labs of the Evaluation Forum, 2024.

[3] W.-w. Yim, A. Ben Abacha, Y. Fu, Z. Sun, F. Xia, M. Yetisgen, M. Krallinger, Overview of the MEDIQA-M3G 2024 shared task on multilingual multimodal medical answer generation, in: T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, D. Bitterman (Eds.), Proceedings of the 6th Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 581–589. URL: https://aclanthology.org/2024.clinicalnlp-1.55/. doi:10.18653/v1/2024.clinicalnlp-1.55.

[4] T. F. Bishop, M. J. Press, J. L. Mendelsohn, L. P. Casalino, Electronic communication improves access, but barriers to its widespread adoption remain 32 (????) 10.1377/hlthaff.2012.1151.

[5] C. A. Sinsky, T. D. Shanafelt, J. A. Ripp, The electronic health record inbox: Recommendations for relief 37 (2024) 4002–4003.

[6] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of imageclef 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.

[7] W. wai Yim, Y. Fu, Z. Sun, A. B. Abacha, M. Yetisgen-Yildiz, F. Xia, Dermavqa: A multilingual visual question answering dataset for dermatology, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2024. URL: https://api.semanticscholar.org/CorpusID:273232728.

[8] 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).

[9] A. D. Karishma Thakrar, Shreyas Basavatia, Ds@gt at mediqa-magic 2025, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[10] N. P. H. Le, H. P. D. Huy, H. T. D. Nhat, H. T. Minh, H3n1 at mediqa-magic 2025: Dermosegdiff and dermkem for comprehensive dermatology ai, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[11] K. B. Desai, V. Hiregoudar, I. Kulkarni, R. Dhane, P. Desai, S. C, U. Mudenagudi, R. Tabib, The kasukabe defense group at mediqa-magic 2025: Clinical visual question answering with resource-efficient multi-modal learning, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[12] T. Le, T. Ngo, K. Nguyen, T. Dang, T. Pham, T. Nguyen, Anastasia at mediqa-magic 2025: A multi-approach segmentation framework with extensive augmentation, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[13] K. Tewari, A. Verma, S. Pal, Irel, iit(bhu) at mediqa-magic 2025: Tackling multimodal dermatology with clipseg-based segmentation and bert-swin question answering, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[14] B. Mallanaikar, S. Kekare, P. Desai, S. C, U. Mudenagudi, R. Tabib, A. Savalkar, A. S. Handi, P. Desai, S. Varur, Kle1 at mediqa-magic 2025, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[15] A. Bozorgpour, Y. Sadegheih, A. Kazerouni, R. Azad, D. Merhof, Dermosegdiff: A boundary-aware segmentation diffusion model for skin lesion delineation, 2023. URL: https://arxiv.org/abs/2308.02959. arXiv:2308.02959.

[16] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, J. Qin, Boundary-Aware Transformers for Skin Lesion Segmentation, Springer International Publishing, 2021, p. 206–216. URL: http://dx.doi.org/10.1007/978-3-030-87193-2_20. doi:10.1007/978-3-030-87193-2_20.

[17] S. Warfield, K. H. Zou, W. M. Wells, Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation, IEEE Transactions on Medical Imaging 23 (2004) 903–921. URL: https://api.semanticscholar.org/CorpusID:3025202.