Published in Proceedings of the 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, 5-8 September 2022, which should be cited to refer to this work.
DOI: 10.1007/978-3-031-13643-6_31

Overview of the ImageCLEF 2022: Multimedia Retrieval in Medical, Social Media and Nature Applications

Bogdan Ionescu[1], Henning Müller[2], Renaud Péteri[3], Johannes Rückert[4], Asma Ben Abacha[5], Alba G. Seco de Herrera[6], Christoph M. Friedrich[4], Louise Bloch[4], Raphael Brüngel[4], Ahmad Idrissi-Yaghir[4], Henning Schäfer[4], Serge Kozlovski[7], Yashin Dicente Cid[8], Vassili Kovalev[7], Liviu-Daniel Ştefan[1], Mihai Gabriel Constantin[1], Mihai Dogariu[1], Adrian Popescu[9], Jérôme Deshayes-Chossart[9], Hugo Schindler[9], Jon Chamberlain[6], Antonio Campello[10], and Adrian Clark[6]

[1] Politehnica University of Bucharest, Bucharest, Romania
bogdan.ionescu@upb.ro
[2] University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
[3] University of La Rochelle, La Rochelle, France
[4] University of Applied Sciences and Arts Dortmund, Dortmund, Germany
[5] Microsoft, Redmond, Washington, USA
[6] University of Essex, Colchester, UK
[7] Institute for Informatics, Minsk, Belarus
[8] University of Warwick, Coventry, UK
[9] Université Paris-Saclay, CEA, LIST, F-91120, Palaiseau, France
[10] Wellcome Trust, London, UK

Abstract. This paper presents an overview of the ImageCLEF 2022 lab that was organized as part of the Conference and Labs of the Evaluation Forum – CLEF Labs 2022. ImageCLEF is an ongoing evaluation initiative (first run in 2003) that promotes the evaluation of technologies for annotation, indexing and retrieval of visual data with the aim of providing information access to large collections of images in various usage scenarios and domains. In 2022, the 20th edition of ImageCLEF runs four main tasks: (i) a *medical* task that groups two previous tasks, i.e., caption analysis and tuberculosis prediction, (ii) a *social media* aware task on estimating potential real-life effects of online image sharing, (iii) a *nature* coral task about segmenting and labeling collections of coral reef images, and (iv) a new *fusion* task addressing the design of late fusion schemes for boosting the performance, with two real-world applications: image search diversification (retrieval) and prediction of visual interestingness (regression). The benchmark campaign received the participation of over 25 groups submitting more than 258 runs.

Keywords: Medical image classification · medical image caption analysis · tuberculosis prediction · coral image segmentation and classification · prediction of effects of online image sharing · late fusion for search diversification and interestingness prediction · ImageCLEF lab

1 Introduction

ImageCLEF[11] is the image retrieval and classification lab of the CLEF (Conference and Labs of the Evaluation Forum) conference. ImageCLEF has started in 2003 with only four participants [8]. It increased its impact with the addition of medical tasks in 2004 [7], attracting over 20 participants already in the second year. An overview of ten years of the medical tasks can be found in [22]. It continued the ascending trend, reaching over 200 participants in 2019 and over 110 in 2020 despite the outbreak of the covid-19 pandemic. The tasks have changed much over the years but the general objective has always been the same, i.e., *to combine text and visual data to retrieve and classify visual information.* Tasks have evolved from more general object classification and retrieval to many specific application domains, e.g., nature, security, medical, Internet. A detailed analysis of several tasks and the creation of the data sets can be found in [29]. ImageCLEF has shown to have an important impact over the years, already detailed in 2010 [39, 40].

Since 2018, ImageCLEF uses the crowdAI platform, now migrated to AIcrowd[12] from 2020, to distribute the data and receive the submitted results. The system allows having an online leader board and gives the possibility to keep data sets accessible beyond competition, including a continuous submission of runs and addition to the leader board. Over the years, ImageCLEF and also CLEF have shown a strong scholarly impact that was analyzed in [39, 40]. For instance, the term "ImageCLEF" returns on Google Scholar[13] over 5,990 article results (search on June 13th, 2022). This underlines the importance of evaluation campaigns for disseminating best scientific practices. We introduce here the four tasks that were run in the 2022 edition[14], namely: ImageCLEFmedical, ImageCLEFfusion, ImageCLEFaware, and ImageCLEFcoral.

2 Overview of Tasks and Participation

ImageCLEF 2022 consists of four main tasks with the objective of covering a *diverse range* of multimedia retrieval applications, namely: *medicine*, *social media and Internet*, and *nature* applications. It followed the 2019 tradition [20] of diversifying the use cases [34, 5, 11, 36, 23]. The 2022 tasks are presented as follows:

- **ImageCLEFmedical**. Medical tasks have been part of ImageCLEF every year since 2004. In 2018, all but one task were medical, but little interaction happened between the medical tasks. For this reason, starting with 2019, the medical tasks were focused towards one specific problem but combined as a single task with several subtasks. This allows exploring synergies between the domains:

[11] `http://www.imageclef.org/`
[12] `https://www.aicrowd.com/`
[13] `https://scholar.google.com/`
[14] `https://www.imageclef.org/2022/`

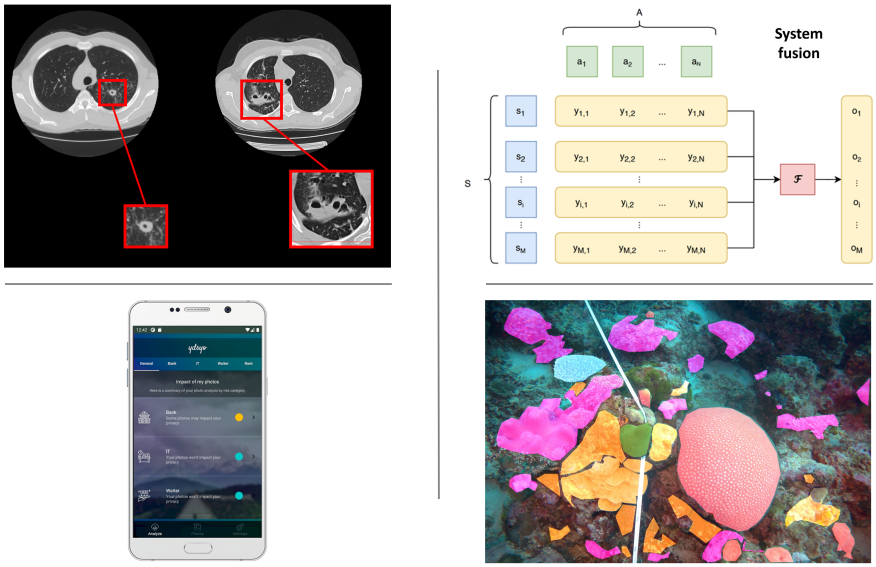

Fig. 1: Sample images from (left to right, top to bottom): ImageCLEFmedical tuberculosis prediction, ImageCLEFfusion with late fusion scheme, ImageCLEFaware with estimating potential real-life effects of online image sharing mobile application, and ImageCLEFcoral with segmenting and labeling collections of coral reef images.

- *Caption*: This is the 6th edition of the task in this format, however, it is based on previous medical tasks. The task is once again running with both the "concept detection" and "caption prediction" subtasks [36], after the former was brought back last year based on the lessons learned in previous editions [17, 14, 18, 31, 32, 30]. The "caption prediction" subtask focuses on composing coherent captions for the entirety of a radiology image, while the "concept detection" subtask focuses on identifying the presence of relevant concepts in the same corpus of radiology images. After a smaller data set of manually annotated radiology images was used last year, the 2022 edition once again uses a larger dataset based on ROCO data [33], which was already used in 2020 and 2019.
- *Tuberculosis*: This is the 6th edition of the task. The main objective is to provide an automatic CT-based evaluation of tuberculosis (TB) patients. This is done by detecting and assessing visual TB-related findings based on the automatic analysis of lung CT scans. Caverns are one of the finding types which need specific attention. Even after successful treatment which fulfills the existing criteria of recovery caverns may still contain colonies of Mycobacterium Tuberculosis that could lead to unpredictable disease relapse. Therefore, finding and describing caverns

helps to evaluate the quality of the treatment and plan recovery and control routines after the active treatment phase. In this year's edition, participants need to solve two subtasks - the first one is cavern detection, and the second one is providing cavern reporting which includes three binary labels: "Thick walls", "Calcification", and "Foci around" [23].

- **ImageCLEFfusion**. This is the 1st edition of the task. The main objective for this task is the development of late fusion or ensembling approaches, that are able to use prediction results from pre-computed inducers in order to generate better, improved prediction outputs. This edition of the task proposes two challenges: a regression challenge that uses media interestingness data, and a retrieval challenge that uses image search result diversification data. The task uses actual inducers developed by real users.
- **ImageCLEFaware**. This was the 2nd edition of the task [24]. The disclosure of personal data is done in a particular context and users are often unaware that their data can be reused in other contexts. It is thus important to give feedback to users about the effects of personal data sharing. The objective was to automatically provide a rating of a visual user profile in different real-life situations. The dataset created specifically for the 2021 edition of the task was expanded in order to make the evaluation more robust. Data were sampled from YFCC100 and were further anonymized in order to comply with GPDR.

Table 1: Key figures regarding participation in ImageCLEF 2022.

Task	Groups that submitted results	Submitted runs	Submitted working notes
Caption	12	157	13
Tuberculosis	6	43	5
Fusion	5	39	4
Aware	3	9	2
Coral	2	11	2
Overall	28	258	26

- **ImageCLEFcoral**. The 4th edition of the task follows the directions of previous years [3, 4, 6]. The task consists on two subtasks which aim to automatically segment and label with types of benthic substrate a collection of coral reef images. The first subtask uses bounding boxes to annotate the images while the second subtask segment the images pixel-wise using polygons. As in the third edition, in 2022 [5] the training and test data form the complete set of images required to form a 3D reconstruction of the environment.

To participate in the evaluation campaign, the research groups had to register by following the instructions on the ImageCLEF 2022 web page[15]. To ease the overall management of the campaign, in 2022 the challenge was organized through the AIcrowd platform[16]. To actually get access to the data sets, the participants were required to submit a signed End User Agreement (EUA). Table 1 summarizes the participation in ImageCLEF 2022, indicated both per task and for the overall lab. The table also shows the number of groups that submitted runs and the ones that submitted a working notes paper describing the techniques used. Teams were allowed to register for participating in several tasks.

After a decrease in participation in 2016, the participation increased in 2017 and 2018, and increased again in 2019. In 2018, 31 teams completed the tasks and 28 working notes papers were received. In 2019, 63 teams completed the tasks and 50 working notes papers were retrieved. In 2020, 40 teams completed the tasks and submitted working notes papers. In 2021, 42 teams completed the tasks and we received 30 working notes papers. In 2022, 28 teams completed the tasks and we received 26 working notes papers. We can observe a drop in participation compared to 2019 and also 2021. The 2022 edition marks the end of the pandemic. Also, one of the medical tasks, i.e., the visual question answering, was not organized this year. Nevertheless, the number of submitted runs is similar to 2021 regardless the fact that less teams submitted, namely 258 vs. 256. Teams were more involved in finding solutions. Overall, even in its 20th anniversary, ImageCLEF continues to provide a strong evaluation benchmark.

In the following sections, we present the tasks. Only a short overview is reported, including general objectives, description of the tasks and data sets, and a short summary of the results. A detailed review of the received submissions for each task is provided with the task overview working notes: Caption [36], Tuberculosis [23], Fusion [11], Aware [34], and Coral [5].

3 The Caption Task

The caption task was first proposed as part of the ImageCLEFmedical [18] in 2016 aiming to extract the most relevant information from medical images. Hence, the task was created to condense visual information into textual descriptions. In 2017 and 2018 [14, 17] the ImageCLEFcaption task comprised two subtasks: concept detection and caption prediction. In 2019 [31] and 2020 [32], the task concentrated on extracting Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUIs) [1] from radiology images. In 2021 [30], both subtasks, concept detection and caption prediction, were running again due to participants demands. The focus in 2021 was on making the task more realistic by using fewer images which were all manually annotated by medical doctors. For the 2022 ImageCLEFmedical Caption task [36], both subtasks are continued albeit with an extended version of the ROCO data set used for both subtasks, which was already used in 2020 and 2019.

[15] https://www.imageclef.org/2022/

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

3.1 Task Setup

The ImageCLEFmedical Caption 2022 [36] follows the format of the previous ImageCLEFmedical caption tasks. In 2022, the overall task comprises two subtasks: "Concept Detection" and "Caption prediction". The concept detection subtask focuses on predicting Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUIs) [1] based on the visual image representation in a given image. The caption prediction subtask focuses composing coherent captions for the entirety of the images.

The detected concepts are evaluated using the balanced precision and recall trade-off in terms of F1-scores, as in previous years. This year, a secondary F1-score based on manually curated concepts regarding image modality and x-ray anatomy was introduced. The predicted captions are evaluated using the BLEU score independent from the first subtask and designed to be robust to variability in style and wording. In addition to the BLEU score, a secondary ROUGE score was provided. After the submission period ended, a number of additional scores were calculated and published: METEOR, CIDEr, SPICE, and BERTScore.

3.2 Data Set

In 2022, an extended subset of the ROCO [33] data set is used for both subtasks, which originates from biomedical articles of the PMC Open Access Subset[17] [35] and was extended with new images added since the last time the data set was updated. In the previous edition, in an attempt to make the task more realistic, the data set contained a smaller number of real radiology images annotated by medical doctors which resulted in high-quality concepts. Additional data of similar quality is hard to acquire and so it was decided to return to the data set already used in 2020 and 2019. From the captions, UMLS® concepts were generated and concepts regarding anatomy and image modality were manually validated for all images.

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

- *Training set* including 83,275 radiology images and associated captions and concepts.
- *Validation set* including 7,645 radiology images and associated captions and concepts.
- *Test set* including 7,645 radiology images.

3.3 Participating Groups and Submitted Runs

In the sixth edition of the ImageCLEFmedical Caption task, 20 teams registered and signed the End-User-Agreement that is needed to download the development data. 12 teams submitted 157 runs for evaluation (all 12 teams submitted

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

Table 2: Performance of the participating teams in the ImageCLEFmedical 2022 concept detection subtask. The best run per team is selected. Teams with previous participation in 2021 are marked with an asterisk.

Team	Institution	F1-Score
AUEB-NLP-Group*	Department of Informatics, Athens University of Economics and Business, Athens, Greece	0.4511
CMRE-UoG (fdallaserra)	Canon Medical Research Europe, Edinburgh, UK and University of Glasgow, Glasgow, UK	0.4505
CSIRO*	Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Herston, Queensland, Australia and CSIRO Data61, Imaging and Computer Vision Group, Pullenvale, Queensland, Australia and Queensland University of Technology, Brisbane, Queensland, Australia	0.4471
eecs-kth	KTH Royal Institute of Technology, Stockholm, Sweden	0.4360
vcmi	University of Porto, Porto, Portugal and INESC TEC, Porto, Portugal	0.4329
PoliMi-ImageClef	Politecnico di Milano, Milan, Italy	0.4320
SSNSheerinKavitha	Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, India	0.4184
IUST_NLPLAB	School of Computer Engineering, Iran University of Science and Technology, Tehran, Islamic Republic Of Iran	0.3981
Morgan_CS	Morgan State University, Baltimore, MD, USA	0.3520
kdelab*	KDE Laboratory, Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan	0.3104
SDVA-UCSD	San Diego VA HCS, San Diego, CA, USA	0.3079

working notes) attracting more attention than in 2021. Each of the groups was allowed a maximum of 10 graded runs per sub task. 11 teams participated in the concept detection subtask this year, 3 of those teams also participated in 2021. 10 teams submitted runs to the caption prediction subtask, 4 of those teams also participated in 2021. Overall, 9 teams participated in both subtasks, two teams participated only in the concept detection subtask and one team participated only in the caption prediction subtask.

In the concept detection subtasks, the groups used primarily multi-label classification systems and image retrieval systems, much like in the 2021 challenge. Multi-label classification systems outperformed retrieval-based systems for most

Table 3: Performance of the participating teams in the ImageCLEFmedical 2022 caption prediction subtask. The best run per team is selected. Teams with previous participation in 2021 are marked with an asterisk.

Team	Institution	BLEU Score
IUST_NLPLAB	School of Computer Engineering, Iran University of Science and Technology, Tehran, Islamic Republic Of Iran	0.4828
AUEB-NLP-Group*	Department of Informatics, Athens University of Economics and Business, Athens, Greece	0.3221
CSIRO*	Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Herston, Queensland, Australia and CSIRO Data61, Imaging and Computer Vision Group, Pullenvale, Queensland, Australia and Queensland University of Technology, Brisbane, Queensland, Australia	0.3114
vcmi	University of Porto, Porto, Portugal and INESC TEC, Porto, Portugal	0.3058
eecs-kth	KTH Royal Institute of Technology, Stockholm, Sweden	0.2917
CMRE-UoG (fdallaserra)	Canon Medical Research Europe, Edinburgh, UK and University of Glasgow, Glasgow, UK	0.2913
kdelab*	KDE Laboratory, Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan	0.2782
Morgan_CS	Morgan State University, Baltimore, MD, USA	0.2549
MAI_ImageSem*	Institute of Medical Information and Library, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China	0.2211
SSNSheerinKavitha	Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, India	0.1595

of the teams who experimented with both, and while the winner was a multi-label classification approach, the second placing team with an F1-score only 0.0006 less than the winning team, used an image retrieval system based on the winning approach from last year.

In the caption prediction subtask, most teams experimented with Transformer-based architectures and image retrieval systems. Only one team used a multi-label classification approach, and it achieved by far the best BLEU score. However, it did not score as well on most of the other employed metrics. Transfer

Learning has frequently been used for pre-training, from a variety of different data sets.

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

3.4 Results

This year's models for concept detection do not show an increased F1-score compared to last year, however due to the much larger data set and number of concepts used in this year's challenge, this is not surprising. Comparing it to the 2020 results, where a data set of similar size was used, the F1-scores show a clear improvement. There are no radically new approaches used in this year's concept detection subtask, but the teams experimented with, optimised and re-combined many different existing techniques and created competitive solutions using both multi-label classification systems and image retrieval systems.

Similar to the concept detection, the BLEU scores in the caption prediction subtask are overall lower compared to last year, which can be explained by the larger data set and more varied captions. Since there was no caption prediction subtask running in 2020, no comparable scores for a similar data set exist. An in-depth analysis is presented in [36].

3.5 Lessons Learned and Next Steps

This year's caption task of ImageCLEFmedical once again ran with both subtasks, concept detection and caption prediction. It returned to a larger, ROCO-based data set for both challenges after a smaller, manually annotated data set was used last year. It attracted 12 teams who submitted 157 runs overall, a stronger participation compared to last year. Some changes were introduced for the scores, with a secondary F1-score related to manually curated concepts for image modality and x-ray anatomy added to the concept detection, and several new scores added to the caption prediction subtask which was appreciated by the teams as it highlighted the difficulty of evaluating caption similarity and showed that models performing worse on the BLEU score could perform better in several of the other metrics instead.

With the bigger data set, most teams were more successful in training multi-label classification models compared to image retrieval models for the concept detection. For the caption prediction, most teams used Transformer-based models, but the winning models in terms of the BLEU score was a multi-label classification model.

For next year's ImageCLEFmedical Caption challenge, some possible improvements include adding more manually validated concepts like increased anatomical coverage and directionality information, reducing recurring captions, more fine-grained CUI filters, improving the caption pre-processing, and using a different primary score for the caption prediction challenge, since the BLEU score has some disadvantages which were highlighted by this year's caption prediction results. It will also be important to make sure that no models are used

that were pre-trained on PubMedCentral data, since these models will already have seen the original captions.

4 The Tuberculosis Task

Tuberculosis (TB) is a bacterial infection caused by a germ called Mycobacterium tuberculosis. More than a century after its discovery, the disease remains a persistent threat and one of the top 10 causes of death worldwide according to the WHO [41]. The bacteria usually attack the lungs and generally TB can be cured with antibiotics. However, the different types of TB require different treatments, and therefore detection of the specific case characteristics is an important real-world task.

In the previous editions of this task, the setup evolved from year to year. In the first two editions [14, 16] participants had to detect Multi-drug resistant patients (MDR subtask) and classify the TB type (TBT subtask) both based only on the CT image. After 2 editions it was concluded to drop the MDR subtask because it seemed impossible to solve based only on the image, and the TBT subtask was also suspended because of a very little improvement in the results between the 1st and the 2nd editions. At the same time, most of the participants obtained good results in the severity scoring (SVR) subtask introduced in 2018. In the 3d edition Tuberculosis task [15] was restructured to allow usage of the uniform dataset, and included two subtasks - a continued Severity Score (SVR) prediction subtask and a new subtask based on providing an automatic report (CT Report) on the TB case. In the 4th edition [25], the SVR subtask was dropped and the automated CT report generation task was modified to be lung-based rather than CT-based. In the 5th edition [24], the task organizers have decided to discontinue the CTR task and brought back to life the Tuberculosis Type classification task from the 1st and 2nd ImageCLEFmed Tuberculosis editions to check if recent Machine Learning and Deep Learning methods allow improving previous rather low results.

In this year's edition [23] the task was dedicated to the caverns detection and report, which were split into two subtasks. The first subtask (Caverns Detection) focused on detection, i.e., participants must detect lung cavern regions in lung CT images associated with lung tuberculosis. The problem is important because even after successful treatment which fulfills the existing criteria of recovery the caverns may still contain colonies of Mycobacterium Tuberculosis that could lead to unpredictable disease relapse. The second subtask (Caverns Report) was the classification of caverns. Participants must predict 4 binary features of caverns suggested by experienced radiologists.

4.1 Task Setup

In this task, participants had to automatically detect lung cavern regions in lung CT images associated with lung tuberculosis in the first subtask, and predict 3 binary features of caverns suggested by experienced radiologists. So the first

subtask was a 3D object detection task, and the second one was a multi-label classification problem.

4.2 Data Set

In this edition, separate data sets were provided for each subtask. The Caverns Detection dataset contained 559 train and 140 test cases, while the Caverns Report data included just 60 train and 16 test cases due to labelled data scarcity. Each CT image corresponded to one unique patient. For all patients, we provided 3D CT images with a slice size of 512×512 pixels and a variable number of slices (the median number was 128). All train CTs for both subtasks were accompanied by caverns area bounding boxes (if any), and labelling of caverns was provided for Cavern Report subtask. Since bounding boxes were provided for all CTs, participants were welcomed to use data from one subtask for the another.

Same as in the previous year, for all patients we provided two versions of automatically extracted masks of the lungs obtained using the methods described in [13, 27].

4.3 Participating Groups and Submitted Runs

In 2022, 6 groups from 5 countries submitted at least one run. 4 group participated in each task, and 2 groups participated in both tasks. Similar to the previous editions, each group could submit up to 10 runs. 43 scored runs were submitted in total (17 for Caverns Detection and 26 for Caverns Report).

All groups used 2D or/and 3D CNNs in both tasks. For the Caverns Detection subtask one group tried both 3D approach using customized 3D Retina U-Net based model and projection-wise 2D approach using YOLO v5 detection netwoks; another group reported 2D slice-wise approach using the YOLO v3. For the Caverns Report subtask three participants reported usage of 3D-only approach, two of them utilized custom 3D CNN, and one used ResNet34 with convolutional block attention model (CBAM); one group used slice-wise approach, but in addition to 2D CNN (EffcientNet, DenseNet) also used SRGAN for data preprocessing. The majority of participants used transfer learning techniques where possible, and executed some pre-processing steps, such as resizing, grouping, normalization, slice filtering etc.

4.4 Results

The Caverns Detection task was scored using the mean average precision at the different intersection over union (IoU) thresholds score. The Caverns Report task was evaluated as a multi-label classification problem and scored using mean AUC as primary score and minimum AUC as secondary score. Tables 4 and 5 shows the final results for each group's best run in both tasks. More detailed results, including metric description and other performance measures, are presented in the overview article [23].

Table 4: Results obtained by the participants of the Caverns Detection task. Only the best run of each participant is reported.

Group name	*Institution*	*map_iou*
CSIRO	Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Herston, Queensland, Australia and CSIRO Data61, Imaging and Computer Vision Group, Pullenvale, Queensland, Australia and Queensland University of Technology, Brisbane, Queensland, Australia	0.504
SenticLab.UAIC	Alexandru Ioan Cuza University of Iasi, Romania	0.295
KDE-lab	KDE Laboratory, Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan	0.185
SDVA-UCSD	San Diego VA HCS, San Diego, CA, USA	0.000

Table 5: Results obtained by the participants of the Caverns Report task. Only the best run of each participant is reported.

Group name	*Institution*	*Mean AUC*	*Min AUC*
SDVA-UCSD	San Diego VA HCS, San Diego, CA, USA	0.687	0.513
KDE-lab	KDE Laboratory, Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan	0.658	0.317
KL_BP_SSN	Sri Sivasubramaniya Nadar college of Engineering, Chennai, India	0.536	0.413
SSN_Dheepak_Kavitha	SSN College of Engineering, Chennai, India	0.461	0.256

4.5 Lessons Learned and Next Steps

The results obtained in the task cannot be compared to the previous editions, since it's the first appearance of caverns-dedicated tasks. Furthermore, this is the first time for the TB task when we switched from classification problems to the detection problem.

The best result of Caverns Detection subtask was achieved by the CSIRO group using a custom neural network with 3D Retina U-Net-based architecture in a combination with developed plane-based bounding box merging postprocessing routine. The second-ranked SenticLab.UAIC group used nodule detection CNN. The 3rd ranked KDE-lab group used slice-wise analysis with YOLO v3 CNN.

The best result of the Caverns Report subtask was achieved by the SDVA-UCSD group using the 3D CBAM Resnet model and a semi-supervised training strategy which allowed involving data set provided in the detection task. The second-ranked KDE-lab group used slice-wise analysis using pre-trained 2D CNN (EffcientNet, DenseNet) and also used resolution increase using SRGAN as a preprocessing step. The 3rd ranked SSN_Dheepak_Kavitha group used custom 3D CNN.

Results analysis shows, that the best scores are reasonably high for both subtasks, and the top score for the Caverns Detection is better than we expected taking into account the complexity of the 3D detection problem. Based on the participants' approach analysis we can note that both winning solutions used advantages of volumetric analysis to the contrary of previous task editions, where projection-based approaches were more effective. As a result, we can conclude that despite a rather low number of participants this year, we see interesting approaches with quite a high score, so in general, the task is successful and its outcome is informative and useful.

Possible updates for future editions of caverns-related TB task should consider: (i) extending data set sizes and labels count for caverns report; (ii) switching from detection to segmentation problem.

5 The Fusion Task

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

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

5.1 Task Setup

This first edition of the ImageCLEFfusion task [11] consists of two challenges: a regression challenge involving media interestingness (ImageCLEFfusion-int)

for which we provide output data from 29 base models, and a retrieval challenge involving result diversification (ImageCLEFfusion-div) for which we provide outputs data from 56 inducers. Participants were asked to develop late fusion learning strategies based on the outputs of the inducers associated with the media samples for each of the subtask. Evaluation was performed using MAP@10 for the ImageCLEFfusion-int task, and F1@20 and Cluster Recall@20 for ImageCLEFfusion-div task. Participants were invited to submit for either or both tasks.

5.2 Data Set

The ImageCLEFfusion-int task uses data from the Interestingness10k dataset [10], specifically, the image-based prediction data associated with the 2017 MediaEval Predicting Media Interestingness task [12]. We provide prediction outputs from the 29 systems submitted during this benchmarking task, dividing the available data into 1877 data samples for the training of the proposed fusion systems and 558 for testing.

On the other hand, the ImageCLEFfusion-div task uses data associated with the e Retrieving Diverse Social Images dataset [21], specifically the DIV150Multi challenge [19]. The retrieval outputs provided from the 56 systems are divided into 60 queries for the training data and 63 queries for the testing set.

In both training sets, we provide the inducer outputs, the necessary scripts for metric computation, the performance for each of the inducers according to the official metrics, and ground truth data. For the testing sets we only provide the inducer outputs. It is also important to note that participants were not allowed to use external inducers, being limited only to the ones we provide, as our intention is to have a fair assessment of the performance of the late fusion approach in itself, without changing the inducer set.

Table 6: Participation in the ImageCLEF-int 2022 task: the best score from all runs for each team.

Team	*#Runs*	*MAP@10*
AIMultimediaLab	5	0.2192
ssn_it	1	0.1106
UECORK	8	0.1097

5.3 Participating Groups and Submitted Runs

Three teams submitted runs for each task, while only one team participated in both tasks. A total of 14 runs are submitted for the interestingness task, while the diversification task is more successful, with 25 submitted runs.

The analysis of the submitted methods shows two important types of approaches proposed by the participants for this task. The first significant type is

Table 7: Participation in the ImageCLEF-div 2022 task: the best score from all runs for each team.

Team	*#Runs*	*F1@20*	*CR@20*
AIMultimediaLab	5	0.6216	0.4916
klssncse	10	0.5634	0.4414
shreya_sriram	10	0.5604	0.4373

based on weighting the inducer output by implementing several different techniques. For example, one group used a simple grid search based on the performance of inducers on the training set, where higher weights are assigned to better-performing inducers. Other weighted approaches use a learning method for determining the optimal inducer weights, including learning methods based on Genetic Algorithms, Particle Swarm Optimization, and Trust Region Constrained Optimization.

The second type of approach is based on passing the inducer prediction outputs through a learning mechanism that provides the final fusion results, thus learning the way inducers interact for a given sample. In this category, some participants proposed implementing sets of traditional learning methods like kNN, Classification and Regression Trees, or SVR, while others chose neural networks as the base for the fusion engine, including approaches based on DeepFusion, MLP models, and Keras Regressors. Finally, it is worth noting that one team proposed a method where the output of several DNN-based fusion engines is passed through a final stage represented by a voting regressor.

5.4 Results

The results are presented in Table 6 for the interestingness task, and Table 7 for the diversification task. In both tasks, the best performance is achieved with a DeepFusion type approach [9], submitted by the AIMultimediaLab team. The best performance for the ImageCLEF-int task is a MAP@10 value of 0.2192, while for the ImageCLEF-div task a F1@20 of 0.6216 and a CR@20 of 0.4916 is achieved.

Overall, while results for the diversification task seem to be higher than those recorded in the interestingness task, it is important to note that the improvement over the provided inducers is greater for the interestingness task. Specifically, the improvement in the interestingness task over the average inducer performance, which is a MAP@10 value of 0.0946 is 131%. For the diversification task, the average inducer performance is an F1@20 value of 0.5313, thus the submitted systems show an improvement of almost 17%. While this may be the result of greater initial performance on the diversification task, it is also worth to note that the degree of complexity associated with the diversification task and its inducer outputs is greater.

5.5 Lessons Learned and Next Steps

The results presented this year are encouraging, especially considering the fact that all teams performed above the performance of the average inducers. A large variety of approaches, ranging from simple statistical methods to more complex approaches that require learning inducer interactions, like SVMs, classification and regression trees, and deep neural networks.

For the next edition of this task, we believe it is very important to continue with these two datasets, as this will allow us to study the year-to-year improvement of the proposed fusion techniques. Also, we will study the possibility of adding another dataset, that will target another complex type of machine learning task, whether it is a multi-class classification task, or multi-label classification.

6 The Aware Task

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

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

6.1 Task Setup

This is the second edition of the task and consists of one challenge. Participants are provided with automatic object detections for the images and with object ratings per situation. Then, the objective is to propose a ranking of user profiles which is as close as possible to the crowdsourced one. Data were split into 600/200/200 profiles for training/validation and test. The Pearson correlation coefficient between manual and automatic profile rankings was used to evaluate the quality of proposed runs. The final scores were calculated by averaging correlations obtained for individual situations.

6.2 Data Set

A data set of 1,000 user profiles with 100 photos per profile was created and annotated with an "appeal" score for four real-life situations via crowdsourcing. The modeled situations are demands for: a bank credit, an accommodation, a job as an IT engineer, a job as a waiter. Participants to the experiment were asked to provide a global rating of each profile in each situation modeled using a 7-points Likert scale ranging from "strongly unappealing" to "strongly appealing". The averaged "appeal" score was used to create a ground truth composed of ranked users in each modeled situation. User profiles are created by repurposing a subset of the YFCC100M dataset [38].

Situations are modeled by crowdsourcing visual objects ratings. Similar to profile crowdsourcing, object ratings are collected for each situation using a 7-points Likert scale with ratings between -3 (strongly negative influence) to +3 (strongly positive influence). The averaged rating is computed and provided to participants. A Faster R-CNN object detector was trained in order to detect objects in images. The detection dataset combines objects from OpenImages [26], ImageNet [37] and COCO [28]. Only objects with at least one non-zero situation rating were kept. All objects detected in the 100 images of a profile were provided to participants, along with the detection probability and the associated bounding box. Given a situation, the combination of the ratings of objects and of their automatic detection enables the automatic computation of a profile score.

Given the personal nature of the included profiles, the dataset was anonymized in order to comply with GDPR. Participants did not have access to the images, and the user IDs and the object names were hashed.

6.3 Participating Groups and Submitted Runs

Three teams registered for the task this year, all from the SSN College of Engineering, Chennai, India. All three teams submitted a total of nine runs.

6.4 Results

The participants tested a range of techniques to rate and rank user profiles, notably: random forest regressors, extra tree regressors and dense neural networks. Attention was also given to the preprocessing step in order to make the most of the available training data, with different runs using various combinations of object detections, confidence scores, object counts, and/or bounding boxes. The best reported Pearson correlation is 0.544, and was obtained with random forest regressor. The best score reported this year is similar to the one from 2021.

6.5 Lessons Learned and Next Steps

The participation this year was better than last year, but still low. The interaction with participants was smooth, and there were no problems with the dataset usage. The availability of a larger dataset allowed the use of different learning

Table 8: Results of the Aware 2022 task.

Team	*#Runs*	*Pearson*
SSNCSE_KS_NA_AKR_CB	5	0.544
JBTTM	2	0.139
ssnce-cse-JT	2	0.0

techniques, including deep learning ones. The scores reported by participants are interesting, but the task is far from being solved.

For the next edition of the task, we will continue the extension of the dataset to make it more robust and timely. Focus will be put on: (1) further increase the number of user profiles, and (2) use large-scale object detection methods, such as Detic [42], to provide finer-grained profiles.

7 The Coral Task

Marine ecosystem monitoring is a key priority for evaluating ecosystem conditions [2]. Despite a wide range of monitoring programs for tropical coral reefs, there is still a crucial need to establish an effective monitoring process. This process can be made by collecting 3D visual data using autonomous underwater vehicles. The ImageCLEFcoral task organisers have developed a novel multi-camera system that allows large amounts of imagery to be captured by a SCUBA diver or autonomous underwater vehicle in a single dive which will provide useful information for both annotation and further study of the coral.

Previous editions of ImageCLEFcoral in 2019 [3] and 2020 [4] have shown improvements in task performance and promising results on cross-learning between images from geographical regions. The 3rd edition [6] increased the complexity of the task and size of data available to participants through supplemental data, resulting in lower performance than previous years. As with the 3rd edition, in 2022 [5], the training and test data form a complete set of images required to form 3D reconstructions of the marine environment.

7.1 Task Setup

In 2022, the ImageCLEFcoral task followed the format of previous editions [3, 4, 6]. In 2021 participants were again asked to devise and implement algorithms for automatically annotating regions in a collection of images containing several types of benthic substrate, such as hard coral or sponge. As in previous editions, 2022 comprised two sub-tasks: "Coral reef image annotation and localisation" and "Coral reef image pixel-wise parsing" subtasks. The "Coral reef image annotation and localisation" subtask uses bounding boxes, with sides parallel to the edges of the image, for the annotation of regions in a collection of images containing several types of benthic substrates. The "Coral reef image pixel-wise parsing" subtask uses a series of boundary image coordinates which form a single polygon around each identified region in the coral reef images; this has been

dubbed *pixel-wise parsing* (these polygons should not have self-intersections). Participants were invited to make submissions for either or both tasks.

Algorithmic performance is evaluated on the unseen test data using the popular intersection over union metric from the PASCAL VOC[18] exercise. This computes the area of intersection of the output of an algorithm and the corresponding ground truth, normalising that by the area of their union to ensure its maximum value is bounded.

7.2 Data Set

As in previous editions, the data for this ImageCLEFcoral task originates from a growing, large-scale collection of images taken from coral reefs around the world as part of a coral reef monitoring project with the Marine Technology Research Unit at the University of Essex. The images contain annotations of the following 13 types of substrates: Hard Coral – Branching, Hard Coral – Submassive, Hard Coral – Boulder, Hard Coral – Encrusting, Hard Coral – Table, Hard Coral – Foliose, Hard Coral – Mushroom, Soft Coral, Soft Coral – Gorgonian, Sponge, Sponge – Barrel, Fire Coral – Millepora and Algae - Macro or Leaves.

In addition, participants are encouraged to use the publicly available NOAA NCEI data[19] and/or CoralNet[20] to train their approaches. They were also encouraged to explore novel probabilistic computer vision techniques based around image overlap and transposition of data points.

7.3 Participating Groups and Submitted Runs

In 2022, 6 teams registered, of which 2 teams submitted 11 valid runs. Teams were limited to submit 10 runs per subtask. To get a better overview of the submitted runs, the best results for each team are presented in Table 9. Unfortunately, there were no participants to the "Coral reef image pixel-wise parsing" subtask this year. An in-depth analysis is presented in [5].

Table 9: Coral reef image annotation and localisation performance in terms of $MAP0.5IoU$. The best run per team is selected.

Run id	*Team*	$MAP0.5IoU$	$MAP0.0IoU$
183919	HHU	0.396	0.752
185373	UTK	0.003	0.327

[18] `http://host.robots.ox.ac.uk/pascal/VOC/`
[19] `https://www.ncei.noaa.gov/`
[20] `https://coralnet.ucsd.edu/`

7.4 Results

The results from the "Coral reef image annotation and localisation" achieved better higher than in the 2021 edition although they are not directly comparable since the data has been updated in 2022. There was no participation in the "Coral reef image pixel-wise parsing", which is a more complicated task while closer to the real-word problem. More detailed analysis of the results is presented in [6].

7.5 Lessons Learned

As with the 3rd edition, the training and test data formed a complete set of images required to form 3D reconstructions of the marine environment. Unfortunately, no participant has explored yet computer vision techniques based around image overlap and transposition of data points. Therefore, we can still unlock the true potential of the dataset to provide meaningful insights for the analysis of the coral reefs.

8 Conclusion

This paper presents a general overview of the activities and outcomes of the ImageCLEF 2022 evaluation campaign. Four tasks were organised, covering challenges in the medical domain (caption analysis, tuberculosis prediction), social networks and Internet (analysis of the real-life effects of personal data sharing, fusion techniques for retrieval and interestingness prediction), and nature (segmenting and labeling collections of coral images). 28 teams completed the tasks and submitted over 258 runs.

As anticipated already, most of the proposed solutions evolved around state-of-the-art deep neural network architectures. In ImageCLEFcaption, with the bigger data set, most teams were more successful in training multi-label classification models compared to image retrieval models for the concept detection. For the caption prediction, most teams used Transformer-based models, but the winning models in terms of the BLEU score was a multi-label classification model. In ImageCLEFtuberculosis, the results cannot be compared to the previous editions, since it's the first time appearance of caverns-dedicated tasks. Furthermore, this is the first time when we switched from classification problems to the detection problem. The best result was achieved using a custom neural network with 3D Retina U-Net-based architecture in a combination with developed plane-based bounding box merging postprocessing routine. In ImageCLEFfusion, although in its first edition, the results are encouraging. All teams performed above the performance of the average inducer. A large variety of approaches, ranging from simple statistical methods to more complex approaches that require learning inducer interactions, like SVMs, classification and regression trees, and deep neural networks, were explored. In ImageCLEFaware, the participation was better than last year, but still low. The availability of a larger dataset allowed the use of different learning techniques, including deep learning

ones. The scores reported by participants are interesting, but the task is far from being solved. In ImageCLEFcoral, the training and test data formed a complete set of images required to form 3D reconstructions of the marine environment. Unfortunately, no participant has explored yet computer vision techniques based around image overlap and transposition of data points. Therefore, we can still unlock the true potential of the dataset to provide meaningful insights for the analysis of the coral reefs.

ImageCLEF 2022 brought again together an interesting mix of tasks and approaches and we are looking forward to the fruitful discussions at the CLEF 2022 workshop.

Acknowledgements

The ImageCLEFaware and ImageCLEFfusion tasks were supported under the H2020 AI4Media "A European Excellence Centre for Media, Society and Democracy" project, contract #951911. 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. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research **32**(Database-Issue), 267–270 (2004). https://doi.org/10.1093/nar/gkh061
2. Carrillo-García, D.M., Kolb, M.: Indicator framework for monitoring ecosystem integrity of coral reefs in the western caribbean. Ocean Science Journal pp. 1–24 (2022)
3. Chamberlain, J., Campello, A., Wright, J.P., Clift, L.G., Clark, A., García Seco de Herrera, A.: Overview of ImageCLEFcoral 2019 task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org (2019)
4. Chamberlain, J., Campello, A., Wright, J.P., Clift, L.G., Clark, A., García Seco de Herrera, A.: Overview of the ImageCLEFcoral 2020 task: Automated coral reef image annotation. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org (2020)
5. Chamberlain, J., García Seco de Herrera, A., Campello, A., Clark, A.: ImageCLEFcoral task: Coral reef image annotation and localisation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022), LNCS Lecture Notes in Computer Science, Springer, Bologna, Italy (September 5-8 2022)
6. Chamberlain, J., García Seco de Herrera, A., Campello, A., Clark, A., Oliver, T.A., Moustahfid, H.: Overview of the ImageCLEFcoral 2021 task: Coral reef image annotation of a 3d environment. In: CLEF2021 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania (September 21-24 2021)

7. Clough, P., Müller, H., Sanderson, M.: The CLEF 2004 cross–language image retrieval track. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign. Lecture Notes in Computer Science (LNCS), vol. 3491, pp. 597–613. Springer, Bath, UK (2005)
8. Clough, P., Sanderson, M.: The CLEF 2003 cross language image retrieval task. In: Proceedings of the Cross Language Evaluation Forum (CLEF 2003) (2004)
9. Constantin, M.G., Ştefan, L.D., Ionescu, B.: Deepfusion: Deep ensembles for domain independent system fusion. In: International Conference on Multimedia Modeling. pp. 240–252. Springer (2021)
10. Constantin, M.G., Ştefan, L.D., Ionescu, B., Duong, N.Q., Demarty, C.H., Sjöberg, M.: Visual interestingness prediction: A benchmark framework and literature review. International Journal of Computer Vision **129**(5), 1526–1550 (2021)
11. Ştefan, L.D., Constantin, M.G., Dogariu, M., Ionescu, B.: Overview of ImageCLEFfusion 2022 task – Ensembling Methods for Media Interestingness Prediction and Result Diversification. In: CLEF2022 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy (September 5-8 2022)
12. Demarty, C.H., Sjöberg, M., Ionescu, B., Do, T.T., Gygli, M., Duong, N.: Mediaeval 2017 predicting media interestingness task. In: MediaEval workshop (2017)
13. Dicente Cid, Y., Jimenez-del-Toro, O., Depeursinge, A., Müller, H.: Efficient and fully automatic segmentation of the lungs in CT volumes. In: Orcun Goksel, Jimenez-del-Toro, O., Foncubierta-Rodriguez, A., Müller, H. (eds.) Proceedings of the VISCERAL Challenge at ISBI. pp. 31–35. No. 1390 in CEUR Workshop Proceedings (Apr 2015)
14. Dicente Cid, Y., Kalinovsky, A., Liauchuk, V., Kovalev, V., , Müller, H.: Overview of ImageCLEFtuberculosis 2017 - predicting tuberculosis type and drug resistances. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
15. Dicente Cid, Y., Liauchuk, V., Klimuk, D., Tarasau, A., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2019 - automatic ct-based report generation and tuberculosis severity assessment. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 9-12 2019)
16. Dicente Cid, Y., Liauchuk, V., Kovalev, V., , Müller, H.: Overview of ImageCLEFtuberculosis 2018 - detecting multi-drug resistance, classifying tuberculosis type, and assessing severity score. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
17. García Seco de Herrera, A., Eickhoff, C., Andrearczyk, V., , Müller, H.: Overview of the ImageCLEF 2018 caption prediction tasks. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
18. García Seco de Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 medical task. In: Working Notes of CLEF 2016 (Cross Language Evaluation Forum) (September 2016)
19. Ionescu, B., Gînscă, A.L., Boteanu, B., Lupu, M., Popescu, A., Müller, H.: Div150multi: a social image retrieval result diversification dataset with multi-topic queries. In: Proceedings of the 7th international conference on multimedia systems. pp. 1–6 (2016)

20. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), vol. 11438. LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)
21. Ionescu, B., Rohm, M., Boteanu, B., Gînscă, A.L., Lupu, M., Müller, H.: Benchmarking image retrieval diversification techniques for social media. IEEE Transactions on Multimedia **23**, 677–691 (2020)
22. Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems: Overview of the medical image retrieval task at ImageCLEF 2004–2014. Computerized Medical Imaging and Graphics **39**(0), 55 – 61 (2015)
23. Kozlovski, S., Dicente Cid, Y., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2022 - CT-based caverns detection and report. In: CLEF2022 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Bologna, Italy (September 5-8 2022)
24. Kozlovski, S., Liauchuk, V., Dicente Cid, Y., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2021 - CT-based tuberculosis type classification. In: CLEF 2021 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Bucharest, Romania (September 21-24 2021)
25. Kozlovski, S., Liauchuk, V., Dicente Cid, Y., Tarasau, A., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2020 - automatic CT-based report generation. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Thessaloniki, Greece (September 22-25 2020)
26. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J.R.R., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., Ferrari, V.: The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. CoRR **abs/1811.00982** (2018), `http://arxiv.org/abs/1811.00982`
27. Liauchuk, V., Kovalev, V.: Imageclef 2017: Supervoxels and co-occurrence for tuberculosis CT image classification. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
28. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V. Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer (2014). https://doi.org/10.1007/978-3-319-10602-1_48, `https://doi.org/10.1007/978-3-319-10602-1\_48`
29. Müller, H., Clough, P., Deselaers, T., Caputo, B. (eds.): ImageCLEF – Experimental Evaluation in Visual Information Retrieval, The Springer International Series On Information Retrieval, vol. 32. Springer, Berlin Heidelberg (2010)
30. Pelka, O., Ben Abacha, A., García Seco de Herrera, A., Jacutprakart, J., Friedrich, C.M., Müller, H.: Overview of the ImageCLEFmed 2021 concept & caption pre-

diction task. In: CLEF2021 Working Notes. pp. 1101–1112. CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania (September 21-24 2021)
31. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2019 concept prediction task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)
32. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2020 concept prediction task: Medical image understanding. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
33. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): a multimodal image dataset. In: Proceedings of the Third International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS 2018), Held in Conjunction with MICCAI 2018. vol. 11043, pp. 180–189. LNCS Lecture Notes in Computer Science, Springer, Granada, Spain (September 16 2018)
34. Popescu, A., Deshayes-Chossart, J., Schindler, H., Ionescu, B.: Overview of the imageclef 2022 aware task. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022), LNCS Lecture Notes in Computer Science, Springer, Bologna, Italy (September 5-8 2022)
35. Roberts, R.J.: Pubmed central: The genbank of the published literature. Proceedings of the National Academy of Sciences of the United States of America **98**(2), 381–382 (Jan 2001). https://doi.org/10.1073/pnas.98.2.381
36. Rückert, J., Ben Abacha, A., García Seco de Herrera, A., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Müller, H., Friedrich, C.M.: Overview of ImageCLEFmedical 2022 – Caption Prediction and Concept Detection. In: CLEF2022 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy (September 5-8 2022)
37. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y, `https://doi.org/10.1007/s11263-015-0816-y`
38. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.: YFCC100M: the new data in multimedia research. Commun. ACM **59**(2), 64–73 (2016)
39. Tsikrika, T., García Seco de Herrera, A., Müller, H.: Assessing the scholarly impact of ImageCLEF. In: CLEF 2011. pp. 95–106. Springer Lecture Notes in Computer Science (LNCS) (sep 2011)
40. Tsikrika, T., Larsen, B., Müller, H., Endrullis, S., Rahm, E.: The scholarly impact of CLEF (2000–2009). In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 1–12. Springer (2013)
41. World Health Organization, et al.: Global tuberculosis report 2019 (2019)
42. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. arXiv preprint arXiv:2201.02605 (2022)