

Overview of ImageCLEFmedical 2024 – Caption Prediction and Concept Detection

Johannes Rückert^{1,*}, Asma Ben Abacha², Alba G. Seco de Herrera^{3,4}, Louise Bloch^{1,5,6,†}, Raphael Brüngel^{1,5,6,†}, Ahmad Idrissi-Yaghir^{1,5,†}, Henning Schäfer^{7,†}, Benjamin Bracke^{1,5,†}, Hendrik Damm^{1,5,†}, Tabea M. G. Pakull^{7,†}, Cynthia Sabrina Schmidt⁶, Henning Müller^{8,9} and Christoph M. Friedrich^{1,5}

¹Department of Computer Science, University of Applied Sciences and Arts Dortmund, Dortmund, Germany

²Microsoft, Redmond, Washington, USA

³University of Essex, UK

⁴UNED, Spain

⁵Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Germany

⁶Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen, Germany

⁷Institute for Transfusion Medicine, University Hospital Essen, Essen, Germany

⁸University of Applied Sciences Western Switzerland (HES-SO), Switzerland

⁹University of Geneva, Switzerland

Abstract

The ImageCLEFmedical 2024 Caption task on caption prediction and concept detection follows similar challenges held from 2017–2023. 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 50 registered teams, 14 teams submitted 82 graded runs for the two subtasks. Participants mainly used multi-label classification systems for the concept detection subtask, the winning team DBS-HHU utilized an ensemble of four different Convolutional Neural Networks (CNNs). For the caption prediction subtask, most teams used encoder-decoder frameworks with various backbones, including transformer-based decoders and Long Short-Term Memories (LSTMs), with the winning team PCLmed using medical vision-language foundation models (Med-VLFMs) by combining general and specialist vision models.

Keywords

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

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*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); benjamin.bracke@fh-dortmund.de (B. Bracke); hendrik.damm@fh-dortmund.de (H. Damm); tabeamargaretagrace.pakull@uk-essen.de (T. M. G. Pakull); cynthia.schmidt@uk-essen.de (C. S. Schmidt); henning.mueller@hevs.ch (H. Müller); christoph.friedrich@fh-dortmund.de (C. M. Friedrich)

ORCID 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); 0000-0003-4986-7142 (B. Bracke); 0000-0002-7464-4293 (H. Damm); 0009-0009-9802-7167 (T. M. G. Pakull); 0000-0003-1994-0687 (C. S. Schmidt); 0000-0001-6800-9878 (H. Müller); 0000-0001-7906-0038 (C. M. Friedrich)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1. Introduction

ImageCLEF¹ is the image retrieval and classification lab of the Conference and Labs of the Evaluation Forum (CLEF) conference. ImageCLEF 2024 consists of the ImageCLEFmedical, ImageCLEFrecommending, Image Retrieval for Augments (Touché) and ImageCLEFToPicto labs, with the ImageCLEFmedical lab being divided into the subtasks Caption (Image Captioning), VQA (text-to-image generation), MEDIQA-MAGIC (Multimodal And Generative TelemedICine), and GANs (generation of medical images).

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 [10] continued in the same vein, once again using a ROCO-based dataset for both subtasks but switching from BiLingual Evaluation Understudy (BLEU) [11] to BERTScore [12] as the primary evaluation metric for caption prediction. For the 8th edition in 2024, additional metrics as well as an optional explainability extension are introduced for the 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 2024 can be found in Ionescu et al. [13].

This is the 8th edition of the ImageCLEFmedical caption task. Just like in 2016 [1], 2017 [2], 2018 [3], 2021 [7], 2022 [8], and 2023 [10] both subtasks of concept detection and caption prediction are included in ImageCLEFmedical 2024 Caption.

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 2024 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).

In 2024, for the development data, the newly released ROCOV2 [14] dataset, a new iteration of the ROCO [9] dataset, was used, with new images from the PubMed Central[®] (PMC) [15] Open Access subset added for the test set, while images from articles with licenses other than CC BY and CC BY-NC were removed.

This paper presents an overview of the ImageCLEFmedical 2024 Caption task 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 Section 6.

¹<https://www.imageclef.org/> [last accessed: 2024-07-01]

2. Task and Participation

In 2024, 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. This year, an optional new experimental explainability extension has been introduced for the caption prediction task. This extension aims to improve the understanding of the models by asking participants to provide explanations, such as heat maps or Shapley values [16, 17], for a selected number of images. These explanations are manually reviewed to assess their effectiveness and clarity.

In 2024, 50 teams registered and signed the End-User-Agreement that is needed to download the development data. 14 teams submitted 82 graded runs for evaluation (13 teams submitted working notes) attracting a similar number of teams as in 2023 [10], with an overall lower number of graded runs. 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. This year, 9 teams participated in the concept detection subtask, 3 of those teams also participated in 2023 [10]. Of the 11 teams that submitted runs to the caption prediction subtask, 5 also participated in 2023. 3 of the teams participated also in 2022. Overall, 6 teams participated in both subtasks, and 5 teams participated only in the caption prediction subtask. Unlike in 2023, 3 teams participated only in the concept detection subtask.

3. Data Creation

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

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

CC BY [Ali et al. (2020)]



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 2024's ImageCLEFmedical caption task. CC-BY [Ali et al. (2020)]

Like last year, a dataset that originates from biomedical articles of the PMC Open Access Subset² [15] was used and was extended with new images added since the last time the dataset was updated in

²<https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/> [last accessed: 2024-07-01]

Table 1

Participating groups in the ImageCLEFmedical 2024 Caption task and their graded runs submitted to both subtasks: T1-Concept Detection and T2-Caption Prediction. Teams with previous participation in 2023 are marked with an asterisk (*).

Team	Institution	Runs T1	Runs T2
AUEB-NLP-Group* [18]	Department of Informatics, Athens University of Economics and Business, Athens, Greece	10	9
DBS-HHU [19]	Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany	8	2
DS@BioMed [20]	University of Information Technology, Ho Chi Minh City, Vietnam	5	7
SSNMLRGKSR* [21]	Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India	3	–
CS_Morgan* [22]	Computer Science Department, Morgan State University, Baltimore, Maryland	1	9
UACH-VisionLab [23]	Facultad de Ingeniería, Universidad Autónoma de Chihuahua, Chihuahua, Mexico	2	–
MICLab [24]	School of Electrical and Computer Engineering, Universidade Estadual de Campinas, Campinas, Brazil	4	4
Kaprov [25]	Department of CSE, SSN College of Engineering, Chennai, India	1	1
PCLmed* [26]	Peng Cheng Laboratory, Shenzhen, China and ADSPLAB, School of Electronic and Computer Engineering, Peking University, Shenzhen, China	–	3
VIT_Conceptz [27]	Vellore Institute of Technology (VIT), Chennai, India	4	–
KDE-medical-caption* [28]	KDE Laboratory, Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan	–	5
2Q2T [29]	University of Information Technology, Ho Chi Minh City, Vietnam	–	7
DarkCow [30]	Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam	–	3

October 2022. An advantage of using new images for the test set is that contamination of models trained on PMC data is not an issue, since the models in use today were mostly trained prior to 2023. The development dataset for this year consists of the images from the newly released ROCov2 [14] dataset.

Once again, 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 removed.

From the resulting captions, concepts were extracted using the Medical Concept Annotation Toolkit (MedCAT) [31]. MedCAT, which is capable of extracting biomedical concepts from unstructured text, was trained on the Medical Information Mart for Intensive Care (MIMIC)-III dataset [32] and links to Systematized Nomenclature of Medicine and Clinical Terms (SNOMED CT) IDs, which were later mapped to CUIs and Type Unique Identifiers (TUIs) of the UMLS2022AB release³. During concept extraction, concepts were retained only if they exceeded a frequency threshold of 10 occurrences, 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 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,

³https://www.nlm.nih.gov/pubs/techbull/nd22/nd22_umls_2022ab_release_available.html [last accessed: 2024-07-01]

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.

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) [33] classification: cranium, spine, upper extremity/arm, chest, breast/mamma, abdomen, pelvis, and lower extremity/leg. New for last year’s dataset was 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. These concepts were not included in this year’s dataset because the medical expertise and time to both ensure the quality of the directionality concepts for the development dataset as well as validate new directionality concepts on the test set was not available. Table 2 shows statistics about the number of concepts for the datasets of the last three years.

Table 2

Number of unique concepts and average number of concepts per image by split for the ImageCLEFmedical Caption datasets of 2022, 2023, and 2024.

Year	Split	Unique concepts	Concepts per image
2022	train	17,210	4.90
	valid	5126	4.85
	test	4403	4.97
2023	train	2126	3.73
	valid	1946	3.84
	test	1936	3.86
2024	train	1946	3.15
	valid	1752	3.21
	test	700	2.82

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

- *Training set* including 70,108 radiology images and associated captions and concepts, with a total of 220,859 concept occurrences and 1945 unique concepts.
- *Validation set* including 9972 radiology images and associated captions and concepts, with a total of 32,060 concept occurrences and 1751 unique concepts.
- *Test set* including 17,237 radiology images, with a total of 48,563 concept occurrences and 700 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. Both tasks are evaluated separately. The AI4MediaBench⁴ by AIMultimediaLab⁵ was used as the challenge platform. Like last year, participants were unaware of their own scores on the

⁴<https://ai4media-bench.aimultimedialab.ro/> [last accessed: 2024-07-01]

⁵<https://www.aimultimedialab.ro/> [last accessed: 2024-07-01]

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) was 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 [12], 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 Bidirectional Encoder Representations from Transformers (BERT) [34]-based models and matches words by cosine similarity. In this work, the pre-trained model *microsoft/deberta-xlarge-mnli*⁶ was used because it is the model that correlates best with human scoring according to the authors⁷. 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 well with human judgments for this task. First, the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [35] 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) [36] 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) [37] 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 [11], 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. BiLingual Evaluation Understudy with Representations from Transformers (BLEURT) [38] 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 [39] 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 Contrastive Language-Image Pretraining (CLIP) [40], 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. In addition to the reference-free CLIPScore, this evaluation also considers RefCLIPScore [39], an extension that incorporates reference captions. This year, two new domain-specific metrics, MedBERTScore and ClinicalBLEURT [41], have been added to the evaluation. These metrics are tailored for evaluating text in medical contexts and aim to better assess the relevance and accuracy of the generated medical content. MedBERTScore enhances the traditional BERTScore by assigning higher weights to medically relevant terms identified in the text. ClinicalBLEURT is a version of BLEURT fine-tuned on large collections of family medicine and orthopedic notes to better capture

⁶<https://huggingface.co/microsoft/deberta-xlarge-mnli> [last accessed: 2023-07-01]

⁷https://github.com/Tiiiger/bert_score [last accessed: 2023-07-01]

the characteristics of the medical language.

5. Results

For the concept detection and caption prediction subtasks, Tables 3 and 4 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 7, 8 and 9.

5.1. Results for the Concept Detection Subtask

In 2024, 9 teams participated in the concept prediction subtask, submitting 38 graded runs. Table 3 presents the best results for each team achieved in the submissions.

Table 3

Performance of the participating teams in the ImageCLEFmedical 2024 Caption 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 7 in Appendix A.

Group Name	Best Run	F1	Secondary F1	Rank (secondary)
DBS-HHU	601	0.6375	0.9534	1 (1)
auebmlpgroup	644	0.6319	0.9393	2 (2)
DS@BioMed	653	0.6200	0.9312	3 (4)
SSNMLRGKSR	425	0.6001	0.9056	4 (5)
UACH-VisionLab	235	0.5988	0.9363	5 (3)
MICLabNM	681	0.5795	0.8835	6 (6)
Kaprov	558	0.4609	0.7301	7 (7)
VIT_ConceptZ	233	0.1812	0.2647	8 (8)
CS_Morgan	530	0.1076	0.2105	9 (9)

DBS-HHU [19] Dethroning the winners of the last several years, the DBS-HHU team achieved the best F1-scores of 0.6375 (primary) and 0.9534 (secondary) by using an ensemble of four different Convolutional Neural Networks (CNNs): ResNet-152 [42], EfficientNet-B0 [43], DenseNet-201 [44], and Wide ResNet-101-2 [45], all pre-trained on ImageNet [46] and followed by different Feed-Forward Neural Networks (FFNNs). Additionally, they experimented with building a hierarchical system of several models, specifically oriented towards the AUEB-NLP-Group’s approach of prior years. However, these did not beat the best results of their first strategy.

AUEB-NLP-Group [18] The AUEB-NLP-Group based their approach on their past work, which won the competition in the last several years, by combining a CNN (DenseNet [44]) followed by a FFNN classification head which achieved a close second place with a primary F1-score of 0.6319 and a secondary F1-score of 0.9393. They also experimented with CNNs followed by k -Nearest Neighbor (k -NN) models and ensembles which performed slightly worse.

DS@BioMed [20] The DS@BioMed team employed a Shifted Window Transformer v2 (Swin-v2) [47] to achieve an F1-score of 0.6200 and a secondary F1-score of 0.9312. They also experimented with other transformer-based architectures, as well as CNNs and ensembles.

SSNMLRGKSR [21] The SSNMLRGKSR team used a DenseNet-121 [44] CNN for their best approach which achieved a primary F1-score of 0.6001 and a secondary F1-score of 0.9056.

UACH-VisionLab [23] The UACH-VisionLab team used several EfficientNet-B0 [43] models trained for different sub-groups of concepts to achieve a primary F1-score of 0.5988 and a secondary F1-score of 0.9363.

MICLabNM [24] The MICLabNM team employed a VisualT5 image-to-text encoder-decoder architecture coupling a Vision Transformer (ViT) [48] with an encoder-decoder T5 [49] text transformer achieving F1-scores of 0.5795 and 0.8835.

Kaprov [25] The Kaprov team utilized a CNN-LSTM model, achieving a primary F1-score of 0.4609 and a secondary F1-score of 0.7301

VIT_Conceptz [27] The VIT_Conceptz team used a ResNet50 [42] CNN to achieve F1-scores of 0.1812 and 0.2647.

CS_Morgan [22] The CS_Morgan team experimented with a ConvMixer [50] model which consists of a combination of CNN and Transformer architectures achieving F1-scores of 0.1076 and 0.2105.

To summarize, in the concept detection subtasks, the groups used primarily multi-label classification systems, with one team integrating image retrieval systems in some of their approaches. Most teams used CNNs to extract features for images. Some teams explored Transformer-based [51] models, such as ViTs [48], while one team used a ConvMixer [50] architecture, blending convolutional networks and ViTs. The winning team this year utilized an ensemble of four different CNNs.

Comparing this year's concept detection task results to those of the last year's ImageCLEFmedical Caption, a remarkable increase of achieved F1-Scores can be observed. For a direct comparison, last year's winner and now second best AUEB-NLP-Group managed to increase their F1-Score from 0.5223 to 0.6319, close to team DBS-HHU's winning F1-Score of 0.6375. This increase is much smaller for the secondary F1-Score, where the AUEB-NLP-Group increased their score from 0.9258 to 0.9393, and DBS-HHU achieved a new all-time high of 0.9534. By training and evaluating our own baseline model on the data from this year, we could determine that about 0.1 of the difference in primary F1-score is purely due to the new test dataset, which contains a much smaller number of unique concepts (see Table 2). One difference in this year's dataset compared to last year's is that the newly added images were fully used for the test dataset and not split into validation and test, resulting in a larger test dataset. On the other hand, the number of unique concepts in the test dataset is much lower than last year, indicating a difference in the newly added data. The practice of updating the test set with the latest images from the PMC Open Access subset can lead to such complications. Further improvements in primary and secondary F1-score can be attributed to continuous changes and improvements of the challenge dataset, e.g., correction of previous errors and further refinement of quality assurance measures as well as improvements and scaling of the teams' approaches.

5.2. Results for the Caption Prediction Subtask

In this 8th edition, the caption prediction subtask attracted 11 teams which submitted 53 graded runs. Tables 4, 5 and 6 present the results of the submissions.

PCLmed [26] The winning team introduced Medical Vision-Language Foundation Models (Med-VLFM) with Vision Encoder Ensembling (VEE) for better representing the content of medical images and Modality-Aware Adaptation (MAA) to take the inference between vision and text modalities into account. An ensemble of a Explore the limits of Visual representation at scAle (EVA)-ViT-g [52] model which was pre-trained on natural images and a BioMedCLIP [53] model pre-trained on medical images was implemented for image encoding. Pangu- α [54] has been used as the Large Language Model (LLM) for text generation. The model reached a BERTScore of 0.6299 and a ROUGE score of 0.2726 and won the caption prediction task.

CS_Morgan [22] The CS_Morgan team experimented with different Large Multimodal Models (LMMs) like Large Language and Vision Assistant (LLaVA) [55], IDEFICS [56], and MoonDream2⁸. The results of these models are compared to conventional encoder-decoder models like VisionGPT2

⁸<https://huggingface.co/vikhyatk/moondream2> [last accessed: 2024-07-01]

Table 4

Performance of the participating teams in the ImageCLEFmedical 2024 Caption 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 Tables 5 and 6. The full results are shown in Tables 8 and 9 in Appendix A.

Group Name	Best Run	BERTScore	ROUGE	Rank (secondary)
pclmed	634	0.6299	0.2726	1 (1)
CS_Morgan	429	0.6281	0.2508	2 (2)
DarkCow	220	0.6267	0.2452	3 (4)
auebnpgroup	630	0.6211	0.2049	4 (7)
2Q2T	643	0.6178	0.2478	5 (3)
MICLab	678	0.6128	0.2135	6 (6)
DLNU_CCSE	674	0.6066	0.2179	7 (5)
Kaprov	559	0.5964	0.1905	8 (8)
DS@BioMed	571	0.5794	0.1031	9 (11)
DBS-HHU	637	0.5769	0.1531	10 (9)
KDE-medical-caption	557	0.5673	0.1325	11 (10)

Table 5

Performance of the participating teams in the ImageCLEFmedical 2024 Caption caption Prediction subtask for additional metrics BLEU-1, BLEURT, ClinicalBLEURT and METEOR. These correspond to the best BERTScore-based runs of each team, listed in Table 4. The full results are shown in Tables 8 and 9 in Appendix A.

Group Name	Best Run	BLEU-1	BLEURT	ClinicalBLEURT	METEOR
pclmed	634	0.2690	0.3376	0.4666	0.1133
CS_Morgan	429	0.2093	0.3174	0.4559	0.0927
DarkCow	220	0.1950	0.3060	0.4562	0.0889
auebnpgroup	630	0.1110	0.2899	0.4866	0.0680
2Q2T	643	0.2213	0.3139	0.4759	0.0986
MICLab	678	0.1853	0.3067	0.4453	0.0772
DLNU_CCSE	674	0.1512	0.2831	0.4756	0.0704
Kaprov	559	0.1697	0.2951	0.4400	0.0609
DS@BioMed	571	0.0121	0.2202	0.5295	0.0353
DBS-HHU	637	0.1493	0.2710	0.4766	0.0559
KDE-medical-caption	557	0.1060	0.2566	0.5022	0.0386

and CNN-Transformer architectures. The best-performing model of the team was a fine-tuned LLaVA 1.6 Mistral 7B. This model achieved a BERTScore of 0.6281 and a ROUGE score of 0.2508.

DarkCow [30] The DarkCow team obtained a BERTScore of 0.6267 and a ROUGE score of 0.2452. A VinVL [57] model was used to extract object features from the images. These features were combined with more general visual features extracted using a ViT [48] model. ClinicalT5- [58] and Biomedical Bidirectional and Auto-Regressive Transformers (BioBART) [59]-based models were used for the caption generation. The best results were achieved for the BioBART model.

AUEB-NLP-Group [18] The AUEB-NLP-Group’s approach on caption prediction involved four primary systems: The first one employing a InstructBLIP [60] model, and the other ones building up upon it, applying a synthesizer, a rephraser, and an innovative Distance from Median Maximum Concept Similarity (DMMCS) mechanism. One combination of InstructBLIP with DMMCS achieved the team’s best BERTscore of 0.6211 and a ROUGE score of 0.2049.

2Q2T [29] The 2Q2T team used the Bootstrapping Language-Image Pre-training (BLIP) [61] architecture as their main approach, which combines a ViT [48] as the encoder while using BERT [34]

Table 6

Performance of the participating teams in the ImageCLEFmedical 2024 Caption caption Prediction subtask for additional metrics CIDEr, CLIPScore, RefCLIPScore and MedBERTScore. These correspond to the best BERTScore-based runs of each team, listed in Table 4. The full results are shown in Tables 8 and 9 in Appendix A.

Group Name	Best Run	CIDEr	CLIPScore	RefCLIPScore	MedBERTScore
pclmed	634	0.2681	0.8236	0.8176	0.6323
CS_Morgan	429	0.2450	0.8213	0.8155	0.6327
DarkCow	220	0.2243	0.8184	0.8117	0.6292
auebnpgroup	630	0.1769	0.8041	0.7987	0.6261
2Q2T	643	0.2200	0.8271	0.8138	0.6224
MICLab	678	0.1582	0.8159	0.8049	0.6172
DLNU_CCSE	674	0.1688	0.7967	0.7904	0.6130
Kaprov	559	0.1070	0.7922	0.7872	0.6089
DS@BioMed	571	0.0715	0.7756	0.7748	0.5804
DBS-HHU	637	0.0644	0.7842	0.7750	0.5827
KDE-medical-caption	557	0.0384	0.7651	0.7610	0.5697

for text generation. They yielded a BERTScore of 0.6178 and ROUGE score of 0.2478 for caption prediction.

MICLabNM [24] The MICLabNM team used a model that combines a ViT [48] with ClinicalT5 [58], called VisualT5. The approach also features a modified spatial attention module for interpretability, by highlighting important image areas for model decisions. The approach achieved a 0.6129 BERTScore and a ROUGE score of 0.2135 for caption prediction.

DLNU_CCSE The team’s approach achieved a BERTScore of 0.6066 and a ROUGE score of 0.2179, with no working notes submitted by the team.

Kaprov [25] The Kaprov team implemented a combination of a Visual Geometry Group (VGG)-16 [62]-based CNN and a Long Short-Term Memory (LSTM) [63] model for the caption prediction task. The team achieved a BERTScore of 0.5964 and a ROUGE score of 0.1905 on the private test set.

DS@BioMed [20] The best performing-model which was submitted by the DS@BioMed team implemented a combination of a BERT [34] Pre-Training of Image Transformers (BEiT) [64] and an BioBART [59] model. This model incorporated the information which was extracted from the medical images with the concepts extracted in the concept detection task. The team achieved a BERTScore of 0.5794 and a ROUGE score of 0.1031 on the private test set.

DBS-HHU [19] The DBS-HHU team based their caption prediction approach on simple pre-processing (lowercasing, punctuation removal, numbers exchange with number token) to focus on linguistic content. Two models, fine-tuned Generative Image-to-text Transformer (GIT) [65] -base and GIT-large, were then employed for caption generation. Both models achieved nearly equal scores, with the large model achieving the higher BERTscore of 0.5769 and a ROUGE score of 0.1531.

KDE-MED-CAPTION [28] The KDE-MED-CAPTION team implemented a caption retrieval approach. First, a priority-based partitioning was implemented. Afterwards, EfficientNet [43], ResNeXt [66], and ViT [48] models were trained for concept detection. These models were used for feature extraction. Similarity measures were used to compare the extracted features from the test samples with the training samples. The caption of the most similar training sample is predicted for a test sample. The best model submitted by the KDE-MED-CAPTION team reached an BERTScore of 0.5673 and a ROUGE score of 0.1325.

To summarize, in the caption prediction subtask teams primarily utilized encoder-decoder frameworks with various backbones, including transformer-based decoders and LSTMs [63]. ViTs [48] were

commonly employed for feature extraction. Some approaches integrated concept detection into the caption generation process by providing predicted concepts as input to the encoder along with the images. This year saw a notable increase in the use of LLMs such as BioBART [59] and ClinicalT5 [58] and Vision Language Models (VLMs), including LLaVA [55] and IDEFICS [56], with some teams experimenting with visual instruction tuning. Only one team used a retrieval-based approach for this approach. The winning team introduced medical vision-language foundation models (Med-VLFMs) by combining general and specialist vision models to achieve top rankings in the challenge.

This is the second iteration of the caption prediction subtask which used BERTScore and ROUGE as primary and secondary evaluation metrics, after BLEU-1 had been used as the primary evaluation metric in all previous iterations. While some teams were still mainly optimizing for the BLEU-1 score last year, resulting in a wide spread of scores for the different metrics with some teams scoring very strongly in some metrics and very weakly in others, the scores were much more even this year, with the winning approach scoring strongly across all metrics.

Even though last year's winning team CSIRO achieved an all-time high BERTScore of 0.6425, a notable overall increase is visible in returning teams' scores. E.g., this year's winning team PCLmed increased their prior score from 0.6152 to 0.6299. The same applies for other teams CS_Morgan (0.5819 vs. 0.6281), the AUEB-NLP-Group (0.6170 vs. 0.6211), and team DLNU_CCSE (0.6005 vs. 0.6066). Such notable increases are observable for the other scores ROUGE, BLEURT, CIDEr, METEOR, and CLIPScore as well. The main reasons for the improvements are likely continuous improvements of the teams' approaches, while experimentation with new approaches did not yield breakthrough improvements. The newly introduced metrics ClinicalBLEURT and MedBERTScore grant additional insight.

The new optional explainability extension was not adopted by the teams, only the team MICLabNM [24] submitted explainability results after the end of the submission phase.

6. Conclusion

This year's caption task of ImageCLEFmedical once again ran with both subtasks, concept detection and caption prediction. It used the newly released ROCov2 [14] as the development dataset. It attracted 14 teams who submitted 82 graded runs using for the first time the AI4MediaBench platform. For the concept detection task, the F1-score and a secondary F1-score, considering only the manually curated concepts, were used. After changing the primary evaluation metric for the caption prediction subtask from BLEU to BERTScore for last year, additional, more domain-specific metrics were added for this year, one of which may be used as the primary metric for next year. The caption prediction subtask was again more popular than the concept detection subtask this year, with 6 teams participating in both subtasks, 5 teams participating only in the caption prediction subtask, and 3 teams only participating in the concept detection subtask. As before, the teams generally approached the tasks completely separately, with only the DS@BioMed team using the generated concepts for the predicted captions.

Like in the 2023 challenge [10], teams generally used multi-label classification systems for the concept detection subtask, with the winning team using an ensemble of four CNNs. Only one team integrated image retrieval systems in some of their approaches. For the caption prediction subtask, encoder-decoder frameworks were used by most teams, with ViTs being used to extract features. LLMs were increasingly being used to generate and fine-tune the captions. The winning approach used Med-VLFMs by combining general and specialist vision models.

For the concept detection subtask, the overall primary F1-scores increased strongly compared to last year despite very similar approaches being employed by the teams. In addition to continuously improved and scaled-up approaches by the teams, a large part of the improvement can be explained by a lower number of unique concepts in the test set compared to last year.

The same applies for the general view on results of this year's caption prediction task. The top scores were slightly worse for BERTScore, but last year's winners CSIRO [67] did not participate this year. Returning teams improved their scores across the board showing that the dataset for this year is comparable to last year for the caption prediction and that while teams have experimented with many

different approaches including LLMs for caption generation, no breakthrough improvement has been achieved with these new techniques.

For next year's ImageCLEFmedical Caption challenge, some possible improvements include an improved caption prediction evaluation metric which is specific to medical texts, as well as additional metrics for readability and factuality. A comprehensive analysis of different metrics is planned to determine whether they should be used as primary indicators or whether a combination of different metrics would be more appropriate for this task, given the complex nature of evaluating generated captions.

An additional focus will be explainability. The optional extension to the caption prediction subtask where participants were asked to provide explainability results for a small subset of images was not adopted by the participants, with only a single team submitting explainability results after the end of the submission phase. For next year, examples will be provided for how these explainability results could look and it might be extracted into its own subtask.

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, Benjamin Bracke and Raphael Brüngel was partially funded by a PhD grant from the University of Applied Sciences and Arts Dortmund (FH Dortmund), Germany. The work of Ahmad Idrissi-Yaghir, Henning Schäfer, Tabea M. G. Pakull and Hendrik Damm 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] J. Rückert, A. Ben Abacha, A. G. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2023 – caption prediction and concept detection, in: CLEF2023 Working Notes, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, Thessaloniki, Greece, 2023, pp. 1328 – 1346.
- [11] 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.
- [12] 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>.
- [13] B. Ionescu, H. Müller, A. Drăgulescu, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.
- [14] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. S. de Herrera, H. Müller, P. A. Horn, F. Nensa, C. M. Friedrich, ROCov2: Radiology Objects in COntext version 2, an updated multimodal image dataset, Scientific Data (2024). URL: <https://arxiv.org/abs/2405.10004v1>. doi:10.1038/s41597-024-03496-6.
- [15] 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.
- [16] L. S. Shapley, et al., A value for n-person games (1953).
- [17] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Neural Information Processing Systems, volume 30, 2017, pp. 4768 – 4777.
- [18] M. Samprovalaki, A. Chatzipapadopoulou, G. Moschovis, F. Charalampakos, P. Kaliosis, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP group at ImageCLEFmedical 2024, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.
- [19] H. Kauschke, K. Bogomasov, S. Conrad, Predicting captions and detecting concepts for medical images: Contributions of the DBS-HHU team to ImageCLEFmedical caption 2024, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.
- [20] N. N. Nguyen, H. L. Tu, P. D. Nguyen, T. N. Do, T. M. Thai, T. B. Nguyen-Tat, DS@BioMed at ImageCLEFmedical caption 2024: Enhanced attention mechanisms in medical caption generation through concept detection integration, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.
- [21] R. Dhinakaran, S. S. N. Mohamed, K. Srinivasan, SSNMLRGKSR at ImageCLEFmedical caption 2024: Medical concept detection using DenseNet-121 with MultiLabelBinarizer, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.
- [22] M. Hoque, M. R. Hasan, M. I. S. Emon, F. Khalifa, M. M. Rahman, Medical image interpretation with large multimodal models, in: CLEF2024 Working Notes, CEUR Workshop Proceedings,

CEUR-WS.org, Grenoble, France, 2024.

- [23] A. Moncloa-Muro, G. Ramirez-Alonso, F. Martinez-Reyes, Automatic medical concept detection on images: dividing the task into smaller ones, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.
- [24] D. Carmo, L. Rittner, R. Lotufo, VisualT5: Multitasking caption and concept prediction with pre-trained ViT, T5 and customized spatial attention in radiological images, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.
- [25] P. Balasundaram, K. Swaminathan, O. Sampath, P. KM, Concept detection and caption prediction of radiology images using convolutional neural networks, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.
- [26] B. Yang, Y. Yu, Y. Zou, T. Zhang, PCLmed: Champion solution for ImageCLEFmedical 2024 caption prediction challenge via medical vision-language foundation models, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.
- [27] S. Ram, S. Vinoth, R. N. Gopalakrishnan, A. A. Balakumar, L. Kalinathan, T. A. J. Velankanni, Leveraging diverse CNN architectures for medical image captioning: DenseNet-121, MobileNetV2, and ResNet-50 in ImageCLEF 2024, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.
- [28] M. Aono, T. Asakawa, K. Shimizu, K. Nomura, Medical image captioning using CUI-based classification and feature similarity, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.
- [29] T. V. Phan, T. K. Nguyen, Q. A. Hoang, Q. T. Phan, T. B. Nguyen-Tat, MedBLIP: Multimodal medical image captioning using BLIP, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.
- [30] Q. V. Nguyen, Q. H. Pham, D. Q. Tran, T. K.-B. Nguyen, N.-H. Nguyen-Dang, B.-T. Nguyen-Tat, UIT-DarkCow team at ImageCLEFmedical caption 2024: Diagnostic captioning for radiology images efficiency with transformer models, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.
- [31] Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts, R. Bendayan, M. P. Richardson, R. Stewart, A. D. Shah, W. K. Wong, Z. Ibrahim, J. T. Teo, R. J. Dobson, Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit, *Artificial Intelligence in Medicine* 117 (2021) 102083. URL: <https://www.sciencedirect.com/science/article/pii/S0933365721000762>. doi:<https://doi.org/10.1016/j.artmed.2021.102083>.
- [32] 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.
- [33] 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.
- [34] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171 – 4186. URL: <https://aclanthology.org/N19-1423>. doi:10.18653/v1/N19-1423.
- [35] 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>.
- [36] 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.

- [37] 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.
- [38] 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.
- [39] 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.
- [40] 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>.
- [41] A. Ben Abacha, W.-w. Yim, G. Michalopoulos, T. Lin, An investigation of evaluation methods in automatic medical note generation, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2575–2588. URL: <https://aclanthology.org/2023.findings-acl.161>. doi:10.18653/v1/2023.findings-acl.161.
- [42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), 2016, pp. 770 – 778. doi:10.1109/CVPR.2016.90.
- [43] M. Tan, Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: Proceedings of the International Conference on Machine Learning (ICML 2019), 2019, pp. 6105 – 6114.
- [44] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), 2017, pp. 2261 – 2269. doi:10.1109/CVPR.2017.243.
- [45] S. Zagoruyko, N. Komodakis, Wide residual networks, in: Proceedings of the British Machine Vision Conference (BMVC 2016), 2016. doi:10.5244/c.30.87.
- [46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), 2009, pp. 248 – 255. doi:10.1109/CVPR.2009.5206848.
- [47] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin Transformer V2: Scaling up capacity and resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), 2022, pp. 11999 – 12009. doi:10.1109/CVPR52688.2022.01170.
- [48] A. Dosovitskiy, L. Beyer, A. I. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: Proceedings of the International Conference on Learning Representations (ICLR 2021), 2021.
- [49] C. Raffel, N. Shazeer, A. Roberts, K. J. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *Journal of Machine Learning Research* 21 (2020) 1 – 67.
- [50] A. Trockman, J. Z. Kolter, Patches are all you need?, *Transactions on Machine Learning Research* (2023). URL: <https://openreview.net/forum?id=rAnB7JSMXL>.
- [51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin,

//openreview.net/forum?id=p-BhZSz59o4.

- [65] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, L. Wang, GIT: A generative image-to-text transformer for vision and language, Transactions on Machine Learning Research 2022 (2022).
- [66] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), 2017, pp. 5987–5995. doi:10.1109/CVPR.2017.634.
- [67] 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.

A. Full Results

Table 7

Performance of the participating teams in the ImageCLEFmedical 2024 Concept Detection subtask.

Group Name	Run	F1	Secondary F1	Rank (secondary)
DBS-HHU	601	0.6375	0.9534	1 (1)
DBS-HHU	602	0.6375	0.9534	2 (2)
DBS-HHU	603	0.6375	0.9534	3 (3)
auebnlpgroup	644	0.6319	0.9393	4 (8)
DBS-HHU	625	0.6309	0.9488	5 (4)
auebnlpgroup	648	0.6308	0.9321	6 (13)
auebnlpgroup	642	0.6304	0.9333	7 (12)
auebnlpgroup	624	0.6274	0.9376	8 (9)
auebnlpgroup	640	0.6273	0.9416	9 (7)
DBS-HHU	600	0.6269	0.9461	10 (5)
DBS-HHU	604	0.6269	0.9461	11 (6)
auebnlpgroup	619	0.6241	0.9339	12 (11)
auebnlpgroup	654	0.6207	0.9243	13 (15)
DS@BioMed	653	0.6200	0.9312	14 (14)
auebnlpgroup	656	0.6162	0.9218	15 (18)
auebnlpgroup	655	0.6156	0.9234	16 (17)
auebnlpgroup	651	0.6136	0.9239	17 (16)
DS@BioMed	652	0.6108	0.9193	18 (19)
DS@BioMed	365	0.6090	0.9177	19 (20)
DS@BioMed	364	0.6090	0.9177	20 (21)
SSNMLRGKSR	425	0.6001	0.9056	21 (22)
SSNMLRGKSR	422	0.6001	0.9056	22 (23)
UACH-VisionLab	235	0.5988	0.9363	23 (10)
MICLabNM	681	0.5795	0.8835	24 (24)
MICLabNM	680	0.5594	0.8568	25 (25)
SSNMLRGKSR	421	0.5463	0.7969	26 (29)
MICLabNM	275	0.5343	0.8133	27 (28)
UACH-VisionLab	290	0.5292	0.8422	28 (26)
MICLabNM	679	0.5282	0.8325	29 (27)
Kaprov	558	0.4609	0.7301	30 (30)
DBS-HHU	610	0.3417	0.4477	31 (31)
DBS-HHU	616	0.3413	0.4340	32 (32)
VIT_ConceptZ	233	0.1812	0.2647	33 (33)
VIT_ConceptZ	471	0.1812	0.2647	34 (34)
VIT_ConceptZ	487	0.1785	0.2536	35 (35)
VIT_ConceptZ	488	0.1143	0.2308	36 (36)
CS_Morgan	530	0.1076	0.2105	37 (37)
DS@BioMed	242	0.0019	0.0032	38 (38)

Table 8

Performance of the participating teams in the ImageCLEFmedical 2024 Caption Prediction for the metrics BERTScore, ROUGE, BLEU-1, BLEURT, ClinicalBLEURT, and METEOR.

Group Name	Run	BERTScore	ROUGE	BLEU-1	BLEURT	ClinicalBLEURT	METEOR
pclmed	634	0.6299	0.2726	0.2690	0.3376	0.4666	0.1133
CS_Morgan	429	0.6281	0.2508	0.2093	0.3174	0.4559	0.0927
DarkCow	220	0.6267	0.2452	0.1950	0.3060	0.4562	0.0889
CS_Morgan	527	0.6254	0.2454	0.2076	0.3165	0.4435	0.0892
CS_Morgan	526	0.6250	0.2440	0.2049	0.3153	0.4438	0.0898
pclmed	633	0.6235	0.2717	0.2680	0.3386	0.4671	0.1121
CS_Morgan	525	0.6230	0.2380	0.1951	0.3096	0.4358	0.0854
pclmed	632	0.6227	0.2690	0.2650	0.3365	0.4654	0.1110
auebnlpgroup	630	0.6211	0.2049	0.1110	0.2899	0.4866	0.0680
auebnlpgroup	635	0.6210	0.2047	0.1108	0.2895	0.4870	0.0680
auebnlpgroup	646	0.6210	0.2044	0.1107	0.2900	0.4872	0.0678
auebnlpgroup	647	0.6210	0.1807	0.0860	0.2846	0.5021	0.0580
DarkCow	243	0.6200	0.2139	0.1685	0.2913	0.4597	0.0751
2Q2T	643	0.6178	0.2478	0.2213	0.3139	0.4759	0.0986
2Q2T	682	0.6178	0.2478	0.2213	0.3139	0.4759	0.0986
CS_Morgan	613	0.6173	0.2178	0.1559	0.2976	0.4487	0.0730
CS_Morgan	529	0.6166	0.2160	0.1827	0.3058	0.4534	0.0760
2Q2T	683	0.6165	0.2501	0.2353	0.3153	0.4748	0.1018
auebnlpgroup	650	0.6159	0.1936	0.1050	0.2859	0.4874	0.0638
CS_Morgan	528	0.6157	0.2237	0.1741	0.3005	0.4339	0.0770
auebnlpgroup	564	0.6153	0.2052	0.1274	0.2920	0.4844	0.0698
MICLab	678	0.6128	0.2135	0.1853	0.3067	0.4453	0.0772
auebnlpgroup	605	0.6114	0.1889	0.1147	0.2796	0.4834	0.0616
auebnlpgroup	639	0.6111	0.1827	0.0744	0.2717	0.5212	0.0515
auebnlpgroup	577	0.6107	0.1838	0.0751	0.2706	0.5158	0.0513
2Q2T	512	0.6106	0.2353	0.2069	0.3209	0.4459	0.0884
2Q2T	684	0.6092	0.2342	0.2148	0.3243	0.4467	0.0893
2Q2T	592	0.6091	0.2341	0.2148	0.3243	0.4468	0.0892
2Q2T	595	0.6091	0.2341	0.2148	0.3243	0.4468	0.0892
MICLab	676	0.6072	0.1922	0.1480	0.2905	0.4608	0.0642
DLNU_CCSE	674	0.6066	0.2179	0.1512	0.2831	0.4756	0.0704
DarkCow	221	0.5994	0.2363	0.2323	0.2954	0.4597	0.0989
Kaprov	559	0.5964	0.1905	0.1697	0.2951	0.4400	0.0609
MICLab	274	0.5888	0.1933	0.1626	0.2864	0.4443	0.0617
DLNU_CCSE	675	0.5839	0.1844	0.1579	0.2756	0.4524	0.0594
DS@BioMed	571	0.5794	0.1031	0.0121	0.2202	0.5295	0.0353
DS@BioMed	563	0.5794	0.1031	0.0121	0.2202	0.5295	0.0353
DBS-HHU	637	0.5769	0.1531	0.1493	0.2710	0.4766	0.0559
DBS-HHU	645	0.5769	0.1531	0.1493	0.2710	0.4766	0.0559
KDE-medical-caption	557	0.5673	0.1325	0.1060	0.2566	0.5022	0.0386
KDE-medical-caption	544	0.5665	0.1273	0.1151	0.2513	0.5220	0.0438
KDE-medical-caption	424	0.5646	0.1223	0.1030	0.2439	0.5082	0.0413
KDE-medical-caption	423	0.5646	0.1223	0.1030	0.2439	0.5082	0.0413
KDE-medical-caption	460	0.5630	0.1199	0.1035	0.2410	0.5240	0.0406
DS@BioMed	555	0.5580	0.1355	0.0600	0.2606	0.5239	0.0548
DS@BioMed	556	0.5580	0.1355	0.0600	0.2606	0.5239	0.0548
DLNU_CCSE	673	0.5462	0.0924	0.0982	0.2279	0.5167	0.0306
CS_Morgan	614	0.5458	0.1184	0.1024	0.2447	0.4501	0.0351
DS@BioMed	313	0.4454	0.0950	0.0899	0.3122	0.6271	0.0504
DS@BioMed	465	0.4454	0.0950	0.0899	0.3122	0.6271	0.0504
DS@BioMed	314	0.4433	0.0952	0.0893	0.3351	0.6231	0.0508
CS_Morgan	615	0.4143	0.0442	0.0289	0.2614	0.6769	0.0199
MICLab	677	0.3739	0.0823	0.0510	0.1601	0.4985	0.0181

Table 9

Performance of the participating teams in the ImageCLEFmedical 2024 Caption Prediction for the metrics CIDEr, CLIPScore, RefCLIPScore, and MedBERTScore.

Group Name	Run	CIDEr	CLIPScore	RefCLIPScore	MedBERTScore
pclmed	634	0.2681	0.8236	0.8176	0.6323
CS_Morgan	429	0.2450	0.8213	0.8155	0.6327
DarkCow	220	0.2243	0.8184	0.8117	0.6292
CS_Morgan	527	0.2241	0.8208	0.8143	0.6315
CS_Morgan	526	0.2199	0.8242	0.8147	0.6300
pclmed	633	0.2597	0.8231	0.8169	0.6254
CS_Morgan	525	0.2034	0.8227	0.8121	0.6298
pclmed	632	0.2521	0.8217	0.8162	0.6242
auebnpgroup	630	0.1769	0.8041	0.7987	0.6261
auebnpgroup	635	0.1762	0.8040	0.7986	0.6260
auebnpgroup	646	0.1758	0.8041	0.7988	0.6261
auebnpgroup	647	0.1459	0.7936	0.7912	0.6291
DarkCow	243	0.1585	0.8132	0.8014	0.6233
2Q2T	643	0.2200	0.8271	0.8138	0.6224
2Q2T	682	0.2200	0.8271	0.8138	0.6224
CS_Morgan	613	0.1708	0.8166	0.8067	0.6262
CS_Morgan	529	0.1619	0.8151	0.8071	0.6243
2Q2T	683	0.2204	0.8284	0.8137	0.6212
auebnpgroup	650	0.1597	0.7990	0.7948	0.6212
CS_Morgan	528	0.1730	0.8193	0.8075	0.6246
auebnpgroup	564	0.1728	0.8045	0.7968	0.6197
MICLab	678	0.1582	0.8159	0.8049	0.6172
auebnpgroup	605	0.1305	0.8037	0.7962	0.6174
auebnpgroup	639	0.1293	0.7858	0.7845	0.6141
auebnpgroup	577	0.1292	0.7832	0.7826	0.6134
2Q2T	512	0.1923	0.8215	0.8147	0.6169
2Q2T	684	0.1948	0.8226	0.8141	0.6162
2Q2T	592	0.1950	0.8226	0.8141	0.6161
2Q2T	595	0.1950	0.8226	0.8141	0.6161
MICLab	676	0.1229	0.7989	0.7915	0.6142
DLNU_CCSE	674	0.1688	0.7967	0.7904	0.6130
DarkCow	221	0.1442	0.8244	0.8100	0.6016
Kaprov	559	0.1070	0.7922	0.7872	0.6089
MICLab	274	0.1082	0.7688	0.7694	0.5963
DLNU_CCSE	675	0.0859	0.7562	0.7506	0.5921
DS@BioMed	571	0.0715	0.7756	0.7748	0.5804
DS@BioMed	563	0.0715	0.7756	0.7748	0.5804
DBS-HHU	637	0.0644	0.7842	0.7750	0.5827
DBS-HHU	645	0.0644	0.7842	0.7750	0.5827
KDE-medical-caption	557	0.0384	0.7651	0.7610	0.5697
KDE-medical-caption	544	0.0499	0.7615	0.7577	0.5700
KDE-medical-caption	424	0.0449	0.7608	0.7580	0.5683
KDE-medical-caption	423	0.0449	0.7608	0.7580	0.5683
KDE-medical-caption	460	0.0425	0.7592	0.7551	0.5674
DS@BioMed	555	0.1043	0.7999	0.7948	0.5487
DS@BioMed	556	0.1043	0.7999	0.7948	0.5487
DLNU_CCSE	673	0.0145	0.6913	0.6989	0.5517
CS_Morgan	614	0.0288	0.6853	0.6924	0.5563
DS@BioMed	313	0.0425	0.7757	0.7675	0.4282
DS@BioMed	465	0.0425	0.7757	0.7675	0.4282
DS@BioMed	314	0.0449	0.7850	0.7736	0.4308
CS_Morgan	615	0.0034	0.6665	0.6698	0.4062
MICLab	677	0.0092	0.6366	0.6614	0.3714