Overview of the ImageCLEF 2024: Multimedia retrieval in medical applications

Bogdan Ionescu[1], Henning Müller[2], Ana-Maria Drăgulinescu[1], Johannes Rückert[3], Asma Ben Abacha[4], Alba García Seco de Herrera[5,6], Louise Bloch[3], Raphael Brüngel[3], Ahmad Idrissi-Yaghir[3], Henning Schäfer[8], Cynthia Sabrina Schmidt[7], Tabea M.G. Pakull[8], Hendrik Damm[3], Benjamin Bracke[3], Christoph M. Friedrich[3], Alexandra-Georgiana Andrei[1], Yuri Prokopchuk[9], Dzmitry Karpenka[9], Ahmedkhan Radzhabov[9], Vassili Kovalev[9,10], Cécile Macaire[11], Didier Schwab[11], Benjamin Lecouteux[11], Emmanuelle Esperança-Rodier[11], Wen-Wai Yim[4], Yujuan Fu[12], Zhaoyi Sun[12], Meliha Yetisgen[12], Fei Xia[11], Steven A. Hicks[13], Michael A. Riegler[13], Vajira Thambawita[13], Andrea Storås[13], Pål Halvorsen[13], Maximilian Heinrich[14], Johannes Kiesel[14], Martin Potthast[15], and Benno Stein[14]

[1] National University of Science and Technology Politehnica Bucharest
bogdan.ionescu@upb.ro
[2] University of Applied Sciences Western Switzerland (HES-SO), Switzerland
[3] Department of Computer Science, University of Applied Sciences and Arts Dortmund, Germany
[4] Microsoft, USA
[5] University of Essex, UK
[6] UNED, Spain
[7] Institute for Artificial Intelligence in Medicine, University Hospital Essen
[8] Institute for Transfusion Medicine, University Hospital Essen, Germany
[9] Belarusian National Academy of Sciences, Belarus
[10] Belarus State University, Belarus
[11] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[12] University of Washington, USA
[13] SimulaMet, Norway
[14] Bauhaus-Universität Weimar, Germany
[15] University of Kassel, hessian.AI, and ScaDS.AI, Germany

Abstract. This paper presents an overview of the ImageCLEF 2024 lab, organized as part of the Conference and Labs of the Evaluation Forum – CLEF Labs 2024. ImageCLEF, an ongoing evaluation event since 2003, encourages the evaluation of technologies for annotation, indexing and retrieval of multimodal data. The goal is to provide information access to large collections of data across various usage scenarios and domains. In 2024, the 22st edition of ImageCLEF runs three main tasks: (i) a *medical task*, continuing the caption analysis, Visual Question Answering for colonoscopy images alongside GANs for medical images, and medical dialogue summarization; (ii) a novel task related to *image retrieval/generation for arguments* for visual communication, aimed at augmenting the effectiveness of arguments; and (iii)ToPicto, a new task focused on *translating natural language*, whether spoken or textual, into a sequence

of pictograms. The benchmarking capaign was a real success and received the participation of over 35 groups submitting more than 220 runs.

Keywords: Medical text summarization · medical image caption analysis · visual question answering · Generative Adversarial Networks (GANs) · image retrieval · translation of neural language · ImageCLEF lab

1 Introduction

This paper presents the overview of the ImageCLEF 2024 lab, part of the Conference and Evaluation Forum - CLEF Labs 2024. Started in 2003, ImageCLEF[1] is an ongoing evaluation initiative that promotes the evaluation of technologies for annotation, indexing, and retrieval of visual data, facilitating information access to image collections across diverse domains. Over the years, ImageCLEF has continually adapted to emerging trends, adding tasks ranging from general object classification and retrieval to specialized application areas such as medical imaging, social media, nature, and security.

Over the years, ImageCLEF and also CLEF have shown a strong scholarly impact that was assessed in [45, 46]. For instance, the term "ImageCLEF" returns on Google Scholar[2] over 7,390 article results (search on June 11, 2024). This underlines the importance of the evaluation campaigns for disseminating best scientific practices.

In 2024, the 24$^{\text{th}}$ edition of ImageCLEF features three main tasks: i) a medical task continuing the four sub-tasks from the previous edition [24] (the 8$^{\text{th}}$ edition of the Caption task, the 5$^{\text{th}}$ edition of the MEDIQA task, and the 2$^{\text{nd}}$ editions for GANs and MedVQA tasks), ii) ToPicto, a new task task focusing on augmentative and alternative communication using pictograms, and iii) Image Retrieval for Arguments, a new task for ImageCLEF lab, organized in collaboration with Touché lab.

2 Overview of Tasks and Participation

ImageCLEF 2024 [23] consists of three main tasks to cover a *diverse range* of multimedia retrieval in *medical applications*. It followed the 2019 tradition [25] of diversifying the use cases [35, 44, 51, 41, 20, 3]. The 2024 tasks are presented as follows:

– **ImageCLEFmedical**. Since 2004, the ImageCLEF benchmarking initiative has included medical tasks. However, by 2018, although nearly all tasks were medical, there was minimal interaction between them. Therefore, beginning in 2019, the medical tasks were consolidated into a single task centered around a specific problem, with multiple subtasks. This approach fostered synergies between the different domains:

[1] `http://www.imageclef.org/`
[2] `https://scholar.google.com/`

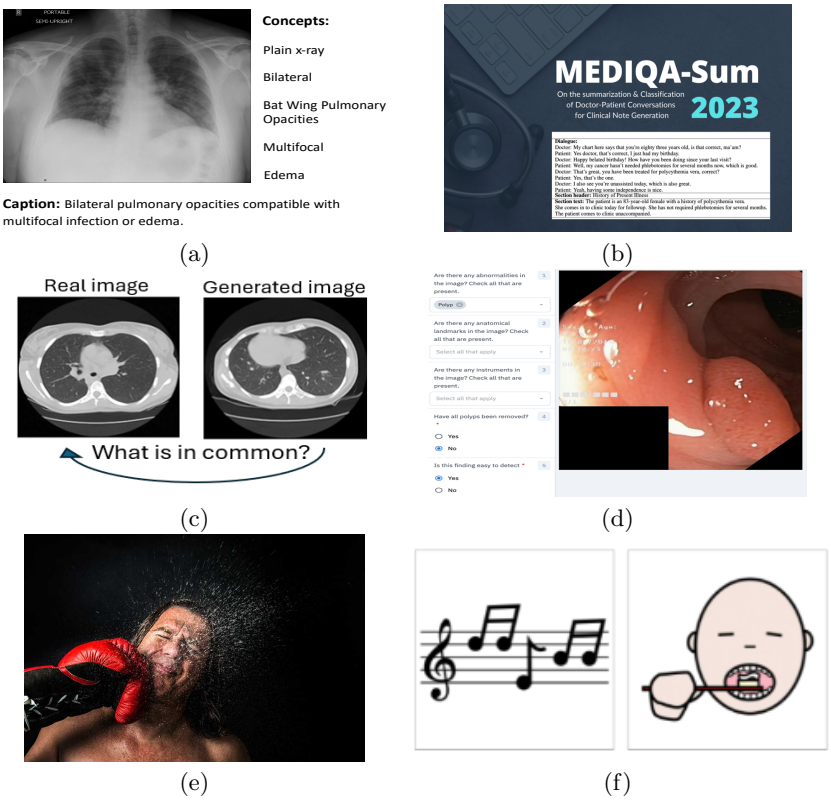

Fig. 1: Sample images: (a) ImageCLEFmedical-caption with an image and the corresponding CUIs and captions, (b ImageCLEFmedical-MEDIQA-MAGIC with an example of doctor-patient conversation, (c) ImageCLEFmedical-GAN with an example of real and generated images, (d) ImageCLEFmedical-VQA with examples of questions and answers in the area of colonoscopy, (e)) Argument-Image with a picture of a boxer conveying the premise that boxing causes injuries[3], (f) ToPicto from left to right: "music", "brush the teeth".

- *Caption*: This is the 8th edition of the task in this format, however, it is based on previous medical tasks. The task is once again running with both the "concept detection" and "caption prediction" subtasks [40], after the former was brought back in 2021 due to participants' demands [18, 34, 41]. The "caption prediction" subtask focuses on composing coherent captions for the entirety of a radiology image, while the "concept detection" subtask focuses on identifying the presence of relevant concepts in the same corpus of radiology images. After a smaller data set of manually annotated radiology images was used in 2021, the 2024 edition

once again uses a larger dataset based on Radiology Objects in COntext version 2 (ROCOv2) [42], which was already used in 2019-2023.

- *MEDIQA-MAGIC*: This is the fifth edition of the MEDIQA tasks and its second edition in the text format. The 2019 MEDIQA task featured several medical natural language semantic retrieval-related tasks, including natural language inference (NLI) classification of MIMIC-III clinical note sentences, recognizing question entailment (RQE) in consumer health questions, and reranking retrieved answers to consumer health questions. Continuing in 2021, the next MEDIQA task resumed hosting one clinical subtask and two consumer-health question-answer related subtasks [7]. Different from the 2019 subtasks, MEDIQA 2021 focused on summarization; summarization of clinical radiology note findings, consumer health questions, and consumer health answers. 2023 edition included clinical dialogue section header classification, short-dialogue note summarization, and full-encounter generation. The MEDIQA-MAGIC 2024 task mirrors the setup of the MEDIQA-M3G task. Participants receive a consumer health textual query along with associated images and are tasked with producing a preliminary doctor response. Responses are evaluated against two reference standards using deltaBLEU [17], BERT-score [53], and UMLS-F1 (F1 score of UMLS concept combined with an assertion label).
- *GANs*: In this edition, we continue to study the first sub-task illustrated in Fig. 1 – "Detect generative models' "fingerprints"– proposed in the previous edition [3] focused on examining the existing hypothesis that GANs generate medical images containing certain "fingerprints" of the authentic images used for generative network training. We extended the task by investigating this hypothesis for two different generative models. Another sub-task is introduced to this second edition — Detect generative models' "fingerprints". The second sub-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 [4].
- *MEDVQA-GI*: The MEDVQA-GI challenge is held for the second time this year with a new goal. One of the new frontiers in AI-driven medical diagnosis is the application of text-to-image generative models. This area integrates language processing and image synthesis to enhance diagnostic capability in the medical field. In this task, we aim to direct the power of artificial intelligence to generate medical images based on text input, along with optimal prompts for off-the-shelf generative models building up on the dataset collected in the first edition of MEDVQA-GI. The objective is to improve the diagnosis and classification of real medical images using AI-generated imagery. The task is divided into two main subtasks

[3] Source: Sweating fighter is punched in the face - gettyimages

– **Image Retrieval/Generation for Arguments** (Argument-Image) This is the third edition of the task. Pictures are a powerful means of visual communication and can be used to enhance the impact of arguments. This observation leads to our task where, given an argument, participants shall find images that help to convey the argument's premise. In this context "convey" is meant in a general manner; it can depict what is described in the argument, but it can also show a generalization (e.g., a symbolic image that illustrates a related abstract concept) or specialization (e.g., a concrete example). To better explain why an image conveys a premise, participants can optionally submit a rationale that helps explain why an image is relevant. This is a joint task with Touché 2024. Details on this task are provided in the Touché overview paper [27]. In Fig. 1 we see an an example submission for an argument, which consists of the premise "Boxing can lead to serious injuries." and the claim "Boxing is a dangerous sport!"
– **ToPicto**. This is the first edition of the task. The objective of ToPicto is to investigate the translation of natural language, either speech or text, into a sequence of pictograms as depicted in Fig 1. Generating pictograms is an emerging and significant domain in natural language processing, with multiple potential applications. It can enable communication with individuals who have disabilities, aid in medical settings for individuals who do not speak the language of a country, and also enhance user understanding in the service industry. Recent advances in artificial intelligence and machine translation have greatly improved performance in text-to-text as well as speech-to-text translations, but they have not been applied to text-to-pictogram and speech-to-pictogram translations before. ImageCLEFtoPicto seeks to bring together linguists, computer scientists, and translators to develop new translation methods. ImageCLEFtoPicto is divided into two subtasks:
  - *Text-to-Picto*: The first proposed subtask focuses on the automatic generation of a corresponding sequence of pictogram terms from a French text.
  - *Speech-to-Picto*: Building on the first subtask, Speech-to-Picto focuses on the two modalities speech and pictograms. The objective is to directly translate speech to a sequence of pictograms without going through the transcription dimension, which is the focus of the speech community with current spoken language translation systems.

In order to participate in the evaluation campaign, research groups were required to register by following the instructions on the ImageCLEF 2024 webpage[4]. This year, the challenges were organized through the AI4Media benchmarking platform[5] based on codalab[6]. Similar to previous editions, participants were required to submit a signed End User Agreement (EUA) to access the datasets. Table 1 summarizes the participation in ImageCLEF 2024, indicated the statistics both per task and for the overall lab. The table also shows the

[4] `https://www.imageclef.org/2024/`
[5] `https://ai4media-bench.aimultimedialab.ro/`
[6] `https://github.com/AIMultimediaLab/Ai4media-Bench`

Table 1: Key figures regarding participation in ImageCLEF 2024.

Task	Groups that submitted results	Submitted runs	Submitted working notes
Caption	14	82	13
MEDIQA-MAGIC	3	22	3
GANs	11	100	10
MedVQA	2	6	2
Argument-Image	2	8	2
ToPicto	4	7	4
Overall	33	225	34

number of groups that submitted runs and the ones that submitted a working notes paper describing the techniques used. Teams were allowed to register for several tasks. Following a decline in participation in 2016, there was an increase in 2017, 2018 and 2019. Specifically, in 2018, 31 teams completed the tasks and 28 working notes papers were submitted. In 2019, the number of participating teams climbed to 63, and we received 50 working papers. In 2020, 40 teams completed the tasks and submitted their working notes papers. In 2022, participation decreases with 28 teams completing the tasks and 26 working notes paper submitted. There was a new increase in 2023 with 47 teams submitting results and 39 working notes papers received. This year's edition of ImageCLEF attracted 36 teams and we received 34 working notes. The number of runs dropped compared to 2022 and 2023 with more teams involved 256 (2022) and 241 (2023) vs 225 (2024). This could be due to teams focusing on developing higher-quality solution and the increased complexity of the tasks this year, which may have required more time and resources per run.

In the following sections, we present the tasks. Only a short overview is reported, including general objectives, description of the tasks and data sets, and a short summary of the results. A detailed review of the received submissions for each task is provided with the task overview working notes: Caption [40], Mediqa [50], GAN [4], MedVQA [21], ToPicto [29] and Image Retrieval for Arguments [27].

3 The Caption Task

The caption task was first proposed as part of the ImageCLEFmedical [18] in 2016, aiming to extract the most relevant information from medical images. Hence, the task was created to condense visual information into textual descriptions. With the exception of 2019 and 2020, when only the concept detection task was offered, the ImageCLEFmedical Caption task has been running since 2017 with two subtasks: concept detection and caption prediction. With a break in 2021, where fewer images which were all manually annotated by medical doctors were used, an extended version of the ROCO data was set was used from 2019 to

2023 [41] for both subtasks, while the 2023 edition switched from BLEU-1 [32] to BERTScore [54] as the primary evaluation metric for caption prediction. In the 2024 edition of the ImageCLEFmedical Caption [40], the data used for both subtasks was based on the newly released ROCOv2 [42] data set.

3.1 Task Setup

The ImageCLEFmedical 2024 Caption [40] follows the format of the previous ImageCLEFmedical Caption tasks. In 2024, the overall task comprises two subtasks: "Concept Detection" and "Caption Prediction". The concept detection sub-task focuses on predicting Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUIs) [12] based on the visual image representation in a given image. The caption prediction subtask focuses on composing coherent captions for the entirety of the images. This year, a new optional, experimental explainability extension has been introduced for the caption prediction task. This extension aims to improve the understanding of the models by asking participants to provide explanations, such as heat maps or Shapley values, for a selected number of images. These explanations are manually reviewed to assess their effectiveness and clarity.

The detected concepts are evaluated using the balanced precision and recall trade-off in terms of F1-scores, as in previous years. Like last year, a secondary F1-score is computed using a subset of concepts that were manually curated and only contain x-ray anatomy and image modality concepts. Similar to last year, BERTScore was used as the primary metric for the evaluation of the caption prediction subtask. BERTScore evaluates the semantic similarity of the predicted captions. In addition to the BERTScore, a secondary ROUGE score, which measures the overlap of content between the predicted captions and reference captions, was provided. After the submission period ended, a number of additional scores were calculated and published: METEOR [5], CIDEr [48], CLIPScore [19], BLEU and BLEURT [43]. This year, two new metrics, MedBERTScore and ClinicalBLEURT [10], were added. These domain-adapted metrics are designed to better assess the relevance and accuracy of generated text in medical contexts, with the goal of improving the precision of evaluations in this specialized field.

3.2 Dataset

In 2024, an updated version of the ROCO dataset, called ROCOv2 [42], is utilized for both subtasks. The ROCOv2 dataset is derived from biomedical articles of the PMC Open Access Subset[7] [38] and was extended with new images added since the last time the dataset was updated. For this year, only CC BY and CC BY-NC licensed images are included. From the captions, UMLS® concepts were extracted, and concepts regarding anatomy and image modality were manually validated for all images.

Following this approach new training, validation, and test sets were provided for both tasks:

[7] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

Table 2: Performance of the participating teams in the ImageCLEFmedical 2024 Caption concept detection subtask. The best run per team is selected. Teams with previous participation in 2023 are marked with an asterisk.

Team	F1	Secondary F1
DBS-HHU	0.6375	0.9534
AUEB-NLP-Group*	0.6319	0.9393
DS@BioMed	0.6200	0.9312
SSNMLRGKSR*	0.6001	0.9056
UACH-VisionLab	0.5988	0.9363
MICLabNM	0.5795	0.8835
Kaprov	0.4609	0.7301
VIT_Conceptz	0.1812	0.2647
CS_Morgan*	0.1076	0.2105

- *Training set* including 70,108 radiology images and associated captions and concepts.
- *Validation set* including 9972 radiology images and associated captions and concepts.
- *Test set* including 17,237 radiology images.

Table 3: Performance of the participating teams in the ImageCLEFmedical 2024 Caption caption prediction subtask. The best run per team is selected. Teams with previous participation in 2023 are marked with an asterisk.

Team	BERTScore	ROUGE
PCLmed	0.6299	0.2726
CS_Morgan	0.6281	0.2508
DarkCow	0.6267	0.2452
AUEB-NLP-Group	0.6211	0.2049
2Q2T	0.6178	0.2478
MICLab	0.6128	0.2135
DLNU_CCSE	0.6066	0.2179
Kaprov	0.5964	0.1905
DS@BioMed	0.5794	0.1031
DBS-HHU	0.5769	0.1531
KDE-medical-caption	0.5673	0.1325

3.3 Participating Groups and Submitted Runs

In the eighth edition of the ImageCLEFmedical Caption task, 50 teams registered and signed the End-User-Agreement that is needed to download the development data. 14 teams submitted 82 graded runs (13 teams submitted working notes)

attracting similar attention to 2023. Similar to last year, participants did not have access to their own scores until after the submission period was over. Of the 9 teams that participated in the concept detection subtask this year, 4 also participated in 2023. Of the 11 teams which submitted runs to the caption prediction subtask, 6 also participated in 2023. Overall, 6 teams participated in both subtasks, and 5 teams participated only in the caption prediction subtask. Unlike in 2023, 3 teams participated only in the concept detection subtask.

In the concept detection subtasks, the groups used primarily multi-label classification systems.

The winning team this year utilized an ensemble of four different CNNs. In the caption prediction subtask, teams primarily utilized encoder-decoder frameworks with various backbones, including transformer-based decoders and LSTMs [22].

The winning team introduced medical vision-language foundation models (Med-VLFMs) by combining general and specialist vision models to achieve top rankings in the challenge.

To get a better overview of the submitted runs, the primary scores of the best results for each team are presented in Tables 2 and 3.

3.4 Results

For the concept detection subtak, the overall F1 scores increased strongly compared to last year despite very similar approaches being employed by the teams. In addition to continuously improved and scaled-up approaches by the teams, some possible explanations for this include an improved and overall larger dataset, a lower number of unique concepts in the test set, and the removal of directionality concepts.

The same applies for the general view on results of this year's caption prediction task. The top scores were slightly worse for BERTScore, but last year's winners CSIRO did not participate this year. Returning teams improved their scores across the board showing that the dataset for this year is comparable to last year and that while teams have experimented with many different approaches including LLMs for caption generation, no breakthrough improvement has been achieved with these new techniques. The new optional explainability extension was not adpoted by the teams, only the team MICLabNM [14] submitted explainability results after the end of the submission phase.

3.5 Lessons Learned and Next Steps

This year's caption task of ImageCLEFmedical once again ran with both subtasks, concept detection and caption prediction. Like last year, it used a ROCOv2-based data set for both challenges. Manually validated concepts for X-ray directionality information introduced last year were removed for this year's dataset. It attracted 14 teams who submitted a total of 82 graded runs, a similar level of participation to last year. Changes were introduced in the evaluation metrics,

with the addition of two new domain-specific metrics, MedBERTScore and ClinicalBLEURT, specifically for the caption prediction task. These additions were made based on feedback received from participants the previous year.

For next year's ImageCLEFmedical Caption challenge, some possible improvements include an improved caption prediction evaluation metric which is specific to medical texts or a combination of different metrics, as well as additional metrics for readability and factuality. The optional explainability extension might be moved into its own subtask for next year.

4 The MEDIQA-MAGIC Task

Since 2019, the MEDIQA shared-tasks have tackled various question-answering and summarization challenges related to medical reasoning, language, and semantics. Its first edition included the classification tasks of clinical note sentence natural language inference and recognizing question entailment, as well as their application towards answer-retrieval re-ranking. In 2021, the MEDIQA challenges focused on monologue-to-monologue summarization tasks, including clinical radiology note findings summarization, consumer health question summarization, and multiple answer summarization [7]. In 2023, two editions were hosted. Both featured problems related to dialogue-to-monologue summarization for clinical note from doctor-patient conversations. Subtasks included short-dialogue section header and note generation, topic-to-note summarization, full-encounter dialogue-to-note generation, and full-encounter note-to-dialogue generation [8, 51]. This year, similarly two related editions were hosted. These tasks revolved around the problem of multi-modal visual question-answering tasks on consumer health dermatology problems. While MEDIQA-M3G [9] (multi-modal, multilingual answer generation), part of the NAACL 2024 ClinicalNLP Workshop focused on short-answers in English, Chinese, and Spanish; the MEDIQA-MAGIC (Multimodal And Generative TelemedICine) task part of ImageCLEF 2024, described here, included in-depth full answer responses for English only.

4.1 Task Setup

The MEDIQA-MAGIC 2024 task follows the setup for the MEDIQA-M3G task. Participants are supplied with a consumer health textual query and associated images. The target objective is to output a draft doctor response. The evaluated responses were graded against two reference standards using deltaBLEU [17], BERTScore [53], and UMLS-F1 (F1 score of UMLS concept combined with an assertion label). For more comprehensive details related to the task, dataset, and results, please refer to the task overview paper [50].

4.2 Dataset

The 2024 MEDIQA-MAGIC challenge used data from the Reddit sub-collection of the DermaVQA dataset [52]. To comply with data usage guidelines, only

post id's and our labels were shared with participants. Participants who registered through Reddit could receive API credentials to access Reddit's data. Afterwards, the participants could use supplied download script[8] to retrieve the original input data. As Reddit users may opt to delete content, the final set of test set id's were determined by the subset of test id's retrieval shortly after the submission deadline. The original labeled dataset included 347, 50, 93 instances for train, valid, and test sets. The final number of test set instances was 78.

4.3 Participating Groups and Submitted Runs

Overall 3 teams participated with a total of 22 runs. The teams came from three different countries and continents (India, Poland, and Taiwan).

4.4 Results

The final results are shown in Table 4. The submitted systems represented a variety of solutions, including using out-of-the-box gemini [1] models (YuanAI), applying small visual language models (VisionQAries), and utilizing visual-language encoders with cosine similarities(IRLab@IIT_BHU). The ranges of scores were co-located in the lower spectrum for all three metrics (100 total for BLEU, and 1.0 for BERTScore and UMLS F1), indicating the difficulty of the task.

Table 4: Performance of the participating teams in the MEDIQA-MAGIC 2024 Answer Generation Task (Best Run).

Team	Institution	BLEU	BERTScore	MEDCON
VisionQAries	IIT (BHU), Varanasi, India	8.969	0.844	0.077
IRLab@IIT_BHU	Poland	4.536	0.839	0.066
YuanAI	Yuan Ze University, Taiwan	4.371	0.856	0.087

The following sections briefly describe the teams' solutions. More information can be found in the overview [50]:

IRLab@IIT_BHU manually labeled instances into 160 categories, passing image and text data through CLIP encoders. Text data went through a Bi-LSTM and vision data through an MLP, with their results averaged to create a label vector. Training involved weighted cosine similarity loss. During inference, the combined embedding matched the closest label embedding. They also used data augmentation with TextGenie and GPT2 for classification.
VisionQAries focused on small multi-modal models, testing direct prompting and fine-tuning on moondream2 and TinyLLaVA models. Fine-tuning moondream2 yielded better BLEU scores than direct prompting.
YuanAI used the Gemini image-to-text model, followed by a LoRA fine-tuned Llama3 to process outputs and queries, generating the final response.

[8] https://github.com/wyim/MEDIQA-MAGIC-2024

4.5 Lessons Learned and Next Steps

This year's ImageCLEF MEDIQA-MAGIC task differed from the 2024 NAACL ClinicalNLP MEDIQA-M3G Shared Tasks [9] by using a different dataset and requiring participants to obtain Reddit credentials, which may have deterred some teams. Another major difference was the longer answer lengths, averaging 90 words compared to 12, increasing the challenge in answer generation and evaluation.The competition used a codabench-based platform for easier submissions and result computation, with an API for automatic participant data download. This year required GitHub code submissions, unlike last year's requirement for run-able code, resulting in less complete documentation. Future editions may use Codabench's real-time inference to ensure clean, run-able code without manual effort.

This task required extensive free-text answers, unlike other visual question-answering tasks with 1-2 word responses, and allowed multiple correct answers, presenting challenges in natural language evaluation. Future editions will address specialty to consumer health multi-modal problems and experiment with evaluation methods for longer text and multiple correct answers.

5 The GANs Task

Biomedical imaging has advanced significantly in recent years due to the convergence of machine learning (ML) and artificial intelligence (AI) technologies, particularly through the development of generative models like Generative Adversarial Networks (GANs). These models have proven effective in producing synthetic images that mimic real biomedical images, creating new opportunities for study and application. Synthetic images produced by these models offer several potential advantages in the biomedical domain, including augmenting existing datasets to address data scarcity and imbalances, which is especially valuable given the difficulty, cost, and time involved in obtaining large amounts of labeled medical data. Moreover, AI algorithms benefit from synthetic images by reducing dependency on real patient data, thus mitigating privacy concerns.

5.1 Task Setup

This is the second edition of the task and consists of two sub-tasks. In addition to the sub-task presented in the previous edition, "Identify training data fingerprints" [3], we have introduced the second sub-task entitled "Detect generative models' fingerprints". The objective of the first sub-task was to detect "fingerprints" within the synthetic biomedical image data to determine which real images were used in training to produce the generated images. The task is formulated as follows:

- *given two sets that contain generated and real images, the participants are requested to employ machine learning and/or deep learning models to determine for each set which of the real images were used to train the model to generate the provided synthetic images.*

The second sub-task explores the hypothesis that generative models imprint unique fingerprints on generated images. The focus is on understanding whether different generative models or architectures leave discernible signatures within the synthetic images they produce. By providing a set of synthetic The task was formulated as follows:

- *given a set of generated images and the number of generative models used, the participants are required to group the images based on the model that generated them.*

5.2 Dataset

The benchmarking image data consists of axial slices of 3D CT images extracted from a bigger dataset of about 8000 lung tuberculosis patients. Considering this, some of the slices may appear pretty "normal" whereas the others may contain certain lung lesions including severe ones. These images are stored as 8-bit/pixel PNG images with dimensions of 256x256 pixels. The artificial slice images are 256x256 pixels in size. The dataset for the first sub-task consisted of both real and generated images, while the dataset for the second sub-task consisted in synthetic images only generated using different generative models.

5.3 Participating Groups and Submitted Runs

Overall, 23 teams registered to both tasks. Among them, 10 teams completed the first sub-task and submitted their runs, while 7 teams completed the second sub-task (including the task organizing team). Notably, 6 teams were common to both sub-tasks, demonstrating consistency across the tasks. When it comes to submitting the working notes, one team did not submit them, resulting in an adherence rate of 90.90%.

5.4 Results

For the first sub-task, "Identify training data fingerprints", a variety of methods were employed, ranging from advanced image preprocessing techniques to deep learning models. Various techniques such as binarization, histogram equalization, feature extraction, noise reduction, noise addition, colorization were used to accentuate distinct features. Different neural network architectures, including ResNet, MobileNet and autoencoders were used for feature extraction and classification. The task was evaluated as a binary-class classification problem and the evaluation was carried out by measuring the F1-score, the official evaluation metric of this year's edition. The results are presented in Table 5.

For the second sub-task, "Detect generative models fingerprints", most teams used pre-trained deep learning models such as ResNet, DenseNet, EfficientNet, MobileNetV2, VGG, and Inception for feature extraction. These models were chosen for their proven efficacy in capturing complex patterns and hierarchical features in images. A variety of clustering algorithms were employed across the

Table 5: The results obtained by the participating teams to the first sub-task proposed by ImageCLEFmedical GANs – Identify training data fingerprints.

Rank	Team	ID #	F1-score	Rank	Team	ID #	F1-score
#1	Inoue Koki	892	0.666	#28	csmorgan	884	0.5
#2	Inoue Koki	896	0.663	#29	AI Multimedia Lab	901	0.499
#3	Inoue Koki	891	0.663	#30	Biomedical Imaging Goa	874	0.499
#4	Inoue Koki	894	0.66	#31	Biomedical Imaging Goa	873	0.497
#5	Inoue Koki	895	0.638	#32	csmorgan	883	0.496
#6	Inoue Koki	890	0.631	#33	csmorgan	886	0.492
#7	AI Multimedia Lab	909	0.627	#34	KDE-med-lab	854	0.488
#8	Inoue Koki	893	0.626	#35	csmorgan	879	0.483
#9	SDVAHCS/UCSD	848	0.624	#36	csmorgan	878	0.47
#10	SDVAHCS/UCSD	849	0.606	#37	Shitongcao	833	0.462
#11	Robot	844	0.603	#38	KDE-med-lab	857	0.46
#12	Shitongcao	834	0.598	#39	KDE-med-lab	853	0.455
#13	Shitongcao	836	0.598	#40	KDElab	897	0.454
#14	AI Multimedia Lab	905	0.538	#41	Shitongcao	835	0.451
#15	Biomedical Imaging Goa	898	0.531	#42	Shitongcao	839	0.448
#16	Shitongcao	838	0.529	#43	KDE-med-lab	856	0.443
#17	AI Multimedia Lab	906	0.527	#44	Biomedical Imaging Goa	876	0.43
#18	Robot	841	0.524	#45	Robot	845	0.429
#19	AI Multimedia Lab	903	0.515	#46	Biomedical Imaging Goa	877	0.385
#20	Biomedical Imaging Goa	875	0.515	#47	Robot	846	0.35
#21	SDVAHCS/UCSD	850	0.511	#48	Robot	842	0.314
#22	KDE-med-lab	852	0.51	#49	Robot	843	0.312
#23	AI Multimedia Lab	904	0.51	#50	Robot	847	0.312
#24	Robot	840	0.503	#51	Shitongcao	837	0.255
#25	AI Multimedia Lab	902	0.502	#52	AI Multimedia Lab	908	0.2358
#26	SDVAHCS/UCSD	851	0.501	#53	Shitongcao	832	0.2
#27	csmorgan	881	0.5	#54	KDE-med-lab	855	0.019

methods. K-means was the most commonly used clustering algorithm, but other techniques like hierarchical clustering, Gaussian Mixture Models (GMM), and t-SNE were also applied to group the extracted features based on their similarities. Many approaches involved combining multiple models or techniques to enhance robustness. Adjusted Rand Index (ARI) was the official evaluation metric of the competition and the results are presented in Table 6 More detailed results, including methods presentation and other performance measures, are presented in the overview article [4].

5.5 Lessons Learned and Next Steps

The second edition of the ImageCLEFmedical GANs task introduced two sub-tasks for participants: a prediction-based task utilizing both real and generated images and a clustering task using only generated images. This task provided insights into the complexities of working with synthetic medical images. Participants employed a variety of methods, including advanced image preprocessing techniques, deep learning models, and clustering algorithms for the two proposed sub-tasks.

Future editions of the task will expand the scope by incorporating a wider variety of data and generation methods to better reflect real-world applications and address existing limitations. Furthermore, new tasks will be introduced to

Table 6: The results obtained by the participating teams to the second sub-task proposed by ImageCLEFmedical GANs – Detect generative models' fingerprints

Rank	Team	ID #	ARI	Rank	Team	ID #	ARI
#1	SDVAHCS/UCSD	545	1	#24	Csmorgan	451	0.267530
#2	AI Multimedia Lab	330	0.997085	#25	Csmorgan	458	0.232390
#3	AI Multimedia Lab	327	0.996517	#26	KDE-med-lab	237	0.226339
#4	AI Multimedia Lab	326	0.934709	#27	Csmorgan	456	0.178545
#5	AI Multimedia Lab	331	0.900844	#28	KDE-med-lab	248	0.166582
#6	Csmorgan	447	0.9000159	#29	KDE-med-lab	257	0.123426
#7	SDVAHCS/UCSD	550	0.885478	#30	KDE-med-lab	271	0.091818
#8	SDVAHCS/UCSD	590	0.877797	#31	KDE-med-lab	258	0.060058
#9	SDVAHCS/UCSD	548	0.851990	#32	KDE-med-lab	254	0.045286
#10	SDVAHCS/UCSD	549	0.851362	#33	KDE-med-lab	270	0.038242
#11	Csmorgan	446	0.813749	#34	KDE-med-lab	259	0.014388
#12	AI Multimedia Lab	334	0.722857	#35	KDE-med-lab	480	0.013856
#13	AI Multimedia Lab	333	0.654021	#36	SDVAHCS/UCSD	546	0.003375
#14	AI Multimedia Lab	335	0.645386	#37	Csmorgan	454	0.001776
#15	Biomedical Imaging Goa	307	0.638117	#38	Csmorgan	453	0.001313
#16	SDVAHCS/UCSD	547	0.577203	#39	KDE-med-lab	236	0.000816
#17	SDVAHCS/UCSD	225	0.577203	#40	GAN-Amis	516	0.000079
#18	AI Multimedia Lab	332	0.552682	#41	Biomedical Imaging Goa	323	0.000046
#19	AI Multimedia Lab	329	0.5037	#42	GAN-Amis	518	-0.000010
#20	Biomedical Imaging Goa	321	0.434414	#43	GAN-Amis	520	-0.000546
#21	Csmorgan	452	0.365604	#44	GAN-Amis	277	-0.000615
#22	AI Multimedia Lab	328	0.329388	#45	GAN-Amis	513	-0.000993
#23	Biomedical Imaging Goa	324	0.272975	#46	GAN-Amis	517	-0.002019

explore different aspects of privacy and security in synthetic medical data and alternative evaluation metrics will be investigated to ensure a more comprehensive assessment of the methodologies employed.

6 The MEDVQA-GI Task

The second iteration of the MedVQA-GI challenge introduces a new goal that focuses on the use of generative models of text to image in medical diagnosis. This combines natural language processing and image generation to potentially improve diagnostic processes in healthcare by providing more comprehensive datasets that can be used for machine learning training. In contrast to last year's focus on a visual question answering task that required retrieving images or masks from user questions, this year's task aims to use generative models to create synthetic medical images from textual inputs. Participants are tasked generating the synthetic images using existing generative models developed using a dataset derived from last years MedVQA-GI challenge.

6.1 Task Setup

This year, the competition is divided into two subtasks: Image Synthesis (IS) and Optimal Prompt Generation (OPG). Participants are welcome to submit entries for one or both tasks, with no restrictions on the number of submissions.

The IS task challenged participants to use text-to-image generative models to create a dataset of medical images from textual descriptions. The objective is to

produce accurate visual representations of various medical conditions described in text. For example, given the description "An early-stage colorectal polyp," participants are expected to generate an image that precisely reflects the text description.

The OPG task asked participants to build prompts that guide the generation of images meeting specific medical imaging requirements. This task tests the ability to develop prompts that result in images accurately matching predefined categories, emphasizing the model's capability to produce precise and clinically relevant images. For more comprehensive details on the tasks, datasets, and evaluation metrics, please see the task overview paper [21].

6.2 Dataset

The dataset used for this year's challenge is based on data developed for last years challenge, which is based on the HyperKvasir dataset [13] and the Kvasir-Instrument dataset [26] datasets. Participants were provided with a development dataset consisting of $2,000$ image and text pairs, and a list of $5,000$ prompts to generate their results. The development data was organized with a directory containing the images and CSV files containing the prompts and connection to the image filenames. For testing, we provided a list of prompts that participants used to generate their synthetic images.

6.3 Results

Overall, we had a total of six runs submitted to Task 1 and none to Task 2, where each team submitted three runs and the results are shown in Table 7. Team MMCP [16] employed two distinct methods for image generation: they fine-tuned existing Kandinsky models and developed a Medical Synthesis with Diffusion Model (MSDM), with the latter showing superior results. Team 2 [31] explored three different approaches in their work. Initially, they used a CLIP model to retrieve images closely related to the input prompts rather than generating new ones. Next, they used a fine-tuned stable diffusion model for creating synthetic images. Lastly, they implemented a Low-Rank Adaptation of Large Language Models (LoRA), modifying a stable diffusion model to produce high-quality images that closely match the input specifications. Overall, the best submission goes to Team MMCP [16], who achieved best results on the quantitative metrics and also visually best results.

6.4 Lessons Learned and Next Steps

Overall, we observed a reduction in participation compared to last year. There may be several reasons for this, like the complexity of tasks, change of direction from last year, or a lack of foundational resources among the participants. Addressing these barriers could involve "getting started" scripts and potentially simplifying the challenge structure to attract a broader range of participants.

Table 7: Results for Task 1. Each submission is evaluate dusing the FID and the Inception Score (IS). The FID scores is calculated against the MedVQA testing datasert (Single), GastroVision (Multi), and a combination of the two (Both). The IS score is calculated on a 10-way split of the synthetic images, where we display the mean (avg), standard deviation (sd), and median (med).

Team	Submission	FID (Single)	FID (Multi)	FID (Both)	IS (avg)	IS (std)	IS (med)
MMCP	1	0.125	0.121	0.119	1.773	0.023	1.775
	2	0.120	0.117	0.115	1.791	0.028	1.792
	3	0.086	0.064	0.066	1.624	0.031	1.633
team2	1	0.114	0.128	0.124	1.568	0.025	1.560
	2	0.099	0.064	0.067	2.327	0.065	2.339
	3	0.110	0.073	0.076	2.362	0.050	2.359

7 ToPicto

Several diseases (e.g., Rett syndrome, Cerebral Palsy, Parkinson's Disease) lead to language impairment, which significantly interferes, as a consequence, with the development of language skills (speaking, listening, reading, and writing). Language production and comprehension are impaired. For these specific cases, Augmentative and Alternative Communication (AAC) can be implemented with the use of pictograms [39]. Pictograms, in AAC, refer to an image linked to a concept that can be a single word, a named entity, or a multi-word expression among others. Using pictograms as a communication aid has been proven effective in visualizing syntax, manipulating words, and facilitating language access [15]. Moreover, the use of AAC has a positive social impact for people with language impairment. The French Red Cross has identified a reduction in stress, an improvement in autonomy and health, and greater serenity and enjoyment in daily life. The main objective of this task is to provide a translation in pictogram terms (each linked to a specific pictogram image from the ARASAAC bank[9] from a natural language (speech or text) understandable by the users with language impairments. The translation has to follow a specific structure and should convey the meaning of the input.

7.1 Task Setup

The first edition of the ToPicto task consisted of two subtasks: *Text-to-Picto* and *Speech-to-Picto*. Participants could choose to work on both tasks or just one of them without any obligation to achieve specific results.In the Text-to-Picto subtask, participants were asked to translate a text input into a pictogram sequence. The subtask involved implementing translation techniques and models to generate a specific pictogram sequence. The second subtask, Speech-to-Picto

[9] `https://arasaac.org/`

is the continuation of Text-to-Picto, but focuses on the speech modality. Participants had to generate a pictogram sequence from a speech input. The objective was to adapt current spoken language translation systems, such has in [11] to the pictogram generation.

7.2 Dataset

The dataset consisted of oral transcriptions (for the Text-to-Picto subtask) and audio utterances (for the Speech-to-Picto subtask) translated into sequences of pictogram terms built from the TCOF corpus [2]. The TCOF corpus contains interactions between adults, adults and children, and children themselves, covering a wide range of topics such as debates, everyday situations, and medical consultations. This type of text is representative of the interactions we observe between caregivers (families, medical staff) and individuals who rely on pictograms due to language impairments.

For each utterance, we applied the method of [28] to extract the pictogram sequence. This sequence was carefully developed and evaluated by experts of the pictographic language. The audio files were a maximum of 30 seconds length with a sampling rate of 16 kHz. For the challenge, the dataset was split into three sets, training, validation and test with a 90/5/5 distribution respectively. General statistics about the dataset are presented in Table 8. The resulting data were provided in a JSON format to the participants with the following information: (i)*id*: the unique identifier of the utterance; (ii)*src*: the input sequence (either text or speech); (iii) *tgt*: the target sequence of pictogram terms; (iv) *pictos*: a list of pictogram identifier linked to each pictogram terms.

The *pictos* tag was provided for reference to give an idea of the input with the sequence of pictogram images. Each pictogram image could be obtained from the ARASAAC website from the provided identifier. The dataset will be released shortly after the end of the challenge.

Table 8: General statistics of the ToPicto dataset.

	train	valid	test
# utterances	24,270	1,348	1,350

7.3 Participating Groups and Submitted Runs

A total of 16 teams participated in the ToPicto challenge, with most registering for both tasks. Four teams completed the Text-to-Picto task. Unfortunately, no submissions were received for the Speech-to-Picto subtask. Every team provided their working notes, resulting in a 100% adherence rate.

7.4 Results

In the following section, we only discuss the submission from the Text-to-Picto subtask. The participants employed several models that are based on the same architecture, Transformer [47]. Two teams made use of multilingual pre-trained models, T5 [37] and Helsinki-NLP/opus-mt-ROMANCE-en [10]. Other models, monolingual, were also applied, specifically on the French language with CamemBERT [30] and on the English language with GPT-2 model [36]. A final work implemented an encoder-decoder architecture with LSTM layers. The evaluation was based on metrics commonly used in the translation community. The evaluation process involved comparing the reference pictogram terms sequence with the hypothesis given by the model. Three metrics were computed: BLEU score [33], METEOR [6] and the Picto-term Error Rate (PictoER), which is based on the Word Error Rate metric [49]. The results are presented in Table 9.

Table 9: The results obtained by the participating teams to the Text-to-Picto sub-task of ToPicto.

Rank	Team	Run	BLEU	METEOR	PictoER
#1	TechTitans	3	74.36	87.08	13.90
#2	TechTitans	2	67.85	83.69	17.57
#3	TechTitans	3	66.56	82.89	18.43
#4	InnoVate	2	68.96	83.54	18.51
#5	SSN-MLRG	1	3.41	14.35	141.90
#6	SSN-MLRG	2	3.41	14.35	141.90
#7	InnoVate	2	3.93	25.56	170.80

7.5 Lessons Learned and Next Steps

The first edition of the task introduced two subtasks: generating a coherent sequence of pictogram terms from either a text utterance (Text-to-Picto) or a speech utterance (Speech-to-Picto). This challenge, previously receiving limited attention, was presented to the community for the first time. Participants employed a variety of methods, ranging from multilingual to monolingual pre-trained models, and encoder-decoder architectures, yielding interesting outcomes in translation. However, the Speech-to-Picto subtask did not result in any submissions, likely due to the challenges associated with starting from a speech modality.

Future editions of the task might explore different language sets and various domains, such as the medical field. Additionally, an important aspect of providing comprehensible translations is simplifying the text input beforehand, which could serve as a new subtask in the ToPicto challenge. Finally, the dynamic construction of pictograms using generative models could also be explored.

[10] `https://huggingface.co/Helsinki-NLP/opus-mt-ROMANCE-en`

8 Conclusion

This paper presents an overview and the outcomes of ImageCLEF 2024 benchmarking campaign. Three main tasks were organised, addressing challenges in the medical domain (caption analysis, visual question answering, medical dialogue summarisation, GANs for medical image generation), natural language translation (generating pictogram from speech and text), and image retrieval/-generation for arguments.

Similar to the previous year, the vast majority of solutions provided by the participants were based on machine learning and deep learning techniques. In ImageCLEFmedical – Caption, multi-label classification was common for concept detection, with some teams integrating image retrieval. Encoder-decoder frameworks with transformers and LSTMs were used for caption prediction. In ImageCLEFmedical – MEDIQA-MAGIC, participants used classic algorithms like SVM, KNN, and Random Forest, along with TF-IDF and lemmatization. Pre-trained models like GPT3.5, clinical-BERT, and clinical T5, including their LoRA adaptations, were also utilized. For ImageCLEFmedical GAN, methods included advanced preprocessing, deep learning models, binarization, histogram equalization, and feature extraction. Majority voting and agglomerative clustering improved results. For the second sub-task, pre-trained CNNs were used for feature extraction, with clustering algorithms like k-means, hierarchical clustering, GMM, and t-SNE. For the ImageCLEFmedical-MedvQA, the participants employed transformer-based pre-trained models. In the first edition of the ToPicto task, methods for Text-to-Picto included multilingual and monolingual pretrained models, and encoder-decoder architectures, achieving interesting translation outcomes. ImageCLEF 2024 offered participants and the community a wide range of tasks and methodologies to delve into, highlighting an exciting fusion of approaches.

Future editions of the ImageCLEF tasks hold exciting potential for growth and innovation. They may broaden domains, including tasks to attract more people, and try new methods like generative models for the GANs task. To overcome barriers in participation, like complicated taks, offering resources may be necessary. Additionally, refining evaluation metrics and exploring alternative approaches are crucial for advancing understanding across disciplines. These actions aim to drive progress and foster collaboration in diverse areas of research.

Acknowledgements

The lab is supported under the H2020 AI4Media "A European Excellence Centre for Media, Society and Democracy" project, contract #951911, as well as the ImageCLEFmedical GANs tasks. The work of Louise Bloch, Raphael Brüngel and Benjamin Bracke was partially funded by a PhD grant from the University of Applied Sciences and Arts Dortmund (FH Dortmund), Germany. The work of Ahmad Idrissi-Yaghir, Tabea M. G. Pakull, Hendrik Damm 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). The ToPicto task was funded by the Agence Nationale de la Recherche (ANR) through the project PROPICTO (ANR-20-CE93-0005).

References

1. Gemini models. https://ai.google.dev/gemini-api/docs/models/gemini (2024), accessed: 2024-04-24
2. 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)
3. Andrei, A., 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: CLEF2023 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 18-21 2023)
4. 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 (September 9-12 2024)
5. 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. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005), https://aclanthology.org/W05-0909
6. 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)
7. Ben Abacha, A., Mrabet, Y., Zhang, Y., Shivade, C., Langlotz, C.P., Demner-Fushman, D.: Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In: Proceedings of the 20th Workshop on Biomedical Language Processing, BioNLP@NAACL-HLT 2021, Online, June 11, 2021. pp. 74–85. Association for Computational Linguistics (2021), https://doi.org/10.18653/v1/2021.bionlp-1.8
8. Ben Abacha, A., wai Yim, W., Adams, G., Snider, N., Yetisgen, M.: Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In: ACL-ClinicalNLP 2023 (2023)
9. Ben Abacha, A., Yim, W., Fu, Y., Sun, Z., Xia, F., Yetisgen, M., Krallinger, M.: Overview of the mediqa-m3g 2024 shared tasks on multilingual multimodal medical answer generation. In: NAACL-ClinicalNLP 2024 (2024)
10. Ben Abacha, A., Yim, W.w., Michalopoulos, G., Lin, T.: An investigation of evaluation methods in automatic medical note generation. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) 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

11. 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)
12. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research **32**(Database-Issue), 267–270 (2004). https://doi.org/10.1093/nar/gkh061
13. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific data **7**(1) (2020). https://doi.org/10.1038/s41597-020-00622-y
14. Carmo, D., Rittner, L., Lotufo, R.: VisualT5: Multitasking caption and concept prediction with pre-trained ViT, T5 and customized spatial attention in radiological images. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France (September 9-12 2024)
15. Cataix-Nègre, E.: Communiquer autrement: Accompagner les personnes avec des troubles de la parole ou du langage. De Boeck Superieur (2017)
16. Chaychuk, M.: Mmcp team at imageclefmed 2024 task on image synthesis: Diffusion models for text-to-image generation of colonoscopy images. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France (September 2024)
17. Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J., Dolan, B.: deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 445–450. Association for Computational Linguistics, Beijing, China (Jul 2015)
18. García Seco de Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 medical task. In: Working Notes of CLEF 2016 (Cross Language Evaluation Forum) (September 2016)
19. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. pp. 7514–7528. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.emnlp-main.595, https://doi.org/10.18653/v1/2021.emnlp-main.595
20. 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 2023)
21. 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, Grenoble, France (September 2024)
22. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (nov 1997). https://doi.org/10.1162/neco.1997.9.8.1735, https://doi.org/10.1162/neco.1997.9.8.1735
23. Ionescu, B., Müller, H., Drăgulinescu, A.M., Idrissi-Yaghir, A., Radzhabov, A., Herrera, A.G.S.d., Andrei, A., Stan, A., Storås, A.M., Abacha, A.B., et al.: Advancing multimedia retrieval in medical, social media and content recommenda-

tion applications with imageclef 2024. In: European Conference on Information Retrieval. pp. 44–52. Springer (2024)
24. Ionescu, B., Müller, H., Drăgulinescu, A., Yim, W., Ben Abacha, A., Snider, N., Adams, G., Yetisgen, M., Rückert, J., de Herrera, A.G.S., Friedrich, C.M., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Hicks, S.A., Riegler, M.A., Thambawita, V., Storås, A., Halvorsen, P., Papachrysos, N., Schöler, J., Jha, D., Andrei, A., Radzhabov, A., Coman, I., Kovalev, V., Stan, A., Ioannidis, G., Manguinhas, H., Ştefan, L., Constantin, M.G., Dogariu, M., Deshayes, J., Popescu, A.: Overview of ImageCLEF 2023: Multimedia retrieval in medical, socialmedia and recommender systems applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science LNCS, Thessaloniki, Greece (September 18-21 2023)
25. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), vol. 11438. LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)
26. Jha, D., Ali, S., Emanuelsen, K., Hicks, S.A., Thambawita, V., Garcia-Ceja, E., Riegler, M.A., de Lange, T., Schmidt, P.T., Johansen, H.D., Johansen, D., Halvorsen, P.: Kvasir-Instrument: Diagnostic and Therapeutic Tool Segmentation Dataset in Gastrointestinal Endoscopy. In: Proceedings of the International Conference on MultiMedia Modeling (MMM). pp. 218–229 (2021), https://doi.org/10.1007/978-3-030-67835-7_19
27. Kiesel, J., Çöltekin, Ç., Heinrich, M., Fröbe, M., Alshomary, M., Longueville, B.D., Erjavec, T., Handke, N., Kopp, M., Ljubešić, N., Meden, K., Mirzakhmedova, N., Morkevičius, V., Reitis-Münstermann, T., Scharfbillig, M., Stefanovitch, N., Wachsmuth, H., Potthast, M., Stein, B.: Overview of Touché 2024: Argumentation Systems. In: Goeuriot, L., Mulhem, P., Quénot, G., Schwab, D., Soulier, L., Nunzio, G.M.D., Galuščáková, P., de Herrera, A.G.S., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024). Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York (Sep 2024)
28. Macaire, C., Dion, C., Arrigo, J., Lemaire, C., Esperança-Rodier, E., Lecouteux, B., Schwab, D.: A multimodal French corpus of aligned speech, text, and pictogram sequences for speech-to-pictogram machine translation. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 839–849. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.76
29. Macaire, C., Esperança-Rodier, E., Lecouteux, B., Schwab, D.: Overview of ImageCLEFToPicto 2024 – Investigating the Translation of Natural Language into Pictograms. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France (September 9-12 2024)

30. Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B.: CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7203–7219. Association for Computational Linguistics, Online (Jul 2020), https://www.aclweb.org/anthology/2020.acl-main.645
31. Oluwafemi Ojonugwa, E.P., Rahman, M., Khalifa, F.: Advancing ai-powered medical image synthesis: Insights from medvqa-gi challenge using clip, fine-tuned stable diffusion, and dream-booth + lora. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France (September 2024)
32. 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. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (jul 2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040
33. 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. pp. 311–318 (2002)
34. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2020 concept prediction task: Medical image understanding. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
35. Popescu, A., Deshayes-Chossart, J., Schindler, H., Ionescu, B.: Overview of the imageclef 2022 aware task. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022), LNCS Lecture Notes in Computer Science, Springer, Bologna, Italy (September 5-8 2022)
36. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
37. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020), http://jmlr.org/papers/v21/20-074.html
38. Roberts, R.J.: Pubmed central: The genbank of the published literature. Proceedings of the National Academy of Sciences of the United States of America **98**(2), 381–382 (Jan 2001). https://doi.org/10.1073/pnas.98.2.381
39. Romski, M., Sevcik, R.A.: Augmentative communication and early intervention: Myths and realities. Infants & Young Children **18**(3), 174–185 (2005)
40. 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)
41. 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)
42. 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 (2024). https://doi.org/10.1038/s41597-024-03496-6, https://arxiv.org/abs/2405.10004v1
43. Sellam, T., Das, D., Parikh, A.P.: BLEURT: learning robust metrics for text generation. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. pp. 7881–7892. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.704, https://doi.org/10.18653/v1/2020.acl-main.704
44. Ştefan, L.D., Constantin, M.G., Dogariu, M., Ionescu, B.: Overview of imagecleffusion 2023 task - testing ensembling methods in diverse scenarios. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 18-21 2023)
45. 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) (sep 2011)
46. 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)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
48. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 4566–4575. IEEE Computer Society (2015). https://doi.org/10.1109/CVPR.2015.7299087, https://doi.org/10.1109/CVPR.2015.7299087
49. 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)
50. Yim, W., Ben Abacha, A., Fu, Y., Sun, Z., Yetisgen, M., Xia, F.: Overview of the mediqa-magic task at imageclef 2024: Multimodal and generative telemedicine in dermatology. In: CLEF 2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France (September 9-12 2024)
51. Yim, W., Ben Abacha, A., Snider, N., Adams, G., Yetisgen, M.: Overview of the mediqa-sum task at imageclef 2023: Summarization and classification of doctor-patient conversations. In: CLEF 2023 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 18-21 2023)
52. Yim, W., Fu, Y., Sun, Z., Ben Abacha, A., Yetisgen, M., Xia, F.: Dermavqa: A multilingual visual question answering dataset for dermatology. CoRR (2024)
53. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. ArXiv **abs/1904.09675** (2019)
54. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with BERT. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020), https://openreview.net/forum?id=SkeHuCVFDr