Overview of the ImageCLEF 2023: Multimedia Retrieval in Medical, Social Media and Internet Applications

Bogdan Ionescu[1], Henning Müller[2], Ana-Maria Drăgulinescu[1], Wen-Wai Yim[3], Asma Ben Abacha[3], Neal Snider[4], Griffin Adams[5], Meliha Yetisgen[6], Johannes Rückert[7], Alba García Seco de Herrera[8], Christoph M. Friedrich[7], Louise Bloch[7], Raphael Brüngel[7], Ahmad Idrissi-Yaghir[7], Henning Schäfer[8], Steven A. Hicks[9], Michael A. Riegler[9], Vajira Thambawita[9], Andrea M. Storås[9], Pål Halvorsen[9], Nikolaos Papachrysos[10], Johanna Schöler[10], Debesh Jha[11,9], Alexandra-Georgiana Andrei[1], Ioan Coman[1], Vassili Kovalev[13,12], Ahmedkhan Radzhabov[13], Yuri Prokopchuk[13], Liviu-Daniel Ştefan[1], Mihai-Gabriel Constantin[1], Mihai Dogariu[1], Jérôme Deshayes[14], and Adrian Popescu[14]

[1] Politehnica University of Bucharest, Romania bogdan.ionescu@upb.ro
[2] University of Applied Sciences Western Switzerland (HES-SO), Switzerland
[3] Microsoft, USA
[4] Microsoft/Nuance, USA
[5] Columbia University, USA
[6] University of Washington, USA
[7] University of Applied Sciences and Arts Dortmund, Germany
[8] University of Essex, UK
[9] SimulaMet, Norway
[10] Sahlgrenska University Hospital, Sweden
[11] Northwestern University, USA
[12] Belarus State University, Belarus
[13] Belarusian National Academy of Sciences, Belarus
[14] CEA LIST, France

Abstract. This paper presents an overview of the ImageCLEF 2023 lab, which was organized in the frame of the Conference and Labs of the Evaluation Forum – CLEF Labs 2023. ImageCLEF is an ongoing evaluation event that started in 2003 and that encourage the evaluation of the technologies for annotation, indexing and retrieval of multimodal data with the goal of providing information access to large collections of data in various usage scenarios and domains. In 2023, the 21st edition of ImageCLEF runs three main tasks: (i) a *medical* task which included the sequel of the caption analysis task and three new tasks, namely, GANs for medical images, Visual Question Answering for colonoscopy images, and medical dialogue summarization; (ii) a sequel of the *fusion* task addressing the design of late fusion schemes for boosting the performance, with two real-world applications: image search diversification (retrieval) and prediction of visual interestingness (regression); and (iii) a sequel of the *social media* aware task on potential real-life effects awareness of online image sharing. The benchmark campaign was a real success and received the participation of over 45 groups submitting more than 240 runs.

Keywords: Medical text summarization · medical image caption analysis · visual question answering · Generative Adversarial Networks (GANs) · late fusion for search diversification and interestingness prediction · prediction of effects of online image sharing · ImageCLEF lab

1 Introduction

Started in 2003 with only four participants [14], ImageCLEF[15] is the image retrieval and classification lab of the CLEF (Conference and Labs of the Evaluation Forum) conference and it rapidly increased its impact when the medical tasks were included in 2004 [13]. Then, over 20 participants were attracted. Its growing trend lead to more than 200 participants in 2019 and even more than 110 in 2020 during the COVID-19 pandemic. Even though the tasks were added, changed or discontinued, the general objective remained the same, i.e., *to combine multimodal data to retrieve and classify visual information.* Tasks have evolved along the time from more general object classification and retrieval to specific application domains, e.g., medical, Internet and social media, nature, and even security. In [32], one presents a thorough analysis of several tasks and the creation of the data sets. ImageCLEF impact over the years was assessed in [46, 47].

Starting with 2018, ImageCLEF used the crowdAI platform, which migrated to AIcrowd[16] from 2020, to distribute the data sets and receive the submitted runs. The system allowed the assignment of an online leader board and gave the opportunity to keep the data sets accessible beyond competition, including a continuous submission of runs and addition to the leader board. In 2023, the ImageCLEF team developed its own system, as migrating to the AI4Media[17] benchmarking platform (based on Codalab[18]). Over the years, ImageCLEF and also CLEF have shown a strong scholarly impact that was assessed in [46, 47]. For instance, the term “ImageCLEF” returns on Google Scholar[19] over 6,850 article results (search on June 26th, 2023). This underlines the importance of the evaluation campaigns for disseminating best scientific practices. We introduce here the three tasks that were run in the 2023 edition[20], namely: ImageCLEFmedical, ImageCLEFfusion, and ImageCLEFaware.

2 Overview of Tasks and Participation

ImageCLEF 2023 consists of three main tasks with the objective of covering a *diverse range* of multimedia retrieval applications, namely: *medicine*, *social me-*

[15] `http://www.imageclef.org/`

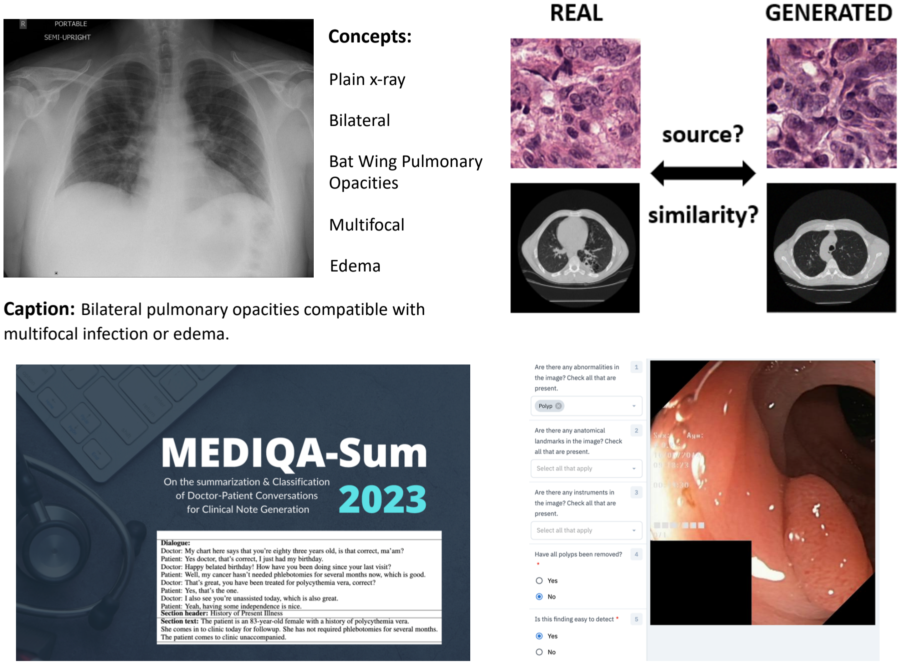
[16] `https://www.aicrowd.com/`

[17] `https://www.ai4media.eu/`

[18] `https://codalab.org/`

[19] `https://scholar.google.com/`

[20] `https://www.imageclef.org/2023/`


R
PORTABLE
SEMI-UPRIGHT
REAL
GENERATED
source?
similarity?
MEDIQA-Sum
On the summarization & Classification
of Doctor-Patient Conversations
for Clinical Note Generation
2023

Concepts:

Plain x-ray

Bilateral

Bat Wing Pulmonary Opacities

Multifocal

Edema

Caption: Bilateral pulmonary opacities compatible with multifocal infection or edema.

Dialogue:
Doctor: My chart here says that you're eighty three years old, is that correct, ma'am?
Patient: Yes doctor, that's correct, I just had my birthday.
Doctor: Happy belated birthday! How have you been doing since your last visit?
Patient: Well, my cancer hasn't needed phlebotomies for several months now, which is good.
Doctor: That's great, you have been treated for polycythemia vera, correct?
Patient: Yes, that's the one.
Doctor: I also see you're unassisted today, which is also great.
Patient: Yeah, having some independence is nice.

Section header: History of Present Illness

Section text: The patient is an 83-year-old female with a history of polycythemia vera. She comes in to clinic today for followup. She has not required phlebotomies for several months. The patient comes to clinic unaccompanied.

1 Are there any abnormalities in the image? Check all that are present.
Polyp

2 Are there any anatomical landmarks in the image? Check all that are present.
Select all that apply

3 Are there any instruments in the image? Check all that are present.
Select all that apply

4 Have all polyps been removed? *
○ Yes
● No

5 Is this finding easy to detect *
● Yes
○ No

Fig. 1: Sample images from (left to right, top to bottom): ImageCLEFmedical-caption with an image and the corresponding CUIs and captions, ImageCLEFmedical-GAN with an example of real and generated images, ImageCLEFmedical-Mediqa with an example of doctor-patient conversation, and ImageCLEFmedical-VQA with examples of questions and answers in the area of colonoscopy.

dia, and *Internet* applications. It followed the 2019 tradition [27] of diversifying the use cases [38, 45, 51, 42, 24, 1]. The 2023 tasks are presented as follows:

– **ImageCLEFmedical**. Since 2004, in the frame of ImageCLEF benchmarking, medical tasks were organised. Despite the fact that in 2018, for example, all but one task were medical, one could remark little interaction between the medical tasks. Consequently, starting with 2019, the medical tasks were focused towards one specific problem but combined as a single task with several subtasks. In this way, one could allow synergies between the domains:
  - *MEDIQA-Sum*: This is the fourth edition of the MEDIQA tasks and its first 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 ques-

tions [7]. Continuing in 2021, the next MEDIQA task resumed hosting one clinical subtask and two consumer-health question-answer related subtasks[5]. Different from the 2019 subtasks, MEDIQA 2021 focused on summarization; summarization of clinical radiology note findings, consumer health questions, and consumer health answers. This year's MEDIQA tasks include clinical dialogue section header classification, short-dialogue note summarization, and full-encounter generation. This task is introduced as part of the ImageCLEF challenges as an experimental precursor to a multimodal image and dialogue summarization task[51]. An overlapping dataset with an additional dialogue generation task was part of the ACL 2023 Clinical NLP MEDIQA-CHAT challenge [8].

- *Caption*: This is the 7th 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 [42], after the former was brought back in 2021 due to participants' demands [21, 18, 22, 35, 36, 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 2023 edition once again uses a larger dataset based on ROCO data [37], which was already used in 2019, 2020, and 2022.
- *GANs*: This is the first edition of the task [1]. The objective of the task is to investigate the hypothesis that generative models generate medical images that exhibit resemblances to the images employed during their training. This addresses concerns surrounding the privacy and security of personal medical image data in the context of generating and utilizing artificial images in various real-world scenarios. The task aims to identify distinctive features or "fingerprints" within synthetic biomedical image data, allowing us to determine which real images were used during the training process to generate the synthetic images.
- *MEDVQA-GI*: Analysis of gastrointestinal images and videos is a very popular topic in both the medical and computer science community. Usually, research and methods focus on images as a single modality. The MEDVQA-GI [24] introduces the task of visual question answering (VQA) [6, 3, 4, 20] in the field of GI endoscopy extending the modalities with text. The idea is that through the combination of text and image data, the output of the analysis gets easier to use by medical experts. For the task, a new dataset based on previously published open datasets [12, 29, 30] was developed. The extended dataset has additional data corresponding to questions regarding the type of examinations, anomaly location, number of findings, colors of the findings, to name a few.

– **ImageCLEFfusion**. This is the 2nd edition of the task [44, 45]. The main objective for this task is the development of late fusion or ensembling approaches, that are able to use prediction results from pre-computed inducers in order to generate better, improved prediction outputs. The present iteration of this task encompasses three distinct challenges: the continuation of the previous year's regression challenge utilizing media interestingness data, the continuation of the retrieval challenge involving image search result diversification data, and the addition of a new multi-label classification task focused on concepts detection in medical data. Notably, the tasks employ inducers that have been developed by actual users, ensuring their real-world applicability.

– **ImageCLEFaware**. This was the 3rd edition of the task and it focuses on personal data disclosure-awareness as users' data can be reused in other contexts when they share it for specific purposes. Consequenly, the feedback to the users is very important when dealing with the effects of personal data sharing. The objective of the task resided in automatically providing a rating of a visual user profile in different real-life situations. The dataset created specifically for the 2021 edition of the task was expanded in order to make the evaluation more robust. Data were sampled from YFCC100 dataset and were further anonymized in order to comply with GPDR.

Table 1: Key figures regarding participation in ImageCLEF 2023.

Task	Groups that submitted results	Submitted runs	Submitted working notes
Caption	13	116	12
Mediqa	12	48	12
GANs	8	40	9
MedVQA	12	14	4
Fusion	2	23	2
Aware	0	0	0
Overall	47	241	39

In order to participate in the evaluation campaign, the research groups had to register by following the instructions on the ImageCLEF 2023 web page[21]. In 2022, the challenge was organized through the AIcrowd platform[22] to ease the overall management of the campaign, but in 2023 we setup our own registration and submission system and next year we will use the AI4Media platform based on codalab[23] to manage the benchmarking campaign. As in previous year, to actually get access to the data sets, the participants were required to submit a signed End User Agreement (EUA). Table 1 summarizes the participation in ImageCLEF 2023, indicated the statistics both per task and for the overall lab. The table also shows the number of groups that submitted runs and the ones

[21] https://www.imageclef.org/2023/

[22] https://www.aicrowd.com/

[23] https://github.com/AIMultimediaLab/AI4Media-EaaS-prototype-Py2-public

that submitted a working notes paper describing the techniques used. Teams were allowed to register for several tasks.

After a decrease in participation in 2016, the participation increased in 2017 and 2018, and increased again in 2019. In 2018, 31 teams completed the tasks and 28 working notes papers were received. In 2019, 63 teams completed the tasks and 50 working notes papers were retrieved. In 2020, 40 teams completed the tasks and submitted working notes papers. In 2021, 42 teams completed the tasks and we received 30 working notes papers. In 2022, 28 teams completed the tasks and we received 26 working notes papers. In 2023, 47 teams submitted the results and we received 39 working notes, thus experiencing the revival of the campaign. Also, visual question answering, not organized in 2022, was retaken this year focusing on the text modality. Nevertheless, the number of submitted runs dropped compared to 2021 and 2022 with more teams involved 258 (2021) and 256 (2022) vs 241 (2023). This could be due to the fact that the teams were focused on finding higher-quality solutions at the expense of the numer of the runs. Thus, ImageCLEF continues to provide a strong evaluation benchmark for the community.

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 [41], Mediqa [51], GAN [1], MedVQA [24], and Fusion [45].

3 The Caption Task

The caption task was first proposed as part of the ImageCLEFmedical [22] in 2016 aiming to extract the most relevant information from medical images. Hence, the task was created to condense visual information into textual descriptions. In 2017 and 2018 [18, 21], the ImageCLEFcaption task comprised two subtasks: concept detection and caption prediction. In 2019 [35] and 2020 [36], the task concentrated on the the concept detection task, extracting Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUIs) [11] from radiology images. In 2021 [34], both subtasks, concept detection and caption prediction, were running again due to participants' demands. The focus in 2021 was on making the task more realistic by using fewer images which were all manually annotated by medical doctors. For the 2022 ImageCLEFmedical Caption task [41], both subtasks were continued albeit with an extended version of the ROCO data set used for both subtasks, which was already used in 2020 and 2019. The 2023 edition of ImageCLEFmedical caption [42] continues in the same vein, once again using a ROCO-based data set for both subtasks, but switching from BLEU [33] to BERTScore [52] as the primary evaluation metric for caption prediction.

3.1 Task Setup

The ImageCLEFmedical Caption 2023 [42] follows the format of the previous ImageCLEFmedical caption tasks. In 2023, 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) [11] 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.

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 was manually curated and only contains x-ray anatomy, directionality, and image modality concepts. For the first time this year, BERTScore was used as the primary metric for the evaluation of the caption prediction subtask, replacing the BLEU score, which had been used in previous years. BERTScore evaluates the semantic similarity of the predicted captions, whereas BLEU focuses more on n-gram overlap. 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 [2], CIDEr [49], CLIPScore [23], BLEU and BLEURT [43].

3.2 Data Set

In 2023, an extended subset of the ROCO [37] data set is used for both subtasks. The ROCO data set originates from biomedical articles of the PMC Open Access Subset[24] [40] and was extended with new images added since the last time the data set 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. New for this year was the addition of manually validated x-ray directionality concepts.

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

- *Training set* including 60,918 radiology images and associated captions and concepts.
- *Validation set* including 10,437 radiology images and associated captions and concepts.
- *Test set* including 10,473 radiology images.

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

Table 2: Performance of the participating teams in the ImageCLEFmedical 2023 concept detection subtask. The best run per team is selected. Teams with previous participation in 2022 are marked with an asterisk.

Team	Institution	F1-Score
AUEB-NLP-Group*	Department of Informatics, Athens University of Economics and Business, Athens, Greece	0.5223
KDE-Lab_Med*	KDE Laboratory, Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan	0.5074
VCMI*	University of Porto, Porto, Portugal and INESC TEC, Porto, Portugal	0.4998
IUST_NLPLAB*	School of Computer Engineering, Iran University of Science and Technology, Tehran, Islamic Republic Of Iran	0.4959
Clef-CSE-GAN-Team	SSN College Of Engineering, Chennai, India	0.4957
CS_Morgan*	Computer Science Department, Morgan State University, Baltimore, Maryland	0.4834
SSNSheerinKavitha	Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, India	0.4649
closeAI2023	Baidu Intelligent Health Unit, Beijing, China and Peng Cheng Laboratory, Shenzhen, China	0.0900
SSN_MLRG	Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, India	0.0173

3.3 Participating Groups and Submitted Runs

In the seventh edition of the ImageCLEFmedical Caption task, 27 teams registered and signed the End-User-Agreement that is needed to download the development data. 13 teams submitted 116 graded runs (12 teams submitted working notes) attracting similar attention to 2022. Each of the groups was allowed a maximum of 10 graded runs per subtask. Unlike last year, participants did not have access to their own scores until after the submission period was over. 9 teams participated in the concept detection subtask this year, 6 of those teams also participated in 2022. 13 teams submitted runs to the caption prediction subtask, 7 of those teams also participated in 2022. Overall, 9 teams participated in both subtasks, and four teams participated only in the caption prediction subtask. Unlike in 2022, no teams participated only in the concept detection subtask.

In the concept detection subtasks, the groups used primarily multi-label classification systems, with image retrieval systems consistently performing worse for teams who experimented with them. One team successfully used an image retrieval system as a fallback when the multi-label classification system did not

Table 3: Performance of the participating teams in the ImageCLEFmedical 2023 caption prediction subtask. The best run per team is selected. Teams with previous participation in 2022 are marked with an asterisk.

Team	Institution	BERTScore
CSIRO*	Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Herston, Queensland, Australia and CSIRO Data61, Imaging and Computer Vision Group, Pullenvale, Queensland, Australia and Queensland University of Technology, Brisbane, Queensland, Australia	0.6413
closeAI2023	Baidu Intelligent Health Unit, Beijing, China and Peng Cheng Laboratory, Shenzhen, China	0.6281
AUEB-NLP-Group*	Department of Informatics, Athens University of Economics and Business, Athens, Greece	0.6170
PCLmed	Peng Cheng Laboratory, Shenzhen, China and ADSPLAB, School of Electronic and Computer Engineering, Peking University, Shenzhen, China	0.6152
VCMI*	University of Porto, Porto, Portugal and INESC TEC, Porto, Portugal	0.6147
KDE-Lab_Med*	KDE Laboratory, Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan	0.6145
SSN_MLRG	Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, India	0.6019
DLNU_CCSE	Unknown	0.6005
CS_Morgan*	Computer Science Department, Morgan State University, Baltimore, Maryland	0.5819
Clef-CSE-GAN-Team	SSN College Of Engineering, Chennai, India	0.5816
Bluefield-2023	Toyohashi University of Technology, Aichi, Japan and Toyohashi Heart Center, Aichi, Japan	0.5780
IUST_NLPLAB*	School of Computer Engineering, Iran University of Science and Technology, Tehran, Islamic Republic Of Iran	0.5669
SSNSheerinKavitha*	Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, India	0.5441

predict any concepts. Last year's winners once again achieved the top scores by increasing their ensemble from two to three models.

In the caption prediction subtask, most teams experimented with encoder-decoder frameworks with different backbones and LSTM [25] decoders. Unsurprisingly, teams increasingly used Large Language Models (LLMs) in the decoding step and to help generate or refine captions. BLIP-2 [31] was used for the first time and achieved good results (second and fourth place). One novelty was the use of reinforcement learning to refine and improve upon last year's best solution in terms of BERTScore, which ended up winning this year's competition after the change of primary scores from BLEU to BERTScore.

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 subtask, the overall F1 scores increased compared to last year, which is not surprising considering the reduced number of concepts for this year's edition of the challenge.

While one team experimented with a novel autoregressive multi-label classification system that tries to model relationships between concepts and another team tried training separate models for the different modalities, these experiments did not yield better results compared to the winning approach.

BERTScore and ROUGE scores were used to predict captions. Unlike last year's edition, BERTScore replaced BLEU as the primary score for a more nuanced evaluation of captions. The adoption of BERTScore reflects the intent to prioritize semantic alignment and information preservation in the generated captions and not focus on the frequency of n-gram matches, which is the basis of BLEU.

The aforementioned change of evaluation metrics had a big effect on the outcome of the caption prediction challenge, with last year's winner placing second to last according to the BERTScore evaluation while still winning in terms of the ROUGE, BLEU and METEOR scores with a similar approach as last year. An in-depth analysis is presented in [41].

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 ROCO-based data set for both challenges after a smaller, manually annotated data set was used in 2021. Manually validated concepts for X-ray directionality information was added for this year's dataset and caption pre-processing was kept minimal. It attracted 13 teams who submitted a total of 116 graded runs, a similar level of participation to last year. Some changes were introduced for the scores, with a switch from BLEU to BERTScore as the primary evaluation metric for the caption prediction. As mentioned before, this switch had a large impact on the results, and we will continue to evaluate and explore different possible metrics or combination of metrics, but the evaluation of generated captions remains difficult.

Like last year, most teams were more successful in training multi-label classification models compared to image retrieval models for the concept detection. For the caption prediction, most teams used Transformer-based models [48], with LLMs making an appearance as part of some of the approaches.

For next year's ImageCLEFmedical Caption challenge, some possible improvements include an improved caption prediction evaluation metric which is specific to medical texts, and improving manually validated concept quality with the help of a medical professional. It will also be important to make sure that no models are used that were pre-trained on PubMedCentral data, since these models will already have seen the original captions.

4 The MEDIQA-Sum Task

The MEDIQA tasks aim to pose natural language problems related to medical language and semantics [7]. The first edition hosted the challenges of clinical note sentence NLI, as well as consumer health RQE, and answer retrieval re-ranking. The focus of the last edition, in 2021, involved summarization tasks in the areas of clinical radiology note findings, consumer health question summarization, and multiple answer summarization [5]. In 2023, two editions were hosted. The 2023 ACL Clinical NLP MEDIQA-CHAT challenge included three subtasks including short-dialogue section header and note generation, full-encounter dialogue-to-note generation, and full-encounter note-to-dialogue generation [8]. In the 2023 ImageCLEF edition, the MEDIQA-SUM subtasks included short dialogue-to-topic classification, short dialogue and topic- to note summarization, and full encounter dialogue-to-note summarization [51]

4.1 Task Setup

The MEDIQA-SUM 2023 overall task comprises three sub-tasks:

- (A) dialogue2topic (section header) classification
- (B) dialogue2note summarization given the target section header
- (C) full-encounter dialogue2note summarization.

Subtask A topic classification was evaluated using accuracy. The subtask B snippet summarization was evaluated using the mean of BLEURT, BERTscore, and ROUGE-1; metrics found to be correlated to human evaluation in several independent health summarization datasets [9]. Full-encounter summarization in Subtask C used two metrics: (1) a full-note ROUGE-1 score and (2) an equally weighted division-based (subjective, objective_exam, objective_results, assessment_and_plan) aggregate score of the BLEURT, BERTscore, and ROUGE-1 metric.

Subtask A and B use the same test set. After Subtask A was closed, the gold standard section header was released so that it would be available as input to Subtask B. Code submissions were required at submission. The organizers checked output of code against submitted runs and documented each team's code replicability status.

4.2 Data Set

The 2023 MEDIQA-SUM challenge includes data from two collections: MTS-Dialog [10] and ACI-BENCH [50]. Subtasks A and B consist of 1,201 pairs of conversations and associated section headers and contents; 100 examples in validation, and 200 pairs in test. Subtask C includes full encounters with 67 examples in training, 20 in validation, and 40 in test.

Table 4: Performance of the participating teams in the MEDIQA-Sum 2023 Subtask A on topic classification. The best run per team is selected.

Team	Institution	Accuracy
Cadence	Cadence Solutions, USA	0.820
HuskyScribe	University of Washington, USA	0.815
Tredence	Tredence Inc, India	0.800
StellEllaStars	University of Michigan School of Information, USA	0.765
SSNSheerinKavitha	Sri Sivasubramaniya Nadar College of Engineering, India	0.740
SuryaKiran	Optum, India	0.735
SSNdhanyadivyakavitha	Sri Sivasubramaniya Nadar College of Engineering, India	0.720
ds4dh	University of Geneva, Switzerland	0.710
uetcorn	University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam	0.710
SKKU-DSAIL	Department of Applied Artificial Intelligence, Sungkyunkwan University, South Korea	0.700
MLRG-JBTTM	Sri Sivasubramaniya Nadar College of Engineering, India	0.665

4.3 Participating Groups and Submitted Runs

Overall 12 teams participated with a total of 48 runs. Subtask A included 23 valid submissions among 11 teams. Subtask B included 16 submissions among 7 teams. Subtask C included 9 submissions among 4 teams. At most three runs were allowed per team in each subtask. With the exception of 1 team, all teams participated in Subtask A. Four teams participated in two subtasks. Three teams participated in all three subtasks.

4.4 Results

The best teams achieved 0.8 Accuracy on Subtask A topic classification (Table 4) and an aggregate score of 0.43 for Subtask B (Table 5). The top two systems for Subtask C achieved ROUGE-1 at 0.49 F1 (Table 6) and aggregated scores at 0.44 (Table 7).

Table 5: Performance of the participating teams in the MEDIQA-Sum 2023 Subtask B on dialogue2note summarization. The best run per team is selected.

Team	Institution	Aggregated Score
SuryaKiran	Optum, India	0.573
PULSAR	ASUS AICS / University of Manchester, Singapore/UK	0.569
Tredence	Tredence Inc, India	0.559
HuskyScribe	University of Washington, USA	0.529
uetcorn	University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam	0.481
SKKU-DSAIL	Department of Applied Artificial Intelligence, Sungkyunkwan University, South Korea	0.461
SSNSheerinKavitha	Sri Sivasubramaniya Nadar College of Engineering, India	0.419

Table 6: Performance of the participating teams in the MEDIQA-Sum 2023 Subtask C on full-encounter dialogue2note summarization, ranked by ROUGE-1. The best run per team is selected.

Team	Institution	ROUGE-1
Tredence	Tredence Inc, USA	0.500
uetcorn	University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam	0.498
HuskyScribe	University of Washington, USA	0.470
PULSAR	ASUS AICS / University of Manchester, Singapore/UK	0.294

Table 7: Performance of the participating teams in the MEDIQA-Sum 2023 Subtask C on full-encounter dialogue2note summarization, ranked by the aggregated score. The best run per team is selected.

Team	Institution	Aggregated Score
Tredence	Tredence Inc, USA	0.455
uetcorn	University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam	0.441
HuskyScribe	University of Washington, USA	0.413
PULSAR	ASUS AICS / University of Manchester, Singapore/UK	0.247

Subtask A submissions included classic machine learning algorithms such as SVM, KNN, Random Forest, with some pre-processing such as TF-IDF, lemmatization. This task also featured the use of pre-trained models such as GPT3.5, clinical-BERT, clinical T5, and their low-rank adaptation (LoRA). Eight out of 23 submissions either used additional training data or adjusted data sampling.

Subtask B primarily consisted of pre-trained sequence-to-sequence models such as llama, bart, flan T5, biobart, and their LoRA versions, fine-tuned on the training and validation sets. Eight out of 16 submissions used the gold standard section headers released from Subtask A.

Subtask C submissions had a diverse set of systems that used creative means to circumvent a low-resource generation problem. Specifically, Uetcorn, HuskyScribe, and Tredence all divided the problem into multiple parts. Firstly, relevant parts of the dialogue were grouped together as related to particular sections. Each team used a different method to achieve this; the UETCorn team identified relevant parts of dialogue for specific note section key points (e.g. "chief complaint" or "medications"), using a similarity function between dialogue sentences and a hand-crafted section-specific description; afterwards, several note generation strategies were used for each key point. HuskyScribe built a model classifying smaller dialogue exchanges into the same categories, while Tredence classified dialogues chunked by various window sizes. In the second step, grouped dialogue chunks were sent through a text generator to produce parts of the note. The use of pre-trained models such as BART/BioBART and flan T5 for the generation was typical. The Uetcorn and Tredence team included some section/key-point specific questions as part of the generation input, e.g. (e.g. input: "question: {question} context: {conversation}", output: summary). The Uetcorn team also experimented with a reading comprehension answer extraction based on specially designed key point query (e.g. "names of medication used") and post-processing. The HuskyScribe team additionally used Subtask A data to generate additional synthetic data for training. Finally, the completed note was assembled through concatenation and post-processing. Unlike the other three groups, the PULSAR team employed an end-to-end approach, experimenting with flan T5 and llama models with additional data created using MTSamples data processed through GPT3.5.

With the exception of two runs in subtask A and one team's runs in Subtask B, all submissions were reproducible based on participants' submitted code. A more detailed account can be found in the MEDIQA-Sum overview paper [51].

4.5 Lessons Learned and Next Steps

This year's MEDIQA tasks hosted similar problems on an overlapping dataset with the 2023 ACL ClinicalNLP MEDIQA-Chat Shared Tasks [8]. A striking difference between the participants in this edition was that there were no GPT4 submissions. As GPT4 access requires a subscription, we can view the solutions from this evaluation lab as a whole to be solutions constrained to only using open-source or free models and data. A more detailed comparison can be found in the MEDIQA-Sum overview paper [51].

The requirement of code submissions in this year's MEDIQA challenges was successful and ensured that final submissions would be of high quality; it also encouraged the release of open-source code into the community beyond the challenge. In this year's edition, the code was run manually by the organizers. In future editions, we will explore the use of platforms, e.g. https://codalab.org/, that will provide a standard management and packaging pipeline allowing submissions to be more easily and quickly evaluated.

Natural language evaluation is a challenging and active area of research. Evaluation for long documents is even less explored. While we used several metrics associated with human-labeled facts for our dataset (ROUGE-1 and an aggregated BLEURT, BERTScore, ROUGE1 metric), new metrics can be further explored for future challenges.

In our next future edition, we plan on running a multi-modal medical dialogue summarization task - we will use the lessons learned in this edition.

5 The GANs Task

The development of generative models in the area of artificial intelligence in recent years has generated a great deal of attention and creativity, altering many industries and the way we approach different tasks. Task offered an environment for investigating GANs' effects on the creation of synthetic medical images by providing a benchmark to explore the impact of GANs on artificial biomedical image generation. Medical image generation is essential for patient care improvement, healthcare professional education, and medical research. While obtaining genuine patient data can be expensive, insufficient, or ethically problematic, the ability to generate artificial yet realistic biological images can fill these gaps and provide researchers, doctors, and educators more leverage. As a result, generative models have shown to be remarkably effective at producing high-quality images that closely resemble the traits and patterns of real data.

5.1 Task Setup

This is the first edition of the task and consists of one challenge. The task aims to identify distinctive features or "fingerprints" within synthetic biomedical image data, allowing us to determine which real images were used during the training process to generate the synthetic images.

5.2 Data Set

A data set containing axial chest CT scans of lung tuberculosis patients was provided for the task. This means that some of them may appear pretty "normal" whereas the others may contain certain lung lesions including the severe ones. These images are stored in the form of 8 bit/pixel PNG images with dimensions of 256 × 256 pixels. The artificial slice images are 256 × 256 pixels in size. All of them were generated using Diffuse Neural Networks.

- *Development (Train) dataset*: consists of 500 artificial images and 160 real images annotated according to their use in the training of the generative network. Out of the real images, 80 were used during training.

– *Test (Evaluation) dataset* was created in similar way. The only difference is that the two subsets of real images are mixed and no proportion of non-used and used ones has been disclosed. Thus, a total of 10,000 generated and 200 real images are provided.

5.3 Participating Groups and Submitted Runs

Overall, 23 teams registered to the task, 8 of them finalizing the task and submitting runs. A total of 40 runs were received.

5.4 Results

An analysis of the proposed methods shows a great diversity among them, ranging from texture analysis, similarity-based approaches that join inducer predictions like SVM or KNN, to deep learning approaches and even multi-stage transfer learning. More detailed results, including methods presentation and other performance measures, are presented in the overview article [1]. 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 8.

5.5 Lessons Learned and Next Steps

The first edition of the ImageCLEF medical GANs task attracted a total of 8 teams that submitted runs, with all of them completing their submissions by creating a working notes papers. A prediction-based task was proposed to the participants. The best result for the task is an F1-score of 0.802 obtained by VCMI team followed by PicusLabMed with an F1-score of 0.666 and AIMultimediaLab with an F1-score of 0.626. We are pleased to report a high level of diversity in the identification strategies put forth by the participants. Future iterations of this task will diversify various elements, such as datasets and generation techniques, and broaden the study fields of synthetic medical data. We also intend to add more tasks based on various aspects of the security and privacy of the created data.

6 The MedVQA-GI Task

Identifying lesions in colonoscopy images is one of the most popular applications of artificial intelligence in medicine. Until now, the research has focused on single-image or video analysis. With this task, we aim to bring a new aspect to the field by adding multiple modalities to the picture. The main focus of the task will be on visual question answering (VQA) and visual question generation (VQG). The goal is that through the combination of text and image data, the output of the analysis gets easier to use by medical experts. The task has three sub-tasks.

Table 8: Summary on the participant submissions and their results for GAN task.

Group rank	**Group name**	**Submission #**	**F1-score**
#1	VCMI	submission 2	0.802
#2	VCMI	submission 1	0.731
#3	VCMI	submission 3	0.707
#4	PicusLabMed	submission 8	0.666
#5	VCMI	submission 4	0.654
#6	AIMultimediaLab	submission 1	0.626
#7	PicusLabMed	submission 6	0.624
#8	VCMI	submission 5	0.621
#9	Clef-CSE-GAN-Team	submission 1	0.614
#10	VCMI	submission 7	0.613
#11	VCMI	submission 6	0.605
#12	VCMI	submission 10	0.594
#13	AIMultimediaLab	submission 2	0.585
#14	one five one zero	submission 2	0.563
#15	PicusLabMed	submission 9	0.562
#16	PicusLabMed	submission 4	0.552
#17	KDE lab	submission 5	0.548
#18	one five one zero	submission 3	0.522
#19	Clef-CSE-GAN-Team	submission 2	0.521
#20	VCMI	submission 9	0.514
#21	one five one zero	submission 1	0.507
#22	GAN-ISI	submission 5	0.502
#23	GAN-ISI	submission 2	0.489
#24	PicusLabMed	submission 10	0.487
#25	GAN-ISI	submission 3	0.486
#26	GAN-ISI	submission 4	0.483
#27	DMK	submission 1	0.480
#28	PicusLabMed	submission 2	0.470
#29	KDE lab	submission 2	0.469
#30	GAN-ISI	submission 1	0.469
#31	KDE lab	submission 1	0.465
#32	KDE lab	submission 4	0.457
#33	DMK	submission 2	0.449
#34	VCMI	submission 8	0.448
#35	PicusLabMed	submission 1	0.434
#36	Clef-CSE-GAN-Team	submission 3	0.431
#37	PicusLabMed	submission 3	0.419
#38	PicusLabMed	submission 5	0.417
#39	KDE lab	submission 3	0.407
#40	PicusLabMed	submission 7	0.093

For the VQA subtask, the participants need to combine images and text answers to answer the questions. In the VQG subtask, the participants are asked to generate text questions from a given image and answer. Example questions for both VQA and VQG: How many polyps are in the image? Are there any polyps in the image? What disease is visible in the image? The third subtask is the visual location question answering (VLQA), where the participants get an image and a question and are required to answer it by providing a segmentation mask for the object in the question. Example questions are: Where exactly in the image is the polyp? Where exactly in the image is the instrument?

6.1 Task Setup

The task had three sub-tasks that the participants could work on. There was no requirement on which task should be finished or not. For the first sub-task (VQA), participants were asked to generate text answers given a text question and image pair. For subtask 2 (VQG), the task was to generate questions based on a given text answer and image pair. The final subtask (VLQA) asked the participants to segment parts of an image given a text question and image pair. For the different tasks, we used different metrics to evaluate the performance. More details on the tasks and evaluation metrics can be found in the task overview paper [24].

6.2 Data Set

The dataset consisted of images from the GI tract and ground truth regarding specific questions and answers related to the images, and was based on open GI data sets previously published by the organizers [12, 29, 30]. The data set was developed with medical experts having several years of experience working in GI endoscopy. Moreover, segmentation masks were included for subtask 3, since the subtask asked for segmentation masks as answers to input pairs of images and textual questions. For the challenge, the dataset was split in two, a development dataset and a testing dataset. The development dataset contained 2,000 samples (imaged and question-answer pairs), and the testing dataset consisted of 1,949 samples. The participants were only provided with the ground truth for the development dataset. The data and evaluation scripts will be made publicly available after the competition of the challenge.

6.3 Results

In total, 16 valid runs were submitted to the task from 8 different teams. One team did not submit their task description paper. Overall, the teams achieved reasonably good results ranging from an accuracy of around 0.21 to 0.82 for subtask 1. For subtask 3, four teams submitted a solution, and there we observed a large performance difference with IoU ranging from 0.234 to 0.666. For subtask 2, teams only submitted an inverse of subtask 1, which was not a meaningful

Table 9: An overview of the results for each task available at MedVQA-GI.

Team Name	Task 1 (Accuracy)	Task 2	Task 3 (IoU)
wsq4747	0.740	-	0.234
BITM	0.819	-	-
SSNSheerinKavitha	0.441	-	-
SSN_KDC	0.820	-	-
utk	0.471	-	-
VisionQAries	0.548	-	0.666
DLNU_CCSE	0.213	-	-
UIT-Saviors	0.752	-	-

way to approach the task. In future iterations of the task, we will consider this and create a totally separate ground truth in addition to more strict task requirements. Table 9 provides an overview of all teams and their metrics for the different subtasks.

6.4 Lessons Learned and Next Steps

Overall, we observe quite some interest in the task, with many teams signing up. We also experienced that the task was somehow perceived as difficult due to the different modalities. One important lesson we learned is that subtask 2 could have worked better, and teams only submitted an inverse of subtask 1, which was difficult to evaluate in a meaningful way. In conclusion, there was great interest in the task, and it was shown that the problem is complex but not impossible. We plan to extend the ground truth and refine some of the tasks for future iterations.

7 The Fusion Task

The generalization ability and performance of machine learning models show signs of reaching a plateau in many domains, where the performance improvements over the years are not significant. Therefore, exploring the performance and optimizing the efficiency of machine learning methods is important for real-world applications as they can only use limited, noisy data. In this context, fusion methods are gaining popularity by harnessing the complementary knowledge of multiple base models to build more robust and accurate models compared with single models.

Several challenges must be explored by the participants in this task, such as *diversity*, which refers to a set of classifiers that, given the same instance, output different predictions; *voting mechanism*, which regulate how individual outputs from the base models are used during prediction; *dependency*, which refers to the way a base model affects the construction of the next model in the fusion chain; *cardinality*, which refers to the number of individual base models that form the ensemble – one needs to find a balance, as diversity may be reduced if too many models are incorporated in the fusion; the *learning mode* of the base models, which is the property that balance the classifiers' ability to adapt properly to new, previously unseen, data while at the same time retaining the previously learned knowledge.

7.1 Task Setup

This second edition of the ImageCLEFfusion task [45] consists of three challenges: a regression challenge involving media interestingness (ImageCLEFfusion–int) for which we provide output data from 29 inducers, a retrieval challenge involving result diversification (ImageCLEFfusion-div) for which we provide outputs data from 56 inducers, and a multi-label classification task involving concepts detection in medical data (ImageCLEFfusion–cap) for which we provided 84 inducers. Participants were required to devise late fusion learning strategies based on the outputs of the inducers associated with the media samples for each of the subtasks. The evaluation of the participants' submissions was conducted using the Mean Average Precision at 10 (mAP@10) metric for the ImageCLEFfusion–int task, F1 at 20 (F1@20) and Cluster Recall at 20 (Cluster Recall@20) metrics for the ImageCLEFfusion-div task, and the F1 metric for the ImageCLEFfusion–cap task. Participants were encouraged to submit their solutions for all three tasks.

7.2 Data Set

The three tasks in ImageCLEFfusion make use of different datasets and associated challenges. The ImageCLEFfusion–int task focuses on the Interestingness10k dataset [16], specifically utilizing the image-based prediction data from the 2017 MediaEval Predicting Media Interestingness task [17]. In this task, we provide prediction outputs from 29 systems that were submitted during the benchmarking task. To facilitate training and testing, the available data is divided into 1877 samples for training the fusion systems and 558 samples for testing.

On the other hand, the ImageCLEFfusion–div task relies on the Retrieving Diverse Social Images dataset [28], specifically targeting the DIV150Multi challenge [26]. For this task, we provide retrieval outputs from 56 systems, which are further divided into 60 queries for the training data and 63 queries for the testing set.

Lastly, the ImageCLEFfusion–cap task is derived from the ImageCLEF Medical Caption Task [41]. This task involves the extraction of multi-label outputs from 84 inducers. The data used for this task consists of 6101 images for the development set and 1500 images for the testing set.

In the training sets of all three tasks, we provide participants with the inducer outputs, along with the requisite scripts for metric computation. Additionally, the performance of each inducer is disclosed based on the official metrics, and ground truth data is made available. However, for the testing sets, only the inducer outputs are provided. It is crucial to emphasize that participants were strictly prohibited from utilizing external inducers. They were solely permitted to employ the inducers we provided. This constraint ensures a fair assessment of the performance of the late fusion approach, without introducing any alterations to the inducer set.

Table 10: Participation in the ImageCLEF-int 2023 task: the best score from all runs for each team. We also included a baseline that consists of the average performance of all the provided inducers.

Team	*#Runs*	*mAP@10*
SSN CSE-ML [39]	10	0.1331
CS_Morgan [19]	3	0,1287
baseline	-	0.0946

Table 11: Participation in the ImageCLEF-div 2023 task: the best score from all runs for each team. We also included a baseline that consists of the average performance of all the provided inducers.

Team	*#Runs*	*F1@20*	*CR@20*
SSN CSE-ML [39]	10	0.5708	0.449
baseline	-	0.5313	0.414

7.3 Participating Groups and Submitted Runs

Twelve teams have officially registered for the ImageCLEFfusion competition, showcasing a strong level of interest in participating. Out of these teams, two have successfully submitted their runs and fulfilled the competition requirements by providing detailed working notes that outline their methodologies. As for the ImageCLEF-int task, both teams combined have submitted a total of thirteen runs, while one team alone has submitted ten runs for the ImageCLEF-div task. There have been no recorded runs for the ImageCLEFfusion–cap task.

7.4 Results

The results are presented in Table 10 for the interestingness task, and Table 11 for the diversification task. The participating teams employed a diverse range of techniques for the tasks. For the result diversification task, they explored various machine learning algorithms including Elastic Net, Gradient Boosting Regressor, and Decision Tree. In addition, for the image interestingness task, they utilized XGBoost Classifier, k-Nearest Neighbors Classifier, and Decision Tree. A voting classifier and an ensemble learning model based on StackingClassifier were also tested that combined the three base models for each task. The results demonstrate the superiority of the ensemble learning approach over the other tested methods in both subtasks. For the diversification task, the ensemble learning approach achieved an F1 score of 0.5708 and a Cluster Recall (CR) score of 0.449. In the interestingness task, the ensemble learning approach achieved a mean Average Precision at 10 (mAP@10) score of 0.1331.

7.5 Lessons Learned and Next Steps

Despite the reduced number of participants compared to the previous year, with only two teams submitting runs for both the ImageCLEFfusion–int and ImageCLEFfusion–div tasks, the participating teams achieved a performance that surpassed the majority of the participants in the previous year, but still under the state-of-the-art result of the last year achieved by [15].

For the next edition of this task, we believe it is very important to continue with these three datasets, as this will allow us to study the year-to-year improvement of the proposed fusion techniques.

8 The Aware Task

Social networks engage the users to share their personal data in order to interact with other users. The context of the sharing is chosen by the users but they do not have control on further data use. These data are automatically aggregated into profiles which are exploited by social networks to propose personalized advertising/services to users. Depending on their visibility, data can be also consulted by other entities to make decisions which have a high impact on the user's life. It is thus important to give users feedback about the potential real-life effects of their personal data sharing.

We designed a task focused on the automatic rating of visual user profile in four impactful situations. Each profile includes 100 photos and its appeal is manually evaluated via crowdsourcing. Participants are asked to provide automatic visual profile ratings obtained by using a training set which includes visual- and situation-related information. These ratings are then ranked and compared to manual ones in order to assess the feasibility of providing automatic feedback related to the effects of personal photos sharing.

Six teams registered for the task this year, but, unfortunately, none of them submitted runs. Given the low interest for the task, there will be no next edition. However, the datasets and evaluation scripts will be kept available in case other research teams will be interested in working with them later.

9 Conclusion

This paper presents a global picture of the tasks and outcomes of the ImageCLEF 2023 benchmarking campaign. Three main tasks were organised, covering challenges in the medical domain (caption analysis, visual question answering, medical dialogue summarisation, GANs for medical image generation) and social networks and Internet (analysis of the real-life effects of personal data sharing, fusion techniques for retrieval and interestingness prediction). With respect to the previous year, we experienced a 67% increase in the number of teams completing the tasks (28 in 2022 vs. 47 in 2023). They successfully submitted 241 runs and 39 working notes papers.

As in the previous year, almost all solutions provided by the participants were based on machine learning and deep learning techniques. In ImageCLEF-caption, multi-label classification systems were used, as well as image retrieval systems, the latter performing worse. Mediqa task determined the participants to use classic machine learning algorithms such as SVM, KNN, Random Forest, with pre-processing methods such as TF-IDF, lemmatization. In addition, the participants used pre-trained models such as GPT3.5, clinical-BERT, clinical T5, and their low-rank adaptation (LoRA). For ImageCLEF-GAN task, the participants explored a large variety of methods as texture analysis, similarity-based approaches that join inducer predictions like SVM or KNN, and even deep learning approaches and multi-stage transfer learning. For ImageCLEF-MedVQA, the participants employed transformer-based pre-trained models. In ImageCLEFfusion, being at the 2nd edition, the participants explored machine learning algorithms as Elastic Net, Gradient Boosting Regressor, Decision Tree, XGBoost Classifier, and k-Nearest Neighbors Classifier. In ImageCLEFaware, the participation decreased even more and no run was submitted. ImageCLEF 2023 provided to the participants and to the community an interesting symbiosis of tasks and approaches and we are looking forward to participating at the CLEF 2023 workshop and to present the current achievements and the future plans.

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 ImageCLEFaware, ImageCLEFfusion tasks. The work of Louise Bloch and Raphael Brüngel was partially funded by a PhD grant from the University of Applied Sciences and Arts Dortmund (FH Dortmund), Germany. The work of Ahmad Idrissi-Yaghir and Henning Schäfer was funded by a PhD grant from the DFG Research Training Group 2535 Knowledge- and data-based personalisation of medicine at the point of care (WisPerMed).

References

1. 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)

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. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005), `https://aclanthology.org/W05-0909`
3. Ben Abacha, A., Datla, V.V., Hasan, S.A., Demner-Fushman, D., Müller, H.: Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain. In: CLEF 2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
4. Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: VQA-Med: Overview of the medical visual question answering task at imageclef 2019. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)
5. 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`
6. Ben Abacha, A., Sarrouti, M., Demner-Fushman, D., Hasan, S.A., Müller, H.: Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In: CLEF 2021 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania (September 21-24 2021)
7. Ben Abacha, A., Shivade, C., Demner-Fushman, D.: Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In: Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019. pp. 370–379. Association for Computational Linguistics (2019), `https://doi.org/10.18653/v1/w19-5039`
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., wai Yim, W., Michalopoulos, G., Lin, T.: An investigation of evaluation metrics for automated medical note generation (2023)
10. Ben Abacha, A., Yim, W.w., Fan, Y., Lin, T.: An empirical study of clinical note generation from doctor-patient encounters. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 2291–2302. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023), `https://aclanthology.org/2023.eacl-main.168`
11. 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
12. 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
13. Clough, P., Müller, H., Sanderson, M.: The CLEF 2004 cross–language image retrieval track. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign. Lecture Notes in Computer Science (LNCS), vol. 3491, pp. 597–613. Springer, Bath, UK (2005)

14. Clough, P., Sanderson, M.: The CLEF 2003 cross language image retrieval task. In: Proceedings of the Cross Language Evaluation Forum (CLEF 2003) (2004)
15. Constantin, M.G., Ştefan, L.D., Dogariu, M., Ionescu, B.: Ai multimedia lab at imagecleffusion 2022: Deepfusion methods for ensembling in diverse scenarios. In: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Bologna, Italy (2022)
16. Constantin, M.G., Ştefan, L.D., Ionescu, B., Duong, N.Q., Demarty, C.H., Sjöberg, M.: Visual interestingness prediction: A benchmark framework and literature review. International Journal of Computer Vision **129**(5), 1526–1550 (2021)
17. Demarty, C.H., Sjöberg, M., Ionescu, B., Do, T.T., Gygli, M., Duong, N.: Mediaeval 2017 predicting media interestingness task. In: MediaEval workshop (2017)
18. Dicente Cid, Y., Kalinovsky, A., Liauchuk, V., Kovalev, V., , Müller, H.: Overview of ImageCLEFtuberculosis 2017 - predicting tuberculosis type and drug resistances. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
19. Emon, I.S., Rahman, M.: Media interestingness prediction in imagecleffusion 2023 with dense architecture-based ensemble & scaled gradient boosting regressor model. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 18-21 2023)
20. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Müller, H.: Overview of the ImageCLEF 2018 medical domain visual question answering task. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
21. García Seco de Herrera, A., Eickhoff, C., Andrearczyk, V., , Müller, H.: Overview of the ImageCLEF 2018 caption prediction tasks. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
22. 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)
23. 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`
24. 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)
25. 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`
26. Ionescu, B., Gînscă, A.L., Boteanu, B., Lupu, M., Popescu, A., Müller, H.: Div150multi: a social image retrieval result diversification dataset with multi-topic queries. In: Proceedings of the 7th international conference on multimedia systems. pp. 1–6 (2016)

27. 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)
28. Ionescu, B., Rohm, M., Boteanu, B., Gînscă, A.L., Lupu, M., Müller, H.: Benchmarking image retrieval diversification techniques for social media. IEEE Transactions on Multimedia **23**, 677–691 (2020)
29. 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`
30. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: Proceeding of the International Conference on Multimedia Modeling (MMM). vol. 11962, pp. 451–462 (2020), `https://doi.org/10.1007/978-3-030-37734-2_37`
31. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. CoRR **abs/2301.12597** (2023). https://doi.org/10.48550/arXiv.2301.12597, `https://doi.org/10.48550/arXiv.2301.12597`
32. 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)
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. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (jul 2002). https://doi.org/10.3115/1073083.1073135, `https://aclanthology.org/P02-1040`
34. Pelka, O., Ben Abacha, A., García Seco de Herrera, A., Jacutprakart, J., Friedrich, C.M., Müller, H.: Overview of the ImageCLEFmed 2021 concept & caption prediction task. In: CLEF2021 Working Notes. pp. 1101–1112. CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania (September 21-24 2021)
35. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2019 concept prediction task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)
36. 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)
37. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): a multimodal image dataset. In: Proceedings of the Third International Workshop on Large-Scale Annotation of Biomedical Data and Expert

Label Synthesis (LABELS 2018), Held in Conjunction with MICCAI 2018. vol. 11043, pp. 180–189. LNCS Lecture Notes in Computer Science, Springer, Granada, Spain (September 16 2018)

38. 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)
39. Prabavathy, B., Sai, G.G., Kishore, N., Olirva, M., Vaibhav, A.M., Murali, N.S., Harshith, P.S.: Efficient fusion techniques for result diversification and image interestingness tasks. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 18-21 2023)
40. 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
41. 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)
42. 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)
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 2022 task-ensembling methods for media interestingness prediction and result diversification. In: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Bologna, Italy (2022)
45. 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)
46. 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)
47. 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)
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on

Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008 (2017), https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

49. 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
50. wai Yim, W., Fu, Y., Abacha, A.B., Snider, N., Lin, T., Yetisgen, M.: Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation (2023)
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. 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