Overview of ImageCLEFmedical 2023 – Caption Prediction and Concept Detection

Johannes Rückert[1,*], Asma Ben Abacha[2], Alba G. Seco de Herrera[3], Louise Bloch[1,4,5,†], Raphael Brüngel[1,4,5,†], Ahmad Idrissi-Yaghir[1,†], Henning Schäfer[6,†], Henning Müller[7,8] and Christoph M. Friedrich[1,4]

[1]*Department of Computer Science, University of Applied Sciences and Arts Dortmund, Dortmund, Germany*
[2]*Microsoft, Redmond, Washington, USA*
[3]*University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK*
[4]*Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Germany*
[5]*Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen, Germany*
[6]*Institute for Transfusion Medicine, University Hospital Essen, Essen, Germany*
[7]*University of Applied Sciences Western Switzerland (HES-SO), Switzerland*
[8]*University of Geneva, Switzerland*

Abstract

The ImageCLEFmedical 2023 Caption task on caption prediction and concept detection follows similar challenges held from 2017–2022. The goal is to extract Unified Medical Language System (UMLS) concept annotations and/or define captions from image data. Predictions are compared to original image captions. Images for both tasks are part of the Radiology Objects in COntext version 2 (ROCOv2) dataset. For concept detection, multi-label predictions are compared against UMLS terms extracted from the original captions with additional manually curated concepts via the F1-score. For caption prediction, the semantic similarity of the predictions to the original captions is evaluated using the BERTScore. The task attracted strong participation with 27 registered teams, 13 teams submitted 116 graded runs for the two subtasks. Participants mainly used multi-label classification systems for the concept detection subtask, the winning team AUEB-NLP-Group used an ensemble of three CNNs. For the caption prediction subtask, most teams used encoder-decoder architectures, with the winning team CSIRO using an encoder-decoder framework with an additional reinforcement learning optimization step.

Keywords

ImageCLEF, Computer Vision, Multi-Label Classification, Image Captioning, Image Understanding, Radiology

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Corresponding author.
†These authors contributed equally.
johannes.rueckert@fh-dortmund.de (J. Rückert); abenabacha@microsoft.com (A. Ben Abacha); alba.garcia@essex.ac.uk (A. G. Seco de Herrera); louise.bloch@fh-dortmund.de (L. Bloch); raphael.bruengel@fh-dortmund.de (R. Brüngel); ahmad.idrissi-yaghir@fh-dortmund.de (A. Idrissi-Yaghir); henning.schaefer@uk-essen.de (H. Schäfer); henning.mueller@hevs.ch (H. Müller); christoph.friedrich@fh-dortmund.de (C. M. Friedrich)
0000-0002-5038-5899 (J. Rückert); 0000-0001-6312-9387 (A. Ben Abacha); 0000-0002-6509-5325 (A. G. Seco de Herrera); 0000-0001-7540-4980 (L. Bloch); 0000-0002-6046-4048 (R. Brüngel); 0000-0003-1507-9690 (A. Idrissi-Yaghir); 0000-0002-4123-0406 (H. Schäfer); 0000-0001-6800-9878 (H. Müller); 0000-0001-7906-0038 (C. M. Friedrich)

1. Introduction

ImageCLEF[1] is the image retrieval and classification lab of the CLEF (Conference and Labs of the Evaluation Forum) conference. ImageCLEF 2023 consists of the ImageCLEFmedical, ImageCLEFfusion and ImageCLEFaware labs, with the ImageCLEFmedical lab being divided into the subtasks MEDIQA-Sum (natural language semantic retrieval), Caption, GANs (generation of medical images), and MEDVQA-GI (gastrointestinal visual question answering).

The Caption task was first proposed as part of the ImageCLEFmedical [1] in 2016. In 2017 and 2018 [2, 3] the ImageCLEFmedical caption task comprised two subtasks: concept detection and caption prediction. In 2019 [4] and 2020 [5], the task concentrated on the concept detection subtask extracting Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUIs) [6] from radiology images.

In 2021 [7], 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. As additional data of similar quality is hard to acquire, the 2022 ImageCLEFmedical caption task [8] continued with both subtasks albeit with an extended version of the Radiology Objects in COntext (ROCO) [9] dataset used for both subtasks, which was already used in 2020 and 2019. The 2023 edition of ImageCLEFmedical caption continues in the same vein, once again using a ROCO-based dataset for both subtasks but switching from BLEU to BERTScore as the primary evaluation metric for caption prediction.

This paper sets forth the approaches for the caption task: automated cross-referencing of medical images and captions into predicted coherent captions and UMLS concept detection in radiology images as a separate subtask. This task is a part of the ImageCLEF benchmarking campaign, which has proposed medical image understanding tasks since 2003; a new suite of tasks is generated each subsequent year. Further information on the other proposed tasks at ImageCLEF 2023 can be found in Ionescu et al. [10].

This is the 7th edition of the ImageCLEFmedical caption task. Just like in 2016 [1], 2017 [2], 2018 [3], 2021 [7], and 2022 [8] both subtasks of concept detection and caption prediction are included in ImageCLEFmedical Caption 2023. Like in 2022, an extended subset of the ROCO [9] dataset is used, with images that are not licensed CC BY or CC BY-NC removed.

Manual generation of the knowledge of medical images is a time-consuming process prone to human error. As this process requires assistance for the better and easier diagnoses of diseases that are susceptible to radiology screening, it is important that we better understand and refine automatic systems that aid in the broad task of radiology-image metadata generation. The purpose of the ImageCLEFmedical 2023 caption prediction and concept detection tasks is the continued evaluation of such systems. Concept detection and caption prediction information is applicable to unlabelled and unstructured datasets and medical datasets that do not have textual metadata. The ImageCLEFmedical caption task focuses on the medical image understanding in the biomedical literature and specifically on concept extraction and caption prediction based on the visual perception of the medical images and medical text data such as medical caption or UMLS CUIs paired with each image (see Figure 1).

[1] https://www.imageclef.org/ [last accessed: 2023-06-28]

In 2023, for the development data, an extended subset of the ROCO [9] dataset from 2022 was used, with new images from the same source added for the validation and test sets, while images from articles with licenses other than CC BY and CC BY-NC were removed.

This paper presents an overview of the ImageCLEFmedical caption task 2023 including the task and participation in Section 2, the data creation in Section 3, and the evaluation methodology in Section 4. The results are described in Section 5, followed by conclusion in Sections 6.

2. Task and Participation

In 2023, the ImageCLEFmedical caption task consisted of two subtasks: concept detection and caption prediction.

The concept detection subtask follows the same format proposed since the start of the task in 2017 [2]. Participants are asked to predict a set of concepts defined by the UMLS CUIs [6] based on the visual information provided by the radiology images.

The caption prediction subtask follows the original format of the subtask used between 2017 and 2018 [2, 3]. This subtask was paused and it is running again since 2021 because of participant demand. This subtask aims to automatically generate captions for the radiology images provided.

In 2023, 27 teams registered and signed the End-User-Agreement that is needed to download the development data. 13 teams submitted 116 graded runs for evaluation (12 teams submitted working notes) attracting similar attention than in 2022 [8]. Each of the groups was allowed a maximum of 10 graded runs per subtask.

Table 1 shows all the teams who participated in the task and their submitted runs. 9 teams participated in the concept detection subtask this year, 6 of those teams also participated in 2022 [8]. 13 teams submitted runs to the caption prediction subtask, 7 of those teams also participated in 2022. Three of the teams participated also in 2021. 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.

3. Data Creation

Figure 1 shows an example from the dataset provided by the task.

Like last year, a dataset that originates from biomedical articles of the PMC Open Access Subset[2] [22] was used and was extended with new images added since the last time the dataset was updated in October 2021. The overall lower number of images is due to the removal of non-CC BY images (including CC BY-SA and CC BY-ND).

Unlike last year, no extensive caption pre-processing beyond the removal of links was performed to keep the captions as realistic as possible. Captions in languages other than English were also removed.

From the resulting captions, concepts were extracted using the Medical Concept Annotation Toolkit (MedCAT) [23]. MedCAT, which is capable of extracting biomedical concepts from

[2] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ [last accessed: 2023-06-17]

Table 1
Participating groups in the ImageCLEFmedical 2023 caption task and their graded runs submitted to both subtasks: T1-Concept Detection and T2-Caption Prediction. Teams with previous participation in 2022 are marked with an asterisk (*).

Team	Institution	Runs T1	Runs T2
AUEB-NLP-Group* [11]	Department of Informatics, Athens University of Economics and Business, Athens, Greece	10	9
Bluefield-2023 [12]	Toyohashi University of Technology, Aichi, Japan and Toyohashi Heart Center, Aichi, Japan	–	2
Clef-CSE-GAN-Team [13]	SSN College Of Engineering, Chennai, India	1	1
closeAI2023 [14]	Baidu Intelligent Health Unit, Beijing, China and Peng Cheng Laboratory, Shenzhen, China	3	8
CS_Morgan* [15]	Computer Science Department, Morgan State University, Baltimore, Maryland	5	10
CSIRO* [16]	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	–	4
IUST_NLPLAB* [17]	School of Computer Engineering, Iran University of Science and Technology, Tehran, Islamic Republic Of Iran	7	10
KDE-Lab_Med* [18]	KDE Laboratory, Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan	10	10
PCLmed [19]	Peng Cheng Laboratory, Shenzhen, China and ADSPLAB, School of Electronic and Computer Engineering, Peking University, Shenzhen, China	–	5
SSN_MLRG [20]	Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, India	5	1
SSNSheerinKavitha*[20]	Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, India	2	3
VCMI* [21]	University of Porto, Porto, Portugal and INESC TEC, Porto, Portugal	10	7

unstructured text, was trained on the MIMIC-III dataset [24] and links to SNOMED CT IDs, which were later mapped to CUIs and TUIs of the UMLS2022AB release[3]. During concept extraction, concepts were retained only if they exceeded a frequency threshold of 10 occurences, and semantic filters were applied to focus on visually observable and interpretable concepts. For example, concepts of semantic type T029 (Body Location or Region) or T060 (Diagnostic Procedure) are relevant, while concepts of semantic type T054 (Social Behavior) cannot be derived from the image if it would appear in the caption. In addition, manual filtering was

[3]https://www.nlm.nih.gov/pubs/techbull/nd22/nd22_umls_2022ab_release_available.html [last accessed: 2023-06-17]

CC BY [Ali et al. (2020)]

UMLS CUI	UMLS Meaning
C1306645	Plain x-ray
C0030797	Pelvis
C1999039	Anterior-Posterior
C0011900	Diagnosis
C1305773	Entire symphysis pubis
C0036036	Sacroiliac joint structure
C0555898	Sacroiliac
C0301559	Screw

Caption: Anteroposterior pelvic radiograph of a 30-year-old female diagnosed with Ehlers-Danlos Syndrome demonstrating fusion of pubic symphysis and both sacroiliac joints (anterior plating, bone grafting and sacroiliac screw insertion)

Figure 1: Example of a radiology image with the corresponding UMLS® CUIs and caption extracted from the 2023's ImageCLEFmedical caption task. CC-BY [Ali et al. (2020)]

performed to exclude UMLS concepts that were either incorrectly detected by the pipeline or were still not related to the image content in any way after semantic filtering. Blacklisted concepts often include qualifiers that would divert actual interest to, for example, anatomical localization or a pathological process, and would also introduce bias, since qualifiers are used in a highly individual and variable manner. Entity linking systems tend to link concepts with ambiguous synonyms incorrectly, e.g. C0994894 (Patch Dosage Form) may be linked if the caption refers to a region that is patchy. In case of high frequency occurrence of such concepts, they were merged to the correct concept via mapping. Due to the different filtering approach, this year's dataset contains 2,125 concepts compared to 8,374 last year.

Additional concepts were assigned to all images addressing their image modality. Six medical image modalities of concepts were covered: X-ray, Computer Tomography (CT), Magnetic Resonance Imaging (MRI), ultrasound, and Positron Emission Tomography (PET) as well as modality combinations (e.g., PET/CT) as standalone concept. For images of the X-ray modality further concepts on the represented anatomy were assigned, covering specific anatomical body regions of the Image Retrieval in Medical Application (IRMA) [25] classification: cranium, spine, upper extremity/arm, chest, breast/mamma, abdomen, pelvis, and lower extremity/leg. New for this year's dataset is the addition of manually validated directionality concepts for x-ray images. Directionality refers to the x-ray imaging orientation according to IRMA: coronal posteroanterior (PA), coronal anteroposterior (AP), sagittal, or transversal. Each of the described concept extensions were created performing a two-stage process. In the first stage, predictions via classification models were created and assigned as annotations. For modality prediction for all images a model trained on the ROCO dataset [9], and for anatomy prediction for X-ray modality images a model trained on an existing IRMA-annotated image dataset [26] was used. For directionality, roughly 20,000 images were manually annotated to train an initial classifier.

In the second stage, these annotations underwent manual quality control measures, involving correction of faulty predictions and filtering of images that did not represent one of the minded modality or anatomy concepts.

The following subsets were distributed to the participants where each image has one caption and one or more concepts (UMLS-CUI):

- *Training set* including 60,918 radiology images and associated captions and concepts, with a total of 263,091 concept occurrences and 2,125 unique concepts.
- *Validation set* including 10,437 radiology images and associated captions and concepts, with a total of 46,584 concept occurrences and 1,945 unique concepts.
- *Test set* including 10,473 radiology images, with a total of 46,955 concept occurrences and 1,936 unique concepts.

4. Evaluation Methodology

In this year's edition, the performance evaluation for the concept detection subtask is carried out in the same way as last year, while the primary evaluation metric for the caption prediction subtask is changed from BLEU to BERTScore. Both tasks are evaluated separately. AIcrowd was not used as a challenge platform this year, instead participants were asked to upload their submissions to a cloud share file drop, with information about whether each submission was successfully evaluated and announced on a website that was regularly updated. An important difference to the last years was the fact that participants were unaware of their own scores on the test set until after the submission deadline. This was done to avoid teams optimizing their approaches based on test set results, which would amount to information leakage.

For the concept detection subtask, the balanced precision and recall trade-off were measured in terms of F1-scores. Like last year, a secondary F1-score is computed using a subset of concepts that was manually curated. On the one hand, this involves the different image modalities (X-ray, Angiography, Ultrasound, CT, MRI, PET, and Combined such as PET/CT). On the other hand, if applicable, for X-ray also the most prominently depicted body region (cranium, chest, upper extremity, spine, abdomen, pelvis, and lower extremity), and the capture directionality (coronal anteroposterior, coronal posteroanterior, sagittal, and transversal) were involved.

As a pre-processing step for evaluating the second task, all captions were lowercased, punctuation was removed, and numbers were replaced by the token "number". This step ensures uniformity and focuses the evaluation on the linguistic content. The performance of caption prediction is evaluated based on BERTScore [27], which is a metric that computes a similarity score for each token in the generated text with each token in the reference text. It uses the pre-trained contextual embeddings from BERT-based models and matches words by cosine similarity. In this work, the pre-trained model *microsoft/deberta-xlarge-mnli*[4] was used because it is the model that correlates best with human scoring according to the authors[5]. Since evaluating generated text and image captioning is very challenging and should not be based on a single metric, additional evaluation metrics were explored in this year's edition in order to find the metrics that correlate

[4]https://huggingface.co/microsoft/deberta-xlarge-mnli [last accessed: 2023-06-17]
[5]https://github.com/Tiiiger/bert_score [last accessed: 2023-06-17]

well with human judgments for this task. First, the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [28] score was adopted as a secondary metric that counts the number of overlapping units such as n-grams, word sequences, and word pairs between the generated text and the reference. Specifically, the ROUGE-1 (F-measure) score was calculated, which measures the number of matching unigrams between the model-generated text and a reference. All individual scores for each caption are then summed and averaged over the number of captions, resulting in the final score. In addition to ROUGE, the Metric for Evaluation of Translation with Explicit ORdering (METEOR) [29] was explored, which is a metric that evaluates the generated text by aligning it to reference and calculating a sentence-level similarity score. Furthermore, the Consensus-based Image Description Evaluation (CIDEr) [30] metric was also adopted. CIDEr is an automatic evaluation metric that calculates the weights of n-grams in the generated text, and the reference text based on term frequency and inverse document frequency (TF-IDF) and then compares them based on cosine similarity. Another metric used is the BiLingual Evaluation Understudy (BLEU) score [31], which is a geometric mean of n-gram scores from 1 to 4. For this task, the focus was on the BLEU-1 score, which takes into account unigram precision. Compared to last year, BLEURT and CLIPScore were newly introduced. BLEURT (Bilingual Evaluation Understudy with Representations from Transformers.) [32] is specifically designed to evaluate natural language generation in English. It uses a pre-trained model that has been fine-tuned to emulate human judgments about the quality of the generated text. The strength of BLEURT lies in its end-to-end training, which enables it to model human judgments effectively and makes it robust to domain and quality variations. For this evaluation, the BLEURT-20 model was used. CLIPScore [33] is an innovative metric that diverges from the traditional reference-based evaluations of image captions. Instead, it aligns with the human approach of evaluating caption quality without references by evaluating the alignment between text and image content. The metric employs CLIP (Contrastive Language-Image Pretraining) [34], a cross-modal model that has been pre-trained on a massive dataset of 400 million image-caption pairs sourced from the web. The model is used to compute similarity scores between images and text. The introduction of BLEURT and CLIPScore in this edition aims to further align the evaluation process with human judgment.

5. Results

For the concept detection and caption prediction subtasks, Tables 2 and 3 show the best results from each of the participating teams. The results will be discussed in this section. The full list of results are shown in appendix A in tables 5, 6 and 7.

5.1. Results for the Concept Detection subtask

In 2023, 9 teams participated in the concept prediction subtask, submitting 47 graded runs. Table 2 presents the results achieved in the submissions.

AUEB-NLP-Group Like in previous years, the AUEB-NLP-Group submitted the best performing result with a primary F1-score of 0.5223 [11] and a secondary F1-score of 0.9258. The winning approach was an ensemble of three CNNs (EfficientNetB0, DenseNet121, and

Table 2
Performance of the participating teams in the ImageCLEFmedical 2023 Concept Detection subtask. Only the best run based on the achieved F1-score is listed for each team, together with the corresponding secondary F1-score based on manual annotations as well as the team rankings based on the primary and secondary F1-score. The full results are shown in table 5 in appendix A.

Group Name	Best Run	F1	Secondary F1	Rank (secondary)
AUEB-NLP-Group	4	**0.5223**	0.9258	1 (2)
KDE-Lab_Med	10	0.5074	**0.9321**	2 (1)
VCMI	8	0.4998	0.9162	3 (3)
IUST_NLPLAB	7	0.4959	0.8804	4 (6)
Clef-CSE-GAN-Team	1	0.4957	0.9106	5 (4)
CS_Morgan	2	0.4834	0.8902	6 (5)
SSNSheerinKavitha	1	0.4649	0.8603	7 (7)
closeAI2023	2	0.0900	0.2152	8 (8)
SSN_MLRG	3	0.0173	0.1122	9 (9)

EfficientNetB0v2) followed by a feed-forward neural network (FFNN) classification head, which is a very similar approach as last year [35], where an ensemble of two such models won the concept detection subtask. They also experimented with training separate models for the different modalities, which did not lead to better results.

KDE-Lab_Med The KDE-Lab_Med team submitted the second best performing approach, with a primary F1-score of 0.5074 [18] and a secondary F1-score of 0.9321, which was the highest overall secondary F1-score. Their best approach is a single CNN+FFNN model with an EfficientNetV2-M backbone. They experimented with image pre-processing by either converting color to grayscale or colorization of grayscale images by stacking color channels. The latter approach performed better.

VCMI The VCMI team achieved the third place in the concept detection subtask with a primary F1-score of 0.4998 [21] and a secondary F1-score of 0.9162. Their best approach utilizes an autoregressive multi-label classification system with a VGG16 network pre-trained on ImageNet, which instead of using a single classification layer at the end, uses 17 classification layers each predicting 125 concepts. For any images that are not assigned any concepts using this model, an image retrieval system assigns concepts appearing in at least two of the four most similar images in the training data.

IUST_NLPLAB The IUST_NLPLAB team reached a primary F1-score of 0.4959 [17] and a secondary F1-score of 0.8804. They used a multi-label classification system based on the vision-language model PubMedCLIP for their best results.

Clef-CSE-GAN-Team The Clef-CSE-GAN-Team achieved a primary F1-score of 0.4957 [13] and a secondary F1-score of 0.9106. They employed a multi-label classification system with a DenseNet121 backbone.

CS_Morgan The CS_Morgan team reached a primary F1-score of 0.4834 [15] and a secondary

F1-score of 0.8902. Their best approach used a multi-label classification system with a DenseNet121 backbone using CheXNet pre-trained weights.

SSN_MLRG and SSNSheerinKavitha The team SSN_MLRG achieved a primary F1-score of 0.4649 [20] and a secondary F1-score of 0.8603. They employed a multi-label classification system using ConceptNet.

To summarize, 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 predict any concepts [21]. As in 2022, the AUEB-NLP-Group once again achieved the top scores by increasing their ensemble from two to three models [11].

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

While one team experimented with a novel autoregressive multi-label classification system which 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.

5.2. Results for the Caption Prediction subtask

In this seventh edition, the caption prediction subtask attracted 13 teams which submitted 69 graded runs. Tables 3 and 4 present the results of the submissions.

Table 3
Performance of the participating teams in the ImageCLEFmedical 2023 Caption Prediction subtask. Only the best run based on the achieved BERTScore is listed for each team, together with the corresponding secondary ROUGE score as well as the team rankings based on the primary BERTScore and secondary ROUGE score. Additional scores are shown in Table 4. The full results are shown in tables 6 and 7 in appendix A.

Group Name	Best Run	BERTScore	ROUGE	Rank (secondary)
CSIRO	2	**0.6413**	0.2463	1 (3)
closeAI2023	7	0.6281	0.2401	2 (4)
AUEB-NLP-Group	2	0.6170	0.2130	3 (8)
PCLmed	5	0.6152	0.2528	4 (2)
VCMI	5	0.6147	0.2175	5 (7)
KDE-Lab_Med	3	0.6145	0.2223	6 (5)
SSN_MLRG	1	0.6019	0.2112	7 (9)
DLNU_CCSE	1	0.6005	0.2029	8 (10)
CS_Morgan	10	0.5819	0.1564	9 (11)
Clef-CSE-GAN-Team	2	0.5816	0.2181	10 (6)
Bluefield-2023	3	0.5780	0.1534	11 (12)
IUST_NLPLAB	6	0.5669	**0.2898**	12 (1)
SSNSheerinKavitha	4	0.5441	0.0866	13 (13)

Table 4
Performance of the participating teams in the ImageCLEFmedical 2023 Caption Prediction subtask for additional metrics BLEURT, BLEU, METEOR, CIDEr and CLIPScore. These correspond to the best BERTScore-based runs of each team, listed in Table 3. The full results are shown in tables 6 and 7 in appendix A.

Group Name	Best Run	BLEURT	BLEU	METEOR	CIDEr	CLIPScore
CSIRO	4	0.3137	0.1615	0.0798	0.2025	**0.8147**
closeAI2023	7	**0.3209**	0.1846	0.0873	**0.2377**	0.8075
AUEB-NLP-Group	2	0.2950	0.1692	0.0720	0.1466	0.8039
PCLmed	5	0.3166	0.2172	0.0921	0.2315	0.8021
VCMI	5	0.3084	0.1653	0.0734	0.1720	0.8082
KDE-Lab_Med	3	0.3014	0.1565	0.0724	0.1819	0.8062
SSN_MLRG	1	0.2774	0.1418	0.0615	0.1284	0.7759
DLNU_CCSE	1	0.2630	0.1059	0.0557	0.1332	0.7725
CS_Morgan	10	0.2242	0.0566	0.0436	0.0840	0.7593
Clef-CSE-GAN-Team	2	0.2690	0.1450	0.0702	0.1737	0.7893
Bluefield-2023	3	0.2716	0.1543	0.0601	0.1009	0.7837
IUST_NLPLAB	6	0.2230	**0.2685**	**0.1004**	0.1773	0.8068
SSNSheerinKavitha	4	0.2152	0.0749	0.0258	0.0143	0.6873

CSIRO The CSIRO team achieved first place in the caption prediction subtask with a BERTScore of 0.6413 [16] and a ROUGE score of 0.2463. The winning approach, which also reached the highest CLIPScore, consists of an encoder-decoder framework based on the Convolutional Vision Transformer (CVT) as the encoder and DistilGPT2 as the decoder. This approach was already used by them in last year's addition and reached the overall highest BERTScore then. For this year, they added a reinforcement learning step "to optimize the model for the primary metric and the means of conditioning the decoder on the visual features" [16], which further improved the performance and set them apart from the competition.

closeAI The closeAI team reached the second place spot with a BERTScore of 0.6281 [14] and a ROUGE score of 0.2401. Their approach, which reached top scores in the BLEURT and CIDEr metrics, consisted of a BLIP-2 framework with a ViT-g image encoder from EVA-CLIP, a Q-Former and OPT2.7 as the LLM with post-processing to remove duplicate content from the generated captions.

AUEB-NLP-Group The AUEB-NLP-Group reached a BERTScore of 0.6170 [11] and a ROUGE score of 0.2130, placing third. Their best approach is a novel captioning pipeline using a denoising model to rewrite captions produced by a CNN-RNN encoder decoder model using sequence to sequence models BART and T5.

PCLmed The PCLmed team achieved a BERTScore of 0.6152 [19] and a ROUGE score of 0.2528. Much like closeAI, they used a BLIP-2 framework with an EVA-ViT-g encoder, a Query Transformer, and ChatGLM-6B as the LLM with a final beam search with a repetition penalty to generate the captions.

VCMI The VCMI team achieved a BERTScore of 0.6147 [21] and a ROUGE score of 0.2175. They used an encoder-decoder framework with a Data-efficient image Transformer (DeiT) as the encoder and DistilGPT-2 as the decoder for their best results.

KDE-Lab_Med The KDE-Lab_Med team reached a BERTScore of 0.6145 [18] and a ROUGE score of 0.2223. Their best approach is a CNN-RNN system based on Show, Attend, and Tell with a ResNet152 backbone and LSTM as RNN. They also experimented with a Caption Transformer system, which did not perform better.

SSN_MLRG and SSNSheerinKavitha The team SSN_MLRG achieved a BERTScore of 0.6019 [20] and a ROUGE score of 0.2112. They used an encoder-decoder system with DeiT as the encoder and Distilled-GPT2 as the decoder.

DLNU_CCSE The DLNU_CCSE team reached a BERTScore of 0.6005 and a ROUGE score of 0.2029. They used an encoder-decoder framework with a ResNet-101 encoder and an LSTM decoder for their best approach. They did not submit working notes.

CS_Morgan The CS_Morgan team reached a BERTScore of 0.5819 [15] and a ROUGE score of 0.1564. They used an encoder-decoder system with a Vision Transformer (ViT) as the encoder, where the decoder generates keywords which are then transformed into captions by a T5 generative model fine-tuned for this purpose.

Clef-CSE-GAN-Team The Clef-CSE-GAN-Team achieved a BERTScore of 0.5816 [13] and a ROUGE score of 0.2181. They used an encoder-decoder approach with a ResNet101 encoder and an LSTM decoder for their best results.

Bluefield The Bluefield team reached a BERTScore of 0.5780 [12] and a ROUGE score of 0.1534. They first classified the images into six groups roughly corresponding to the imaging modalities, and then used six different CLIP models with a ResNet50 backbone to generate captions for the images.

IUST_NLPLAB Last years winners reached a BERTScore of 0.5669 [17] and the overall best ROUGE score of 0.2898. Like last year, they used a multi-label classification approach where the top 20 words words are returned as the caption in the order of their probability. This system once again achieved top scores in the ROUGE, BLEU, and METEOR scores, but did not perform as well on the remaining metrics, including the primary metric BERTScore.

To summarize, in the caption prediction subtask most teams experimented with encoder-decoder frameworks with different backbones and LSTM decoders. Unsurprisingly, teams increasingly used LLMs in the decoding step and to help generate or refine captions. BLIP-2 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.

The aforementioned change of evaluation metrics had a big effect on the outcome of the 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. We will continue to evaluate and explore different possible metrics or combination of metrics, but the evaluation of generated captions remains difficult.

BERTScore and ROUGE scores were used to predict captions. Unlike the previous edition, BERTScore replaced BLEU as the primary score for a more refined evaluation of the caption task. The adoption of BERTScore reflects the intent to prioritize semantic alignment and information preservation in the generated captions rather than focusing on the frequency of n-gram matches, which is the basis of BLEU.

6. Conclusion

This year's caption task of ImageCLEFmedical once again ran with both subtasks, concept detection and caption prediction. It once again used a ROCO-based dataset with additional manual annotations for X-ray directionality. It attracted 13 teams who submitted 116 graded runs using a cloud file drop instead of the AICrowd platform, which was not available to be used this year. For the concept detection task, the F1-score and a secondary F1-score, considering only the manually curated concepts, were used. After adding a number of additional metrics to the caption prediction task last year, the primary metric was changed from BLEU to BERTScore for this year, hoping to reward semantic similarity instead of just n-gram overlap. The caption prediction subtask was more popular than the concept detection subtask this year, with all 9 teams participating in both subtasks, and four teams participating only in the caption prediction subtask. As before, the teams generally approached the tasks completely separately, not really making use of generated concepts for the predicted captions. Like last year, teams generally used multi-label classification systems for the concept detection subtask, last year's winning team simply scaling up their approach to use three instead of two ensembles to once again reach top scores. Retrieval-based systems were still used by some teams, but were consistently outperformed by multi-label classification systems. For the caption prediction subtask, encoder-decoder frameworks were used by most teams, with LLMs being used to generate or refine the captions by some teams. BLIP-2 was used for the first time and achieved good results. Reinforcement learning helped last year's top scoring team in terms of BERTScore further increase last year's score and take the top spot.

The scores for both tasks have improved compared to last year. For the concept detection subtask, this is partly due to the decreased number of concepts. For caption prediction, BERTScore and ROUGE scores have improved, illustrating the beneficial shift to BERTScore as the primary metric, which emphasizes semantic alignment and information preservation over n-gram frequency.

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.

Acknowledgments

This work was partially supported by the University of Essex GCRF QR Engagement Fund provided by Research England (grant number G026). 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] A. García Seco de Herrera, R. Schaer, S. Bromuri, H. Müller, Overview of the ImageCLEF 2016 medical task, in: Working Notes of CLEF 2016 (Cross Language Evaluation Forum), 2016, pp. 219–232.

[2] C. Eickhoff, I. Schwall, A. G. S. de Herrera, H. Müller, Overview of ImageCLEFcaption 2017 - Image Caption Prediction and Concept Detection for Biomedical Images, in: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017., 2017. URL: http://ceur-ws.org/Vol-1866/invited_paper_7.pdf.

[3] A. G. S. de Herrera, C. Eickhoff, V. Andrearczyk, H. Müller, Overview of the ImageCLEF 2018 Caption Prediction Tasks, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018., 2018. URL: http://ceur-ws.org/Vol-2125/invited_paper_4.pdf.

[4] O. Pelka, C. M. Friedrich, A. G. S. de Herrera, H. Müller, Overview of the ImageCLEFmed 2019 Concept Detection Task, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/paper_245.pdf.

[5] O. Pelka, C. M. Friedrich, A. García Seco de Herrera, H. Müller, Overview of the ImageCLEFmed 2020 concept prediction task: Medical image understanding, in: CLEF2020 Working Notes, volume 1166 of *CEUR Workshop Proceedings*, CEUR-WS.org, Thessaloniki, Greece, 2020.

[6] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research 32 (2004) 267–270. doi:10.1093/nar/gkh061.

[7] O. Pelka, A. Ben Abacha, A. García Seco de Herrera, J. Jacutprakart, C. M. Friedrich, H. Müller, Overview of the ImageCLEFmed 2021 concept & caption prediction task, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021, pp. 1101–1112.

[8] J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2022 – Caption Prediction and Concept Detection, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[9] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology Objects in COntext (ROCO): A Multimodal Image Dataset, in: Intravascular Imaging and Computer Assisted

Stenting - and - Large-Scale Annotation of Biomedical Data and Expert Label Synthesis - 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings, 2018, pp. 180–189. doi:10.1007/978-3-030-01364-6_20.

[10] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, 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, 2023.

[11] P. Kaliosis, G. Moschovis, F. Charalambakos, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP group at ImageCLEFmedical caption 2023, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[12] M. Aono, H. Shinoda, T. Asakawa, K. Shimizu, T. Togawa, T. Komoda, Multi-stage medical image captioning using classification and CLIP, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[13] V. Yeshwanth, P. P, L. Kalinathan, Concept detection and image caption generation in medical imaging, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[14] W. Zhou, Z. Ye, Y. Yang, S. Wang, H. Huang, R. Wang, D. Yang, Transferring pre-trained large language-image model for medical image captioning, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[15] M. R. Hasan, O. Layode, M. Rahman, Concept detection and caption prediction in ImageCLEFmedical caption 2023 with convolutional neural networks, vision and text-to-text transfer transformers, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[16] A. Nicolson, J. Dowling, B. Koopman, A concise model for medical image captioning, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[17] Y. Lotfollahi, M. Nobakhtian, M. Hajihosseini, S. Eetemadi, IUST_NLPLAB at ImageCLEFmedical caption tasks 2023, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[18] H. Shinoda, M. Aono, T. Asakawa, K. Shimizu, T. Komoda, T. Togawa, KDE lab at ImageCLEFmedical caption 2023, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[19] B. Yang, A. Raza, Y. Zou, T. Zhang, Customizing general-purpose foundation models for medical report generation, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[20] S. S. N. Mohamed, K. Srinivasan, SSN MLRG at caption 2023: Automatic concept detection and caption prediction using ConceptNet and vision transformer, in: CLEF2023 Working

Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.
[21] I. Rio-Torto, C. Patrício, H. Montenegro, T. Gonçalves, J. S. Cardoso, Detecting concepts and generating captions from medical images: Contributions of the VCMI team to ImageCLEFmedical caption 2023, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.
[22] R. J. Roberts, PubMed Central: The GenBank of the published literature, Proceedings of the National Academy of Sciences of the United States of America 98 (2001) 381–382. doi:10.1073/pnas.98.2.381.
[23] Multi-domain clinical natural language processing with medcat: The medical concept annotation toolkit, Artificial Intelligence in Medicine 117 (2021) 102083. doi:https://doi.org/10.1016/j.artmed.2021.102083.
[24] A. E. Johnson, T. J. Pollard, L. Shen, L. wei H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data 3 (2016). URL: https://doi.org/10.1038/sdata.2016.35. doi:10.1038/sdata.2016.35.
[25] T. M. Lehmann, H. Schubert, D. Keysers, M. Kohnen, B. B. Wein, The IRMA code for unique classification of medical images, in: H. K. Huang, O. M. Ratib (Eds.), Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation, SPIE, 2003. doi:10.1117/12.480677.
[26] T. Deserno, B. Ott, 15.363 IRMA Bilder in 193 Kategorien für ImageCLEFmed 2009, 2009. URL: https://publications.rwth-aachen.de/record/667225. doi:10.18154/RWTH-2016-06143.
[27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.
[28] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.
[29] M. Denkowski, A. Lavie, Meteor Universal: Language Specific Translation Evaluation for Any Target Language, in: Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, 2014, pp. 376–380. URL: http://aclweb.org/anthology/W14-3348. doi:10.3115/v1/W14-3348.
[30] R. Vedantam, C. L. Zitnick, D. Parikh, CIDEr: Consensus-based image description evaluation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 4566–4575. URL: http://ieeexplore.ieee.org/document/7299087/. doi:10.1109/CVPR.2015.7299087.
[31] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[32] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7881–7892. URL: https://aclanthology.org/2020.acl-main.704. doi:10.18653/v1/2020.acl-main.704.

[33] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, Y. Choi, CLIPScore: A reference-free evaluation metric for image captioning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7514–7528. URL: https://aclanthology.org/2021.emnlp-main.595. doi:10.18653/v1/2021.emnlp-main.595.
[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.
[35] F. Charalampakos, G. Zachariadis, J. Pavlopoulos, V. Karatzas, C. Trakas, I. Androutsopoulos, AUEB NLP group at ImageCLEFmed caption 2022, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

A. Full results

Table 5
Performance of the participating teams in the ImageCLEFmedical 2023 Concept Detection subtask.

Group Name	Run	F1	Secondary F1	Rank (secondary)
AUEB-NLP-Group	4	**0.5223**	0.9258	1 (6)
AUEB-NLP-Group	8	0.5221	0.9276	2 (4)
AUEB-NLP-Group	2	0.5219	0.9220	3 (12)
AUEB-NLP-Group	7	0.5213	0.9277	4 (3)
AUEB-NLP-Group	6	0.5208	0.9154	5 (15)
AUEB-NLP-Group	3	0.5208	0.9235	6 (8)
AUEB-NLP-Group	5	0.5189	0.9195	7 (13)
AUEB-NLP-Group	1	0.5174	0.9307	8 (2)
KDE-Lab_Med	10	0.5074	**0.9321**	9 (1)
KDE-Lab_Med	8	0.5016	0.9223	10 (10)
KDE-Lab_Med	3	0.5000	0.9222	11 (11)
VCMI	8	0.4998	0.9162	12 (14)
KDE-Lab_Med	2	0.4992	0.9235	13 (7)
KDE-Lab_Med	1	0.4980	0.9260	14 (5)
KDE-Lab_Med	9	0.4980	0.9224	15 (9)
IUST_NLPLAB	7	0.4959	0.8804	16 (23)
Clef-CSE-GAN-Team	1	0.4957	0.9106	17 (16)
VCMI	5	0.4928	0.9062	18 (17)
IUST_NLPLAB	5	0.4851	0.8928	19 (19)
CS_Morgan	2	0.4834	0.8902	20 (20)
VCMI	7	0.4793	0.9014	21 (18)
CS_Morgan	5	0.4792	0.8582	22 (26)
VCMI	6	0.4728	0.8738	23 (24)
VCMI	3	0.4676	0.8811	24 (22)
SSNSheerinKavitha	1	0.4649	0.8603	25 (25)
SSNSheerinKavitha	2	0.4611	0.8569	26 (27)
VCMI	1	0.4469	0.8305	27 (30)
AUEB-NLP-Group	10	0.4424	0.8113	28 (33)
IUST_NLPLAB	6	0.4409	0.8121	29 (32)
VCMI	10	0.4387	0.8394	30 (29)
CS_Morgan	1	0.4369	0.8543	31 (28)
VCMI	2	0.4360	0.7582	32 (35)
IUST_NLPLAB	3	0.4357	0.8843	33 (21)
IUST_NLPLAB	4	0.4332	0.8238	34 (31)
IUST_NLPLAB	8	0.4212	0.7718	35 (34)
KDE-Lab_Med	4	0.3991	0.7418	36 (36)
KDE-Lab_Med	5	0.3887	0.7252	37 (37)
IUST_NLPLAB	2	0.3804	0.7250	38 (38)
VCMI	9	0.3327	0.7049	39 (39)
VCMI	4	0.2803	0.5999	40 (40)
KDE-Lab_Med	6	0.1061	0.2263	41 (42)
CS_Morgan	4	0.1008	0.3728	42 (41)
KDE-Lab_Med	7	0.0993	0.2136	43 (45)
closeAI2023	2	0.0900	0.2152	44 (43)
closeAI2023	1	0.0900	0.2152	45 (44)
SSN_MLRG	3	0.0173	0.1122	46 (47)
CS_Morgan	3	0.0061	0.1445	47 (46)

Table 6
Performance of the participating teams in the ImageCLEFmedical 2023 Caption Prediction.

Group Name	Run	BERTScore	ROUGE	BLEURT	BLEU	METEOR	CIDEr	CLIPScore
CSIRO	2	**0.6413**	0.2463	0.3151	0.1589	0.0795	0.2071	0.8143
closeAI2023	7	0.6281	0.2401	0.3209	0.1846	0.0873	0.2377	0.8075
closeAI2023	8	0.6243	0.2517	0.3164	0.1743	0.0882	**0.2586**	0.8071
CSIRO	1	0.6225	0.2430	0.3053	0.2055	0.0898	0.2130	**0.8154**
CSIRO	3	0.6189	0.2347	0.3063	0.1922	0.0844	0.1975	0.8130
AUEB-NLP-Group	2	0.6170	0.2130	0.2950	0.1692	0.0720	0.1466	0.8039
PCLmed	5	0.6152	0.2528	0.3166	0.2172	0.0921	0.2315	0.8021
PCLmed	4	0.6148	0.2533	0.3160	0.2176	0.0922	0.2323	0.8020
AUEB-NLP-Group	3	0.6147	0.2144	0.2878	0.1523	0.0696	0.1583	0.8059
VCMI	5	0.6147	0.2175	0.3084	0.1653	0.0734	0.1720	0.8082
KDE-Lab_Med	3	0.6145	0.2223	0.3014	0.1565	0.0724	0.1819	0.8062
KDE-Lab_Med	9	0.6143	0.2319	0.3064	0.1750	0.0773	0.1990	0.8083
PCLmed	3	0.6142	0.2521	0.3154	0.2228	0.0930	0.2280	0.8027
PCLmed	2	0.6142	0.2521	0.3154	0.2228	0.0930	0.2280	0.8027
VCMI	3	0.6138	0.2181	0.3058	0.1618	0.0723	0.1709	0.8089
KDE-Lab_Med	10	0.6108	0.2152	0.2935	0.1577	0.0694	0.1586	0.8042
VCMI	6	0.6103	0.1948	0.2893	0.1233	0.0602	0.1368	0.7996
AUEB-NLP-Group	4	0.6099	0.2189	0.2991	0.1920	0.0742	0.1447	0.7978
KDE-Lab_Med	7	0.6097	0.2204	0.3004	0.1695	0.0725	0.1609	0.8081
VCMI	4	0.6096	0.1938	0.2888	0.1252	0.0592	0.1244	0.7920
KDE-Lab_Med	4	0.6094	0.2005	0.2767	0.1249	0.0596	0.1321	0.7829
KDE-Lab_Med	1	0.6089	0.2160	0.2979	0.1640	0.0699	0.1519	0.8043
KDE-Lab_Med	2	0.6082	0.2144	0.2912	0.1585	0.0687	0.1569	0.8028
closeAI2023	6	0.6080	0.2439	0.3281	0.2267	0.0938	0.2374	0.8069
PCLmed	1	0.6079	0.2422	0.3108	0.2247	0.0894	0.1839	0.8050
AUEB-NLP-Group	1	0.6065	0.2273	0.3049	0.2061	0.0790	0.1662	0.8026
closeAI2023	5	0.6063	0.2449	**0.3306**	0.2217	0.0948	0.2438	0.8070
AUEB-NLP-Group	8	0.6059	0.1885	0.2730	0.1222	0.0606	0.1276	0.8010
KDE-Lab_Med	8	0.6044	0.2167	0.3011	0.1744	0.0730	0.1605	0.8066
closeAI2023	1	0.6039	0.2333	0.2984	0.1580	0.0751	0.1897	0.7943
closeAI2023	2	0.6039	0.2333	0.2984	0.1580	0.0751	0.1897	0.7943
SSN_MLRG	1	0.6019	0.2112	0.2774	0.1418	0.0615	0.1284	0.7759
DLNU_CCSE	1	0.6005	0.2029	0.2630	0.1059	0.0557	0.1332	0.7725
AUEB-NLP-Group	9	0.5960	0.2155	0.3050	0.2040	0.0807	0.1360	0.8043
closeAI2023	3	0.5939	0.2301	0.3284	0.1947	0.0882	0.2215	0.8050
closeAI2023	4	0.5927	0.2364	0.3305	0.1942	0.0899	0.2232	0.8075
AUEB-NLP-Group	6	0.5880	0.1708	0.2590	0.1341	0.0539	0.0816	0.7569
DLNU_CCSE	2	0.5874	0.1886	0.2704	0.1152	0.0559	0.1115	0.7942
CS_Morgan	10	0.5819	0.1564	0.2242	0.0566	0.0436	0.0840	0.7593
Clef-CSE-GAN-Team	2	0.5816	0.2181	0.2690	0.1450	0.0702	0.1737	0.7893
CS_Morgan	4	0.5791	0.1541	0.2649	0.1331	0.0568	0.1731	0.7772
KDE-Lab_Med	5	0.5789	0.1838	0.2905	0.1484	0.0698	0.0838	0.7826
Bluefield-2023	3	0.5780	0.1534	0.2716	0.1543	0.0601	0.1009	0.7837
Bluefield-2023	2	0.5777	0.1539	0.2714	0.1540	0.0597	0.1048	0.7832
VCMI	8	0.5750	0.1464	0.2682	0.1447	0.0555	0.0732	0.7852
VCMI	1	0.5734	0.1427	0.2648	0.1382	0.0533	0.0676	0.7819
IUST_NLPLAB	6	0.5669	**0.2898**	0.2230	**0.2685**	**0.1004**	0.1773	0.8068

Table 7
Performance of the participating teams in the ImageCLEFmedical 2023 Caption Prediction (continued).

Group Name	Run	BERTScore	ROUGE	BLEURT	BLEU	METEOR	CIDEr	CLIPScore
VCMI	2	0.5647	0.1284	0.2554	0.1243	0.0457	0.0491	0.7664
IUST_NLPLAB	2	0.5647	0.2708	0.2088	0.2412	0.0895	0.1594	0.8049
AUEB-NLP-Group	7	0.5630	0.1682	0.2793	0.1514	0.0656	0.0486	0.7602
IUST_NLPLAB	4	0.5612	0.2797	0.2103	0.2592	0.0954	0.1617	0.8061
IUST_NLPLAB	10	0.5560	0.2750	0.2121	0.2643	0.0959	0.1422	0.8007
CS_Morgan	12	0.5558	0.1272	0.2569	0.1199	0.0344	0.0164	0.7338
IUST_NLPLAB	8	0.5534	0.2687	0.2034	0.2639	0.0946	0.1341	0.8030
CS_Morgan	5	0.5508	0.1070	0.2373	0.1040	0.0351	0.0481	0.7165
IUST_NLPLAB	5	0.5494	0.2898	0.2008	0.2685	0.0996	0.1739	0.8042
CS_Morgan	13	0.5482	0.1144	0.2435	0.1180	0.0323	0.0142	0.6909
IUST_NLPLAB	1	0.5463	0.2708	0.1894	0.2412	0.0887	0.1559	0.8028
IUST_NLPLAB	3	0.5445	0.2797	0.1862	0.2592	0.0945	0.1584	0.8026
SSNSheerinKavitha	4	0.5441	0.0866	0.2152	0.0749	0.0258	0.0143	0.6873
CS_Morgan	6	0.5438	0.1107	0.1817	0.0026	0.0329	0.0925	0.7582
SSNSheerinKavitha	3	0.5436	0.0860	0.2151	0.0746	0.0259	0.0143	0.6858
CS_Morgan	9	0.5419	0.0924	0.1735	0.0406	0.0210	0.0187	0.6821
AUEB-NLP-Group	5	0.5417	0.1682	0.2780	0.1323	0.0639	0.0388	0.7600
IUST_NLPLAB	9	0.5394	0.2750	0.1770	0.2643	0.0961	0.1416	0.7969
IUST_NLPLAB	7	0.5367	0.2687	0.1678	0.2639	0.0947	0.1335	0.7967
CS_Morgan	8	0.5087	0.0264	0.1205	0.0034	0.0107	0.0125	0.6819
KDE-Lab_Med	6	0.4425	0.1079	0.2968	0.0709	0.0528	0.0057	0.7305