# ImageCLEF 2025: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications

**Chapter** · April 2025

DOI: 10.1007/978-3-031-88720-8_60



**47 authors**, including:

Bogdan Ionescu
Universitatea Națională de Știință și Tehnologie Politehnica București
**252** PUBLICATIONS **3,238** CITATIONS

SEE PROFILE

Henning Müller
HES-SO Valais-Wallis
**1,001** PUBLICATIONS **27,452** CITATIONS

SEE PROFILE

Ahmad Idrissi-Yaghir
University of Applied Science and Arts Dortmund
**36** PUBLICATIONS **216** CITATIONS

SEE PROFILE

Radjabov Ahmedkhan
United Institute Of Informatics Problems
**16** PUBLICATIONS **46** CITATIONS

SEE PROFILE

# ImageCLEF 2025: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications

⋆ Bogdan Ionescu[1], Henning Müller[2], Dan-Cristian Stanciu[1], Ahmad Idrissi-Yaghir[3], Ahmedkhan Radzhabov[15], Alba García Seco de Herrera[4], Alexandra Andrei[1], Andrea Storås[11], Asma Ben Abacha[5], Benjamin Bracke[3], Benjamin Lecouteux[6], Benno Stein[7], Cécile Macaire[6], Christoph M. Friedrich[3], Cynthia Sabrina Schmidt[8], Diandra Fabre[6], Didier Schwab[6], Dimitar Dimitrov[9] and Emmanuelle Esperança-Rodier[6], Gabriel Constantin[1], Helmut Becker[8] and Hendrik Damm[3] and Henning Schäfer[8], Ivan Rodkin[12], Ivan Koychev[9], Johannes Kiesel[7], Johannes Rückert[3], Josep Malvehy[16], Liviu-Daniel Ștefan[1], Louise Bloch[3], Martin Potthast[10], Maximilian Heinrich[7], Michael A. Riegler[11], Mihai Dogariu[1], Noel Codella[5], Pål Halvorsen[11], Preslav Nakov[12] Raphael Brüngel[3], Roberto Andres Novoa[17], Rocktim Jyoti Das[13] Steven A. Hicks[11], Sushant Gautam[11], Tabea M. G. Pakull[8], Vajira Thambawita[11], Vassili Kovalev[14,15], Wen-Wai Yim[5], and Zhuohan Xie[12]

[1] National University of Science and Technology Politehnica Bucharest, Romania
[2] University of Applied Sciences Western Switzerland (HES-SO), Switzerland
[3] University of Applied Sciences and Arts Dortmund, Germany
[4] Universidad Nacional de Educación a Distancia (UNED), Spain
[5] Microsoft, USA
[6] Univ. Grenoble Alpes, CNRS, Grenoble INP*, Grenoble, France
[7] Bauhaus-Universität Weimar, Germany
[8] University Hospital Essen, Germany
[9] Sofia University, Bulgaria
[10] University of Kassel, hessian.AI, and ScaDS.AI, Germany
[11] SimulaMet, Norway
[12] Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
[13] Indian Institute of Technology, Delhi
[14] Belarus National Academy of Sciences, Belarus
[15] Belarus State University, Belarus
[16] Hospital Clinic of Barcelona, Spain
[17] Stanford University, USA

**Abstract.** ImageCLEF has been a part of CLEF (Conference and Labs of the Evaluation Forum) for more than 20 years. Started in 2003, ImageCLEF is an evaluation initiative which promotes the evaluation of technologies for annotation, indexing, retrieval, or generation of multimodal data. It provides access to large amounts of challenging data in very diverse use cases like medicine, argumentation, reasoning, generation, or content recommendation. In its 23rd edition, ImageCLEF will

⋆ apart from the general organisers, the authors are listed in alphabetical order.

have four main tasks: (i) a *Medical* task involving concept detection and caption prediction in radiology images, synthetic medical images created with Generative Adversarial Networks (GANs), Visual Question Answering for improving the diagnosis and classification of real medical gastrointestinal images, and multimodal dermatology response generation, (ii) a joint ImageCLEF-Touché task *Image Retrieval/Generation for Arguments* to convey the premise of an argument, (iii) the *ToPicto* task which involves converting either text or speech into a meaningful sequence of pictograms and (iv) a new Multimodal Reasoning task addressing question answering and reasoning generation. In its last edition in 2024, 90 users and 31 unique teams submitted runs, totaling 257 runs, revealing a good impact in the community, similar to previous years.

**Keywords:** Medical AI, image captioning, GANs, Visual Question Answering, response generation, argumentation

## 1 Introduction

Since 2003, the ImageCLEF evaluation campaign has provided the scientific community a space for research and benchmark activities in fields like annotation, indexing, classification and retrieval of multimodal data. The 23rd edition of ImageCLEF takes place in Madrid, Spain, 9-12 September 2025[18], as part of the Conference and Labs of the Evaluation Forum (CLEF) [16, 17]. A series of benchmarking activities have been developed to evaluate various aspects of mono-language and cross-language information retrieval systems [13, 16, 17]. Some of the fields of interest for ImageCLEF include information retrieval in multimedia, machine learning, deep learning, data mining, computer vision, natural language processing, data generation, and multimodal data fusion. Over the years, both ImageCLEF and the CLEF campaign have demonstrated significant scholarly impact [26, 27] and impact in the community overall, with the term ImageCLEF mentioned in 7,330 papers since 2003, according to Google Scholar. This paper presents the four tasks planned for 2025. Some examples can be seen in Fig. 1

## 2 ImageCLEFmedical

The ImageCLEFmedical task is currently at its 21st edition [17]. The 2025 edition introduces 4 medical sub-tasks, namely: (i) the *Caption* task with medical concept detection and caption prediction, (ii) the *GAN* task on synthetic medical images, (iii) *MEDVQA* regarding Visual Question Answering for gastrointestinal data, and (iv) *Mediqa-magic* with a new use-case on multimodal dermatology response generation.

*ImageCLEFmedical-Caption*[19]. The *Caption* task consists of the interpretation of the insights gained from radiology images, and in this 9th edition [7, 8,

[18] https://clef2025.clef-initiative.eu//
[19] https://www.imageclef.org/2025/medical/caption

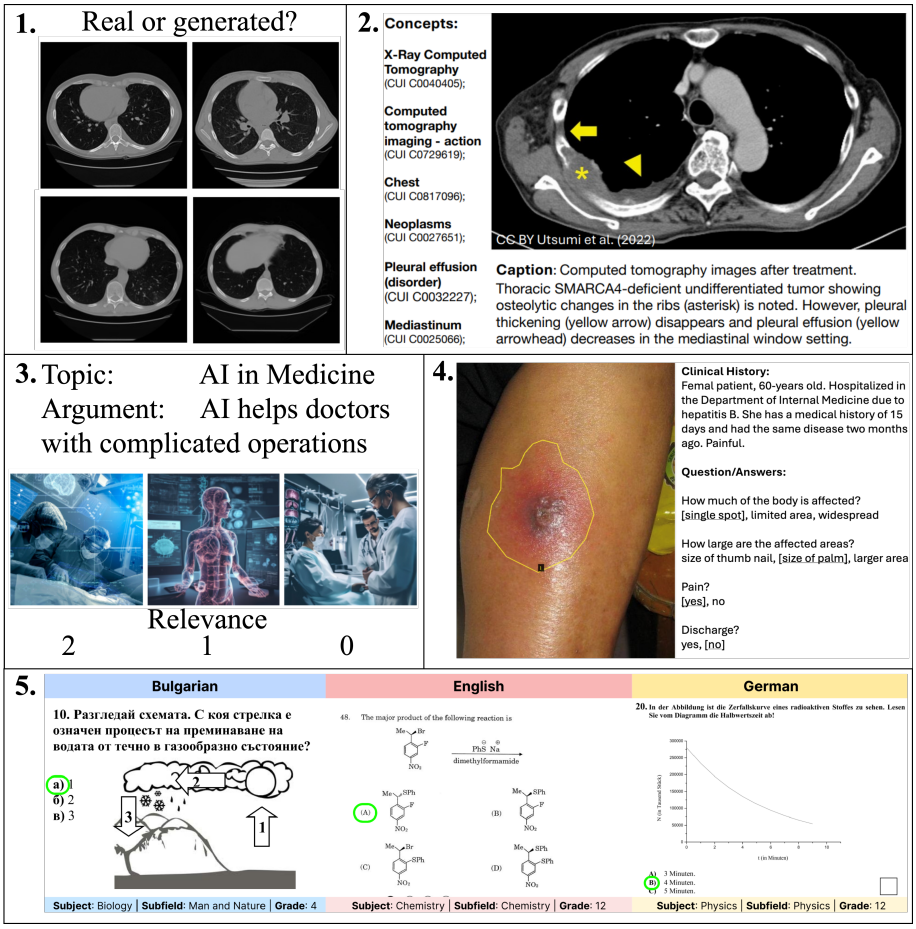


**Fig. 1.** Sample images from (1) ImageCLEFmedical-GAN, with real and generated images, (2) ImageCLEFmedical-Caption, presenting an image with associated concepts and captions, (3) Image Retrieval/Generation for Arguments, showing an argument with possible image submissions, (4) MEDIQA-MAGIC, examples of segmentation and question answers, (5) MultimodalReason, with examples of questions and answers in different languages and at different levels.

18–23], there are two subtasks: concept detection and caption prediction. The *concept detection* subtask aims to develop competent systems that are able to predict the Unified Medical Language System (UMLS®) Concept Unique Identifiers (CUIs) based on the visual image content. The F1-Score [10] will be used for evaluation. This year, the updated ROCOv2 [24] dataset will be used.

In the *caption prediction* subtask, domain-specific models such as BioBERT [14] are utilised to enhance the evaluation using BERTScore. Furthermore, experiments are conducted with other models, including ClinicalBLEURT and Med-

BERTScore [3], to explore alternative scoring methods. Readability scores such as LENS [15] and the Coleman-Liau Index [4] (CLI) are also incorporated to assess the clarity and readability of the generated captions. An explainability extension is offered for both tasks, where participants are asked to provide explanations, such as heatmaps and Shapley values, for a small subset of images that will be manually evaluated.

*ImageCLEFmedical-GAN*[20]. The 3rd edition of the GANs task will continue to explore the quality of synthetic medical images generated by various models, including Generative Adversarial Networks (GANs). Building on the previous two editions [1,2], the sub–tasks will address both the challenges of ensuring privacy and confidentiality in medical imagining and the quality of the generated data itself. The focus remains on investigating potential risks related to data's integrity and the potential for sensitive information to be embedded within the generated images. The sub–task involves analyzing test image datasets to evaluate the likelihood that specific real patient images were used to train the generative models. The provided data will include both real and generated axial slices of CT scans from tuberculosis patients, captured at various stages of their treatment. Metrics like Accuracy, F1 score will be used to determine if specific real images were used in the training of generative models or to identify the particular dataset of real images that contributed to the training of the generative model responsible for creating each synthetic image.

*ImageCLEFmedical-MEDVQA*[21]. The 3rd MEDVQA-GI challenge focuses on the integration of Visual Question Answering (VQA) with synthetic gastrointestinal (GI) data to enhance diagnostic accuracy and AI learning. Building on the previous two editions [11, 12], this year's challenge asks participants to complete two subtasks. The first subtask involves developing algorithms to accurately interpret and respond to questions related to GI images. The aim is for these algorithms to understand the context and details in the images and provide precise answers that assist in medical diagnostics. The second subtask focuses on generating high-fidelity synthetic GI images that closely resemble real medical images. The dataset for this year is an extended version of the collections used in the last two years, including GI images with VQA annotations [9]. The evaluation will differ for each task, which will be a combination of subjective and objective evaluations.

*ImageCLEFmedical-MEDIQA-MAGIC*[22]. In the 2nd MEDIQA-MAGIC task, we will extend on the previous year's dataset and challenge based on multimodal dermatology response generation. Participants will be given a clinical narrative context along with accompanying images. The task is divided into two relevant sub-parts: (i) segmentation of dermatological problem regions, and (ii) providing answers to closed-ended questions.

In the first sub-task, given each image and the clinical history, participants will need to generate segmentations of the regions of interest for the described

[20] https://www.imageclef.org/2025/medical/gans
[21] https://www.imageclef.org/2025/medical/vqa
[22] https://www.imageclef.org/2025/medical/mediqa

dermatological problem. In the second sub-task, participants will be given a dermatological query, its accompanying images, as well as a closed-question with accompanying choices – the task is to select the correct answer to each question.

The dataset is created by using real consumer health users' queries and images; the question schema was created by two certified dermatologists. Segmentation will be evaluated against common metrics such as Jaccard or IOU. Closed question-answering will be evaluated using metrics such accuracy and F1 score.

## 3 Image Retrieval/Generation for Arguments

*Touché-Argument-Images*[23] Images can drastically increase the impact of arguments and make them much more compelling. The 2025 edition marks the fourth edition of the task, having run at the Touché workshop since 2022. For this year's edition, we have streamlined the task and focus on complete arguments. Given an argument, the participants shall find suitable images that convey the argument. Optionally, participants can add captions to explain the meaning of the submitted images. An example of an argument with possible image submissions can be seen in Fig. 1(3).

Participants can choose between two submission styles. (1) Retrieval: We deliver matching images from a focused crawl of around 10,000 images, together with website data and other data such as automatically generated image descriptions. (2) Prompted Generation: Based on the idea of the infinite index [6], participants can submit prompts for an image generator. A dedicated API is provided for this purpose. The submitted images are judged by human experts according to defined relevance criteria.

## 4 ImageCLEFtoPicto

In its 2nd edition, *ImageCLEFtoPicto* [24] involves converting either text or speech into a meaningful sequence of pictograms to be used as a communication aid. This year, the task will be split into 2 sub-tasks: (1) Text-to-Picto Translation, which involves the generation of a sequence of pictogram terms from a French text, using oral transcriptions from the ESLO corpus to create pictogram sequences, (2) Speech-to-Picto translation, a task that focuses on translating speech directly into pictogram sequences without going through the transcription. Using the ESLO corpus, participants will handle speech utterances and generate corresponding pictogram terms, challenging them to integrate speech recognition and translation technologies. This year, the ESLO Corpus [25] will be integrated to enhance the evaluation campaign, focusing specifically on the ESLO 2 sub-corpus, which contains contemporary interactions and offers modern language usage and a wide range of conversational contexts. For evaluation,

[23] https://touche.webis.de/clef25/touche25-web/image-retrieval-generation-for-arguments.html

[24] https://www.imageclef.org/2025/topicto

both automatic metrics like Picto-term Error Rate (PictoER) and the BLEU score will be used, as well as human expert evaluation, based on availability.

## 5 MultimodalReason

*MultimodalReason* is a new task focusing on Multilingual Visual Question Answering (VQA). Given an image of a question with four possible answers, participants must identify the single correct answer, as shown in Fig. 1(4). Furthermore, we provide OCR-extracted text and captions for each questions (but participants can extract their own). The questions available are in 11 different languages, across 20 school disciplines and different difficulty levels. The dataset for the task, called *Exams-V* [5], is publicly available[25]. The goal of the task is to assess the reasoning capabilities of modern LLMs on complex inputs, presented in different languages, across various subjects. The official evaluation measure for the task will be *accuracy*. In addition to the existing dataset, we will provide the participants new unseen test data that follow the variety of languages, subjects, and difficulties of *Exams-V*.

## 6 Conclusions

This paper showcases the tasks proposed for the 23rd edition of the ImageCLEF evaluation campaign, in fields like medicine, data generation, multimodal data processing, data retrieval, natural language processing, just to name a few. ImageCLEF continues to provide very diverse and interesting challenges for researchers, giving them the possibility to assess the performance of their systems with a mix of both established and new tasks. All the tasks provide collections of test data end evaluation benchmarks, enabling researchers to assess the performance of their systems using realistic evaluation frameworks.

**Acknowledgement.**

The work of Louise Bloch, Benjamin Bracke and Raphael Brüngel was partially funded by a PhD grant from the University of Applied Sciences and Arts Dortmund (FH Dortmund), Germany. The work of Hendrik Damm, Tabea M. G. Pakull, Helmut Becker, Ahmad Idrissi-Yaghir and Henning Schäfer was funded by a PhD grant from the DFG Research Training Group 2535 Knowledge- and data-based personalisation of medicine at the point of care (WisPerMed). Image Retrieval/Generation for Arguments task was partially supported by the European Commission under grant agreement GA 101070014 (OpenWebSearch.eu). The work of D. Dimitrov and I. Koychev is partially financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No BG-RRP-2.004-0008. The work of Bogdan Ionescu, Dan-Cristian Stanciu, Alexandra Andrei, Liviu-Daniel Stefan, Gabriel Constantin is funded by UEFISCDI, Solutions Axis, grant SOL4/2024.

[25] https://huggingface.co/datasets/Rocktim/EXAMS-V/tree/main/data

## References


1. Andrei, A., Radzhabov, A., Karpenka, D., Prokopchuk, Y., Kovalev, V., Ionescu, B., Müller, H.: Overview of 2024 imageclefmedical gans task–investigating generative models' impact on biomedical synthetic images. In: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Grenoble, France (2024)
2. Andrei, A.G., Radzhabov, A., Coman, I., Kovalev, V., Ionescu, B., Müller, H.: Overview of imageclefmedical gans 2023 task: identifying training data "fingerprints" in synthetic biomedical images generated by gans for medical image security. In: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023). vol. 3497 (2023)
3. Ben Abacha, A., Yim, W.w., Michalopoulos, G., Lin, T.: An investigation of evaluation methods in automatic medical note generation. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 2575–2588. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.161, https://aclanthology.org/2023.findings-acl.161
4. Coleman, M., Liau, T.L.: A computer readability formula designed for machine scoring. **60**(2), 283–284. https://doi.org/10.1037/h0076540, 515 citations (Crossref/DOI) [2024-10-16] 515 citations (Crossref/DOI) [2024-10-16]
5. Das, R., Hristov, S., Li, H., Dimitrov, D., Koychev, I., Nakov, P.: EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7768–7791. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.420, https://aclanthology.org/2024.acl-long.420
6. Deckers, N., Fröbe, M., Kiesel, J., Pandolfo, G., Schröder, C., Stein, B., Potthast, M.: The Infinite Index: Information Retrieval on Generative Text-To-Image Models. In: Gwizdka, J., Rieh, S.Y. (eds.) ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2023). pp. 172–186. ACM (Mar 2023). https://doi.org/10.1145/3576840.3578327, https://doi.org/10.1145/3576840.3578327
7. Eickhoff, C., Schwall, I., García Seco de Herrera, A., Müller, H.: Overview of ImageCLEFcaption 2017 – the image caption prediction and concept extraction tasks to understand biomedical images. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2017). CEUR Workshop Proceedings, vol. 1866. CEUR-WS.org (2017)
8. García Seco De Herrera, A., Eickhof, C., Andrearczyk, V., Müller, H.: Overview of the ImageCLEF 2018 caption prediction tasks. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2018). CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018)
9. Gautam, S., Storås, A.M., Midoglu, C., Hicks, S.A., Thambawita, V., Halvorsen, P., Riegler, M.A.: Kvasir-VQA: A Text-Image Pair GI Tract Dataset. In: ACM Conferences, pp. 3–12. Association for Computing Machinery, New York, NY, USA (Oct 2024). https://doi.org/10.1145/3689096.3689458
10. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: Advances in Information Retrieval – 27th European Conference on IR Research (ECIR 2005). pp. 345–359. Springer (2005)

11. Hicks, S.A., Storås, A., Halvorsen, P., de Lange, T., Riegler, M.A., Thambawita, V.: Overview of ImageCLEFmedical 2023 – Medical Visual Question Answering for Gastrointestinal Tract. In: CLEF2023 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 18-21 2023)
12. Hicks, S.A., Storås, A., Halvorsen, P., Riegler, M.A., Thambawita, V.: Overview of ImageCLEFmedical 2024 – Medical Visual Question Answering for Gastrointestinal Tract. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Gernoble, France (September 9-12 2024)
13. Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems: Overview of the medical image retrieval task at ImageCLEF 2004–2014. Computerized Medical Imaging and Graphics **39**(0), 55 – 61 (2015)
14. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinform. **36**, 1234 – 1240 (2020). https://doi.org/10.1093/bioinformatics/btz682
15. Maddela, M., Dou, Y., Heineman, D., Xu, W.: LENS: A learnable evaluation metric for text simplification. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 16383–16408. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.905, https://aclanthology.org/2023.acl-long.905, 4 citations (Crossref/DOI) [2024-10-16] 4 citations (Crossref/DOI) [2024-10-16] 2 citations (Crossref/DOI) [2024-09-11]
16. Müller, H., Clough, P., Deselaers, T., Caputo, B. (eds.): ImageCLEF – Experimental Evaluation in Visual Information Retrieval, The Springer International Series On Information Retrieval, vol. 32. Springer, Berlin Heidelberg (2010)
17. Müller, H., Kalpathy-Cramer, J., García Seco de Herrera, A.: Experiences from the ImageCLEF medical retrieval and annotation tasks. In: Information Retrieval Evaluation in a Changing World, pp. 231–250. Springer (2019)
18. Pelka, O., Abacha, A.B., García Seco de Herrera, A., Jacutprakart, J., Friedrich, C.M., Müller, H.: Overview of the ImageCLEFmed 2021 concept & caption prediction task. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2021). CEUR Workshop Proceedings, vol. 2936. CEUR-WS.org (2021)
19. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2019 concept detection task. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2019). CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org (2019)
20. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2020 concept prediction task: Medical image understanding. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2020). CEUR Workshop Proceedings, vol. 2696. CEUR-WS.org (2020)
21. Rückert, J., Ben Abacha, A., G. Seco de Herrera, A., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Bracke, B., Damm, H., Pakull, T.M.G., Schmidt, C.S., Müller, H., Friedrich, C.M.: Overview of ImageCLEFmedical 2024 – Caption Prediction and Concept Detection. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France (September 9-12 2024)
22. Rückert, J., Ben Abacha, A., García Seco de Herrera, A., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Müller, H., Friedrich, C.M.: Overview of ImageCLEFmedical 2022 – Caption Prediction and Concept Detection. In: CLEF2022 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy (September 5-8 2022)

23. Rückert, J., Ben Abacha, A., G. Seco de Herrera, A., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Müller, H., Friedrich, C.M.: Overview of ImageCLEFmedical 2023 – Caption Prediction and Concept Detection. In: CLEF2023 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 18-21 2023)
24. Rückert, J., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Schmidt, C.S., Koitka, S., Pelka, O., Abacha, A.B., de Herrera, A.G.S., Müller, H., Horn, P.A., Nensa, F., Friedrich, C.M.: Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. Scientific Data **11**(1), 688 (June 2024). https://doi.org/10.1038/s41597-024-03496-6, https://doi.org/10.1038/s41597-024-03496-6
25. Serpollet, N., Bergounioux, G., Chesneau, A., Walter, R.: A large reference corpus for spoken french: Eslo 1 and 2 and its variations. In: Proceedings from Corpus Linguistics Conference Series, University of Birmingham (2007)
26. Tsikrika, T., García Seco de Herrera, A., Müller, H.: Assessing the scholarly impact of ImageCLEF. In: CLEF 2011. pp. 95–106. Springer Lecture Notes in Computer Science (LNCS) (2011)
27. Tsikrika, T., Larsen, B., Müller, H., Endrullis, S., Rahm, E.: The scholarly impact of CLEF (2000–2009). In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 1–12. Springer (2013)