Advancing Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications with ImageCLEF 2024

38 authors, including:

Bogdan Ionescu
Polytechnic University of Bucharest
240 PUBLICATIONS   **2,792** CITATIONS

SEE PROFILE

Ana Maria Claudia Dragulinescu
Polytechnic University of Bucharest
61 PUBLICATIONS   **360** CITATIONS

SEE PROFILE

Henning Müller
HES-SO Valais-Wallis
962 PUBLICATIONS   **23,233** CITATIONS

SEE PROFILE

Ahmad Idrissi-Yaghir
University of Applied Science and Arts Dortmund
23 PUBLICATIONS   **88** CITATIONS

SEE PROFILE

Advancing Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications with ImageCLEF 2024

⋆ Bogdan Ionescu[1], Henning Müller[2], Ana Maria Drăgulinescu[1], Ahmad Idrissi-Yaghir[4], Ahmedkhan Radzhabov[16], Alba Garcia Seco de Herrera[5], Alexandra Andrei[1], Alexandru Stan[6], Andrea M. Storås[7], Asma Ben Abacha[8], Benjamin Lecouteux[19], Benno Stein[17], Cécile Macaire[19], Christoph M. Friedrich[4], Cynthia Sabrina Schmidt[10], Didier Schwab[19], Emmanuelle Esperança-Rodier[19], George Ioannidis[6], Griffin Adams[9], Henning Schäfer[10], Hugo Manguinhas[11], Ioan Coman[1], Johanna Schöler[13], Johannes Kiesel[17] Johannes Rückert[4], Louise Bloch[4], Martin Potthast[18], Maximilian Heinrich[17], Meliha Yetisgen[14], Michael A. Riegler[7], Neal Snider[15], Pål Halvorsen[7], Raphael Brüngel[4], Steven A. Hicks[7], Vajira Thambawita[7], Vassili Kovalev[16,12], Yuri Prokopchuk[16], and Wen-Wai Yim[8]

[1] National University of Science and Technology Politehnica Bucharest, Romania
alexandra.andrei@upb.ro
[2] University of Applied Sciences Western Switzerland (HES-SO), Switzerland
[3] CEA LIST, France
[4] University of Applied Sciences and Arts Dortmund, Germany
[5] University of Essex, UK
[6] IN2 Digital Innovations, Germany
[7] SimulaMet, Norway
[8] Microsoft, USA
[9] Columbia University, USA
[10] University Hospital Essen, Germany
[11] Europeana Foundation, Netherlands
[12] Belarus State University, Belarus
[13] Sahlgrenska University Hospital, Sweden
[14] University of Washington, USA
[15] Microsoft/Nuance, USA
[16] Belarus National Academy of Sciences, Belarus
[17] Bauhaus-Universität Weimar, Germany
[18] Leipzig University and ScaDS.AI, Germany
[19] Univ. Grenoble Alpes, CNRS, Grenoble INP*, Grenoble, France

Abstract. The ImageCLEF evaluation campaign was integrated with CLEF (Conference and Labs of the Evaluation Forum) for more than 20 years and represents a Multimedia Retrieval challenge aimed at evaluating the technologies for annotation, indexing, and retrieval of multimodal data. Thus, it provides information access to large data collections in usage scenarios and domains such as medicine, argumentation

---

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

and content recommendation. ImageCLEF 2024 has four main tasks: (i) a *Medical* task targeting automatic image captioning for radiology images, synthetic medical images created with Generative Adversarial Networks (GANs), Visual Question Answering and medical image generation based on text input, and multimodal dermatology response generation; (ii) a joint ImageCLEF-Touché task *Image Retrieval/Generation for Arguments* to convey the premise of an argument, (iii) a *Recommending* task addressing cultural heritage content-recommendation, and (iv) a joint ImageCLEF-ToPicto task aiming to provide a translation in pictograms from natural language. In 2023, participation increased by 67% with respect to 2022 which reveals its impact on the community.

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

1   Introduction

With a tradition of more than 20 years, the ImageCLEF benchmarking campaign provides the scientific community with research activities and evaluation of approaches for annotation, indexing, classification and retrieval of multimodal data. ImageCLEF 2024 is integrated with the Conference and Labs of the Evaluation Forum (CLEF) [18,19], with the 22nd edition being hosted by University of Grenoble Alpes, France, 9–12 September 2024[20]. Considering the experience of the last four successful editions, ImageCLEF 2024 will handle a diversity of applications within the four benchmarking tasks approaching different aspects of mono- and cross-language information retrieval systems [14,18,19] related to the interpretation of the radiology images [25], and testing the hypothesis that the artificial biomedical images contain fingerprints of the original images [6,12], to name a few. The campaign targets multimodal data annotation and retrieval community and researchers from computer vision, image information retrieval and digital image processing fields. From its inception, ImageCLEF demonstrated a meaningful scholarly impact and, currently, there are 420 publications with 3792 citations on Web of Science (WoS). The paper introduces the four tasks planned for 2024, namely: ImageCLEFmedical, ImageCLEFrecommeding, Image Retrieval/Generation for Arguments, and ImageCLEFtoPicto (Fig. 1).

2   ImageCLEFmedical

ImageCLEFmedical task is currently at its 20th edition [19]. The 2024 edition will continue all the medical sub-tasks in 2023, namely: (i) *caption* task with medical concept detection and caption prediction, (ii) *GAN* task on synthetic medical images generated with GANs, (iii) *MEDVQA-GI* task for medical images generation based on text input, (iv) *Mediqa* task with a new use-case on multimodal dermatology response generation.
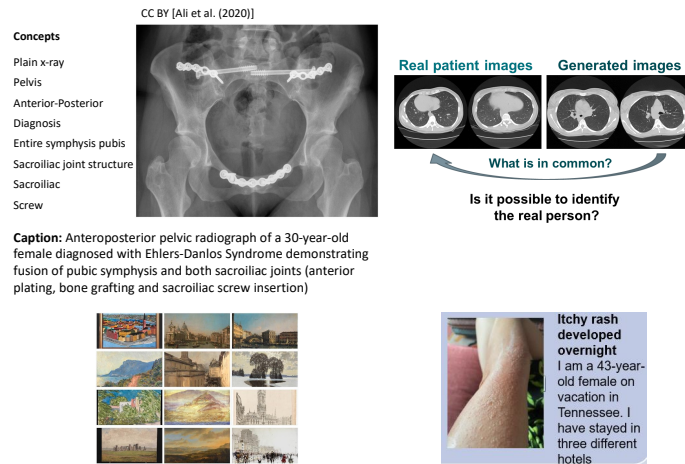
---

[20] https://clef2024.imag.fr/

Fig. 1. Sample images from (left to right, top to bottom): ImageCLEFmedical-caption with an image with the corresponding CUIs and caption, ImageCLEFmedical-GAN with examples of real and generated images, ImageCLEFrecommending with an example of editorial "European landscapes and landmarks" Gallery, and ImageCLEF-Mediqa with example of medical details related to a skin problem and the associated image.

ImageCLEFmedical-caption[21]. The *caption* task consists in the interpretation of the insights gained from radiology images and in this 8th edition [8, 10, 21–23, 27, 28], there are also 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 [11] will be used for evaluation. The *caption prediction* subtask focuses on implementing models to predict captions for given radiology images. To improve the evaluation quality, BERTScore will be refined by integrating domain-specific models such as BioBERT [15]. The scope of scoring may be expanded experimenting with other models, such as ClinicalBLEURT and MedBERTScore [3]. In 2024, an updated version of the Radiology Objects in Context (ROCO) [24] dataset will be used, further extended with new PubMed images. As a novelty in the 8th edition, an optional, experimental explainability extension will be offered for both tasks, and participants will be asked to provide explanations (e.g., heatmaps, Shapley values) for a small subset of images manually evaluated.

ImageCLEFmedical-GAN[22]. The *GANs* task is a relatively new challenge in the ImageCLEFmedical track. We will continue with the first task proposed in the previous edition focused on examining the existing hypothesis that GANs generate medical images containing certain "fingerprints" of the authentic images used for generative network training. The task supposes performing analysis of

---

[21] https://www.imageclef.org/2024/medical/caption
[22] https://www.imageclef.org/2024/medical/gans

test image datasets and assessing the probability with which certain images of real patients were used for training image generators. The participants will test the hypothesis on two different levels, including identifying the source dataset used for training and exploring the problem of detection and isolation of image regions in generated images that inherit the patterns presented in the original ones. The second task explores the hypothesis that generative models imprint unique fingerprints on generated images and whether different generative models or architectures leave discernible signatures within the synthetic images they produce. Similar to the previous year, the 2D gray-scale images being provided depict the axial slices of CT scans of tuberculosis patients taken at different stages of their treatment. In 2024, we continue to use the advanced Diffuse Models [17] along with other generative models for image generation.

ImageCLEFmedical-MEDVQA-GI[23]. In the 2nd MEDVQA-GI challenge, the participants have to generate medical images based on text input, along with optimal prompts for off-the-shelf generative models to improve the diagnosis and classification of real medical images using AI-generated images. The dataset is built up on the one collected in the first edition of MEDVQA-GI [13].The task is divided into two subtasks: i) Image Synthesis (IS) requiring participants to leverage text-to-image generative models to create a rich dataset of medical images derived from textual prompts; 2) Optimal Prompt Generation (OPG) asking to generate optimal textual prompts to guide an off-the-shelf generative model in creating realistic medical images with imaging modalities ranging from magnetic resonance imaging (MRI) and computerised tomography (CT) scans to endoscopic images of various medical conditions. The subjective evaluation is made by medical experts and on how accurately a model trained on these AI-generated images can classify real medical images. The metrics employed are the Fréchet Inception Distance (FID) and standard metrics such as accuracy, precision, recall, and F1 score in both single-center and multi-center datasets.

ImageCLEFmedical-mediqa[24]. The *MEDIQA-magic* task is a new task focusing on multimodal dermatology response generation. Participants are given a clinical narrative context along with accompanying images. They have to generate a textual response answering the health question based on the described history and images. Examples of responses include diagnosis of the problem and suggestions for treatment. The dataset is created by using real consumer health users' queries and images; gold standard responses are composed by certified medical doctors. Response output will be evaluated using metrics such as ROUGE, BLEURT, and BERTScore.

3   Image Retrieval/Generation for Arguments

Touché-Argument-Image[25] gives a set of arguments asking to return several images for each argument, helping to convey the argument's premise. Participants

---

[23] https://www.imageclef.org/2024/medical/vqa

[24] https://www.imageclef.org/2024/medical/mediqa

[25] https://touche.webis.de/clef24/touche24-web/image-retrieval-for-arguments.html

Table 1. Submissions for the argument. The first (retrieved) image could help to convey the "forget"-part of the premise but does not relate to voting. The second image (generated) is rated higher on relevance (1 vs. 2). The third image (generated) does not indicate that someone forgets their ID or is barred from voting (irrelevant, 0).

Topic:	Photo identification at polling stations
Claim:	Legislation to impose restrictive photo ID requirements has the potential to block millions of American voters.
Premise:	People will forget their ID cards and be denied their right to vote.

Images:			
Rationale Relevance:	1	2	0
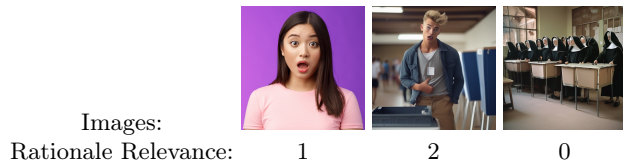
can optionally add a short rationale explaining the meaning of the image and can employ both image retrieval and image generation approaches. We provide a conclusion and a topic as context for the premise (see Table. 1 for an example).

There are three kinds of submissions: (1) Retrieval: Participants can retrieve suitable images from a focused crawl, where we also provide automatically recognised text from the image and text from web pages containing the image. (2) Prompted Generation: Following the idea of the infinite index [7], participants can submit prompts for the Stable Diffusion image generator [26] (we provide a stable API). (3) Direct: Participants can employ other reproducible methods to generate images (e.g., chart generators) to submit them directly. The task follows the classic TREC-style methodology with ranked results judged by human assessors. The performance will be evaluated with nDCG.

4 ImageCLEFrecommending

ImageCLEFrecommending[26] is a task which focuses on content-recommendation for cultural heritage content. Despite current advances in content-based recommendation systems, there is limited understanding of how well these perform and how relevant they are for the final end-users. This task aims to fill this gap by benchmarking different recommendation systems and methods. The task builds upon a key infrastructure for researchers and heritage professionals, namely Europeana [9] and requires participants to devise recommendation methods and systems, apply them in the supplied dataset gathered from Europeana and provide a series of recommendations for items and editorials within two sub-tasks: (i) given a list of items, provide a list of recommended items; (ii) given an editorial (Europeana blog or gallery), provide a list of recommended editorials. A new dataset based on Europeana items and editorials will be provided to the participants including metadata based on the Europeana Data Model schema.

---

[26] https://www.imageclef.org/2024/recommending

Performance will be evaluated based on the recommendations that are provided computing Mean Average Precision at X (Map@X) compared to the ground truth. Moreover, because black-box systems make it difficult for users to assess why the recommendation should be trusted, the systems providing explanation for the results will be awarded additional point.

5   ImageCLEFtoPicto

ImageCLEFtoPicto introduces two new tasks whose objective is to provide a translation in pictograms from a natural language, either from (i) text or (ii) speech understandable by the users, in this case, people with language impairments. ImageCLEFtoPicto is an opportunity for the research community who works in the field of Augmentative and Alternative Communication (AAC), and pictogram translation to gather around two challenging tasks. The task is divided into two subtasks: (i) text-to-pictogram translation focuses on automatically generating a corresponding sequence of pictogram terms from a French text, and (ii) speech-to-pictogram focuses on the two modalities of speech and pictograms. The data of the tasks are taken from Traitement de Corpus Oraux en Français (TCOF) [1], a French speech corpus. The challenge is to directly translate speech to pictogram terms without going through the transcription dimension, which is the focus of the speech community with current spoken language translation systems [4,5]. Both automatic (BLEU [20], METEOR [2], WER [29]) and human evaluation (MQM framework [16]) will be carried out by experts in the domain.

6   Conclusions

The paper highlights the tasks proposed by the 22nd edition of ImageCLEF evaluation campaign. The tasks are refreshed and the participants will meet new challenging use-cases as image retrieval/generation for argumentation or generation of optimal textual prompts to guide an off-the-shelf generative model to create realistic medical images. Moreover, the two joint tasks with Touché and ToPicto increase to a larger extent the diversity of the datasets and metrics which are shared with the community, whereas tasks such as Fusion and Aware will be discontinued. Overall, ImageCLEF2024 will continue to provide the researchers with the possibility to assess the performance of their conceived systems having access to a common evaluation framework to compare their results.

References

1. André, V., Canut, E.: Mise à disposition de corpus oraux interactifs: le projet tcof (traitement de corpus oraux en français). Pratiques. Linguistique, littérature, didactique (147-148), 35–51 (2010)
2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
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. Bérard, A., Besacier, L., Kocabiyikoglu, A.C., Pietquin, O.: End-to-end automatic speech translation of audiobooks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6224–6228. IEEE (2018)
5. Bérard, A., Pietquin, O., Besacier, L., Servan, C.: Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In: NIPS Workshop on end-to-end learning for speech and audio processing. Barcelona, Spain (Dec 2016), https://hal.science/hal-01408086
6. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., Wallace, E.: Extracting training data from diffusion models (2023)
7. 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
8. 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)
9. Europeana Foundation: Europeana (2022), https://www.europeana.eu/
10. 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)
11. 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)
12. Gui, J., Sun, Z., Wen, Y., Tao, D., Ye, J.: A review on generative adversarial networks: algorithms, theory, and applications. IEEE Transactions on Knowledge and Data Engineering (2021). https://doi.org/10.1109/TKDE.2021.3130191
13. 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)

14. 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)
15. 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
16. Lommel, A., Uszkoreit, H., Burchardt, A.: Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. Tradumàtica (12), 0455–463 (2014)
17. Mueller-Franzes, G., Niehues, J.M., Khader, F., Arasteh, S.T., Haarburger, C., Kuhl, C., Wang, T., Han, T., Nolte, T., Nebelung, S., Kather, J.N., Truhn, D.: A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. Scientific Reports **13**(1) (jul 2023). https://doi.org/10.1038/s41598-023-39278-0, https://doi.org/10.1038%2Fs41598-023-39278-0
18. 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)
19. 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)
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL 2002). pp. 311–318 (2002)
21. 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)
22. 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)
23. 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)
24. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): A multimodal image dataset. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, pp. 180–189. Springer (2018)
25. Pelka, O., Nensa, F., Friedrich, C.M.: Annotation of enhanced radiographs for medical image retrieval with deep convolutional neural networks. PLOS ONE **13**(11), e0206229 (2018). https://doi.org/10.1371/journal.pone.0206229
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. CoRR **abs/2112.10752** (2021), https://arxiv.org/abs/2112.10752
27. 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)

28. 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)

29. Woodard, J., Nelson, J.: An information theoretic measure of speech recognition performance. In: Workshop on standardisation for speech I/O technology, Naval Air Development Center, Warminster, PA (1982)