Experiences from the ImageCLEF Medical Retrieval and Annotation Tasks

Henning Müller, Jayashree Kalpathy-Cramer and Alba G. Seco de Herrera

Abstract The medical tasks in ImageCLEF have been run every year from 2004-2018 and many different tasks and data sets have been used over these years. The created resources are being used by many researchers well beyond the actual evaluation campaigns and are allowing to compare the performance of many techniques on the same grounds and in a reproducible way. Many of the larger data sets are from the medical literature, as such images are easier to obtain and to share than clinical data, which was used in a few smaller ImageCLEF challenges that are specifically marked with the disease type and anatomic region. This chapter describes the main results of the various tasks over the years, including data, participants, types of tasks evaluated and also the lessons learned in organizing such tasks for the scientific community.

1 Introduction

ImageCLEF[1] started as the Cross–Language Image Retrieval Task in CLEF (Cross–Language Evaluation Forum[2]) in 2003 (Clough and Sanderson, 2004; Clough et al, 2010). A medical task was added in 2004 (Clough et al, 2005) and has been held every year since then (Kalpathy-Cramer et al, 2015). Several articles and books describe the overall evolution of the tasks and the various approaches that were used

Henning Müller
HES–SO Valais, Sierre, Switzerland, e-mail: henning.mueller@hevs.ch

Jayashree Kalpathy-Cramer
MGH Martinos Center for Biomedical Imaging, Charlestown, MA, USA, e-mail: Kalpathy@nmr.mgh.harvard.edu

Alba G. Seco de Herrera
University of Essex, UK, e-mail: alba.garcia@essex.ac.uk

[1] http://www.imageclef.org/

[2] http://www.clef-campaign.org/

to create the resources and compare the results in much detail (Kalpathy-Cramer et al, 2015; Müller et al, 2010a). Similar to other campaigns such as TREC (Text Retrieval Conference) (Rowe et al, 2010) or TRECvid (The video retrieval task of the Text Retrieval Conference) (Thornley et al, 2011), an important scholarly impact was shown for both ImageCLEF (Tsikrika et al, 2011) and also the overall CLEF campaign (Tsikrika et al, 2013; Angelini et al, 2014). As the impact increases almost exponentially over the years it can be expected that the impact has grown even stronger since these studies were published in 2011 and 2013, respectively. Particularly the resources on medical data have been used by a large number of researchers, as many technical research groups find it hard to access medical data sets if they do not have a close collaboration with medical partners. As Open Science is generally supported strongly by funding organizations and universities, there is a whole field building around making data, tasks and code available and sharing these resources with other researchers. Such Open Science can strongly increase the impact of research projects as well, when sharing data and software.

The data sets and tasks in ImageCLEF have evolved over the years with data sets becoming generally larger and tasks more challenging and complex. Some clinically relevant data sets remain relatively small but this is simply linked to data availability and confidentiality, and also to the cost of annotation. An overall goal of ImageCLEF has always been to create resources that allow for multimodal data access, so combining visual and textual information and possibly structured data. Another objective was to develop tasks that are based on solid grounds and allow for an evaluation in a realistic scenario (Müller et al, 2007). Log files of search systems have been used as well as example cases from teaching files (Müller et al, 2008b) to develop topics for retrieval system evaluation.

Scientific challenges were rare in the multimedia analysis or medical imaging field in the 1990s and 2000s compared to the information retrieval community, where they already started in the 1960s (Cleverdon et al, 1966; Jones and van Rijsbergen, 1975). In medical imaging, systematic benchmarking really started with a few conferences adding challenges in the late 2000s (Heimann et al, 2009) and slightly earlier with the ImageCLEF benchmark but only for visual medical information retrieval. Since around 2010, most major conferences in the field of image analysis and machine learning propose scientific challenges similar to workshops that have been part of conference programs for many years and that usually take one or two days at these conferences. These conference challenges have strongly influenced the field, as many examples show (Menze et al, 2015; Jimenez-del-Toro et al, 2015). Many large data sets and also software are now being shared (via platforms such as GitHub) and used by a large number of researchers to compare techniques on the same grounds.

More recent changes are linked to research infrastructures where an objective was to move the algorithms towards the data rather than the data to the algorithms (Hanbury et al, 2012). This has many advantages when dealing with very large data sets, confidential data, or sources that change and evolve quickly, when creating a fixed data collection is not practical. Several approaches have been presented for creating evaluation frameworks that allow the submission of source code, virtual machines or

Docker containers (Jimenez-del-Toro et al, 2016; Gollub et al, 2012). More generally, such approaches are grouped under the term Evaluation–as–a–Service (EaaS[3]) (Hanbury et al, 2015), and are really an integrated way to share data, source code and computational infrastructures for research. A previous chapter in this volume discusses EaaS in more details.

This chapter analyzes the work done in the ImageCLEF medical tasks from 2004 until 2018 showing how tasks and techniques have evolved. It also gives many links to further resources, as an extremely detailed analysis of the participating techniques is not possible in such a short book chapter. The many references give good starting points for a more detailed analysis. The data sets created in ImageCLEF are also usually used for many years beyond the ImageCLEF challenges and these articles need to be analysed to show the real advances in system performance over the years.

This chapter is organized as follows: Section 2 describes the ImageCLEF tasks, the data sets and the participation. An overview of the main techniques that achieved best results is given in the last part of the section. The main lessons learned are described in Section 3 and conclusions are given in Section 4.

2 Tasks, Data and Participation in the ImageCLEF Medical Tasks over the Years

This section describes the evolution of the tasks over the years, starting with the types of tasks proposed, the data types used, data size available and the participation in the task. A short discussion of the main techniques leading to best results is given.

2.1 Overview of the Medical Tasks Proposed

This section analyzes past data and resources created in the medical tasks of ImageCLEF that have been organized for 15 years. The analysis is based on the overview articles of these years (Clough et al, 2005, 2006; Müller et al, 2008a; Müller et al, 2009; Radhouani et al, 2009; Müller et al, 2010b; Kalpathy-Cramer et al, 2011; Müller et al, 2012; García Seco de Herrera et al, 2013, 2015, 2016; Dicente Cid et al, 2017; Eickhoff et al, 2017; Müller et al, 2006) and is summarized in Table 1.

It can be seen that the first years of ImageCLEF offered mainly general retrieval and then classification tasks. In 2010, a case–based retrieval task that is closer to clinical applications was proposed. In 2014, a first task related to a clinically-relevant set of diseases was introduced (annotation of liver CT images with semantic categories of lesions) and since 2017 a tuberculosis task is similarly related to a real clinical application and need (looking at tuberculosis type and drug resistances of the bacteria in the images alone). Many of the later tasks were much more complex

[3] `http://www.eaas.cc/`

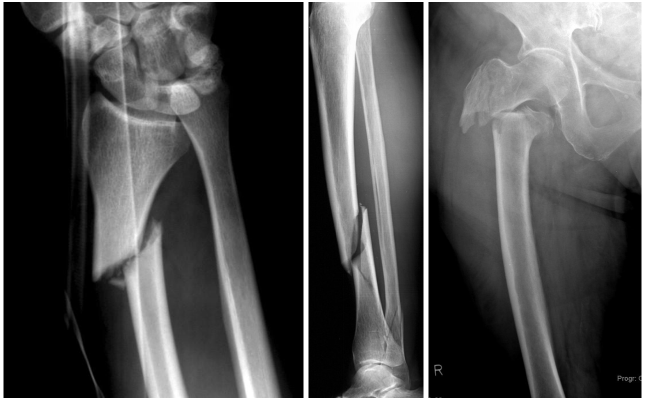
Table 1: Overview of the various tasks that have been performed over the years, ranging from general tasks in the beginning to some disease–oriented task later on that are marked as such.

Task type	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
image–based retrieval	x	x	x	x	x	x	x	x	x	x					
image type classification		x	x	x	x	x									
case–based retrieval							x	x	x	x					
image modality classification							x	x	x	x					
subfigure classification												x	x		
compound figure detection												x	x		
multi-label classification												x	x		
compound figure separation										x		x	x		
liver CT annotation											x	x			
caption prediction												x	x	x	x
tuberculosis classification														x	x
visual question answering															x

and required not only information retrieval competencies and features extraction from images but really targeted approaches towards extracting knowledge from the images. Research groups without a close link with health specialists often reported that it was challenging to estimate performance of their tools. A user analysis of retrieval based on images in the medical open access literature showed that research tasks are required that enrich meta data on images in the literature, as basically no information describing the images is available. The type of image (for example x-ray, CT, MRI, light microscopy image) can be used to filter images before visual image similarity retrieval is employed, as it can strongly focus the search and also use image type-dependent visual features. Such meta data in the images and also filtering are required to build retrieval applications based on the cleaned data. Compound figures are another challenge in the biomedical literature as many journal figures contain several subfigures with varying content and relationships among them because some journals limit the number of figures and this pushes authors to add more content into few figures. Such figures can have subfigures of different types and thus also have parts with the visual appearance of several sub-categories. With the exponential growth of the biomedical literature this can also be considered a priority area for the future, as images are available in almost unlimited quantities (growing exponentially) and getting ground truth is a main challenge. Crowdsourcing has been used for this (Foncubierta-Rodríguez and Müller, 2012) (see Section 2.3).

An example topic with a query in three languages and image examples for the retrieval task in 2005 is shown in Figure 1.

Images:



Query :

Show me all x–ray images showing fractures.
Zeige mir Röntgenbilder mit Brüchen.
Montres–moi des radiographies avec des fractures.

Fig. 1: A query requiring more than visual retrieval but visual features can provide hints to good results (taken from ImageCLEF 2005).

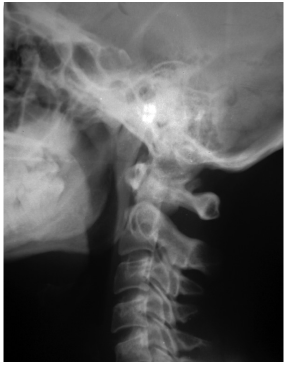
2.2 Data Sets and Constraints for Medical Data

One of the major challenges in medical data analysis is the availability of large-scale resources. Any medical data usage in health institutions needs to be confirmed by local ethics committees and usually requires a targeted application with a clinical application that cannot be modified without changing the ethics agreement. This often limits the size and availability of medical data and ethics committees may completely restrict sharing data, so analyses can only be executed locally on the data. Exceptions are medical teaching files that are created with ethics approval and also the biomedical literature that contains many images that were acquired with ethics approval and are then made available publicly. These two facts also drove the data sets in the medical ImageCLEF tasks. Table 2 shows an overview of the types of images and the number of images or cases that are available in each of the years of ImageCLEF.

Whereas for most tasks the data set size is the number of images, for the tuberculosis task this is the number of volumes. Each volume then consists of around 150–200 slices or images. This explains the seemingly small size, even though the complexity of the tasks has significantly grown with the 3D data set need to be analyzed.

Most data sets are from the biomedical literature because this can make sharing dtaa sets easier. Whereas the initial database of images from radiology journals was already filtered prior to using it and contained almost exclusively clinical images the images of PubMed Central (PMC) had a much larger variability. This variability can also be seen in Figure 2 that shows an example of an image from the biomedical literature with its caption. Further examples are given later in the text.

Image:



Caption :
Lateral view plain radiograph of the cervical spine shows a large ossified structure extending from the base of the skull anterolaterally and caudally to the hyoid bone.

Fig. 2: Example of an image and its caption from the PubMed Central dataset.

One problem with images from the biomedical literature is shown in Figure 3, which contains two compound figures and its parts that were automatically separated in this case. Compound figures are the majority of the content of PMC and their treatment thus has a massive impact on how the overall content of the biomedical literature can be exploited fully automatically. As subfigures can be of very different types the visual content is otherwise mixed and before attributing subfigures to a specific type they need to be separated.

2.3 Relevance Judgements and Ground Truthing

To develop a standard test bed for large and varied data sets, manually generated ground truth or relevance assessments (in the case of retrieval tasks) is basically always needed. Ground truth generation is costly, tedious and time–consuming. It is even more complex when specialists are required for tasks that can not be performed by the general public. Medical doctors are expensive and they often have no time for such ground truthing tasks. Sometimes, medical students can be used or other persons from health professions, and for simple and focused tasks crowdsourcing is a good option. For crowdsourcing, several relevance assessments are usually collected and used to eliminate incoherent results and to obtain a high quality (Clough

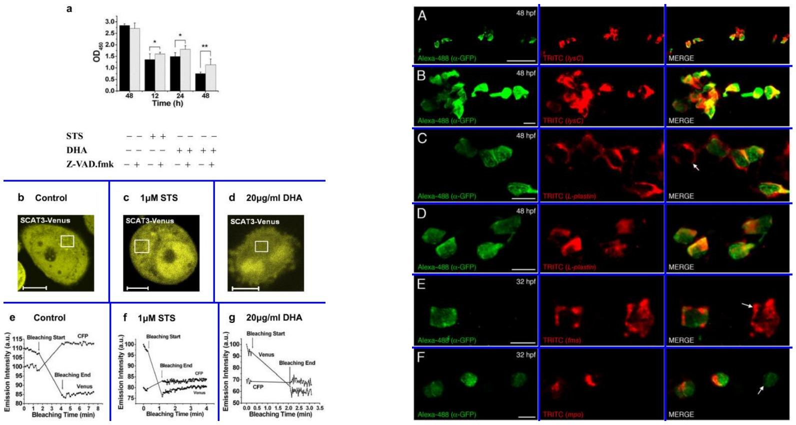


Fig. 3: Examples of successful automatic separation results of compound figures (blue lines separate the subfigures).

and Sanderson, 2013; García Seco de Herrera et al, 2014; García Seco de Herrera et al, 2016). Several assessors agreeing usually means that the results are fine but there also need to be strategies to combine several judgements where disagreement exists.

For retrieval tasks a full judgement of an entire collection is not possible and thus a pooling technique is frequently used (Jones and van Rijsbergen, 1975). Basically all image retrieval experiments in ImageCLEF on larger datasets use pooling, so the top *N* results of all participating runs are put together into a pool per topic and only these documents are judged for relevance. For classification usually the entire collection is classified manually and thus the data sets are often smaller than for the retrieval tasks. With sufficient training of very specific tasks also non–medical staff can be used for the classifications or relevance assessments, so crowdsourcing with quality control is also possible.

For ImageCLEF, the ground truth was in the first years generated by medical doctors, also because the collections were much smaller (500 images in 2004). Then, health science students could be hired, of which many were physians. This was only possible thanks to funding that was avaialble via related research projects. Limited funding was then used for crowdsourcing. In the past few years tasks based on data from the litereature were created where no manual ground truthing was required (for example the caption prediction task) or data sets were obtained where the ground truth was already available (as in the tuberculosis task). Sometimes also a combination of approaches was used, partly with manual judgements and partly with crowdsourcing. More details can be found in the overview papers of the respective tasks.

2.4 Evaluation Measures

The relevance assessments or ground truth are used to quantify system effectiveness (Clough and Sanderson, 2013). Many evaluation measures can be used to assess performance of retrieval or classification tasks based on the number of relevant documents (Buyya and Venugopal, 2005). The trec_eval[4] package is used as a standard tool for text retrieval and it extracts all relevant measures of ImageCLEF and most other benchmarks. Usually early precision and MAP (Mean Average Precision) are used as lead measures. Sometimes BPref (Binary Preference) is added as a measure that takes into account documents that were not judged in the pooling process.

Accuracy is most commonly used to assess classification tasks. When the class distribution is very unbalanced there are also several other measures that are important, for example the geometric mean of the performance on all classes. This highlights a good performance on all classes and not a concentration on good results for a few majority classes, which would favor a good accuracy in this case. For medical tasks, specificity (true negative rate, 1- false positive rate) and sensitivity (true positive rate or recall) are also very frequently used measures. These two measures allow to dscriminate between whether it is important catch all patients with a condition or whether it is more important to limit false positives. Each community thus has its own measures and it is always important to show several measures to analyse different aspects of the performance of participating systems.

However, assessing tasks such as compound figure separation is challenging. In this case a new evaluation approach was developed. The evaluation required to have a minimum overlap for the subfigure division between the ground truth and the data supplied by the groups in their runs (García Seco de Herrera et al, 2013). This allowed for some margin in terms of the separating lines, which is important as there is not one single optimal solution and the judges doing the ground truthing had an important amount of subjectivity.

In general, it is important to have more than one performance measure and ranking to really evaluate several aspects of the participating techniques and to not concentrate all techniques into optimizing a single measure.

2.5 Participants and Submissions

In Table 3 the number of groups that registered for a task and the number of groups that finally submitted results are listed. For some of the years the exact registration numbers were not mentioned in the overview papers and thus we cannot reproduce them anymore. Thus, we used square brackets for these and used the number of submissions as a lower bound of the participation.

[4] `http://trec.nist.gov/trec_eval/`

There has been a general increase in the participation over the years, but many new tasks take one or two years to obtain higher numbers because researchers need to adapt to specific tasks in their research projects. The number of submissions on the other hand has been lower in recent years where many new and more complex tasks were introduced that go beyond simple text retrieval or image classification.

Figure 4 shows the evolution of research groups that registered for the ImageCLEF medical tasks on a per task basis and in the second graphic those groups that submitted results. We can see that the long running tasks had a large number of actual submissions whereas the more recent tasks that have only been organized for 1-2 years have relatively few submissions. The number of registrations actually had some peaks in recent years and it seems to increase over the years in a relatively stable fashion. On the other hand, the percentage of registered users actually submitting results has decreased over this period. Possibly, this can be attributed to a larger availability of benchmarks and data sets for researchers to choose from.

2.6 Techniques used

Whereas first techniques applied in ImageCLEF used mainly simple texture (Gabor filters, Tamura, Co–occurence matrices) and color (color histograms) features extracted from the images in combination with often simple distance measures such as k–nearest neighbors (k–NN), there were also first tests with combinations of text retrieval and visual retrieval techniques (Müller et al, 2005). In general, techniques can clearly be separated into text retrieval and visual analysis techniques, where text retrieval usually led to much better results for the retrieval tasks, whereas in classification tasks often the visual results were better. Best results in the first years (2004–2007) were often obtained using simple feature modelling techniques similar to visual words (Deselaers et al, 2005) or Fisher vectors based on patches in the images and not the global image content alone. These techniques had very good results for several years until more elaborate machine learning approaches such as support vector machines (SVMs) really improved outcomes for all classification tasks (Tommasi et al, 2010). Details of all techniques are impossible to be described here. Often similar techniques led in some cases to very good results and in other cases to poor results depending on how well the techniques were really optimized.

Feature fusion remained another area where many approaches were tested (Depeursinge and Müller, 2010). Often rank–based fusion led to better results than score–based fusion with text retrieval and image retrieval following very different distributions in terms of absolute similarity scores. Both early and late fusion sometimes led to best results, so this might really depend on the exact data and application scenario. Another major advance in terms of techniques was the use of Fisher vectors (Clinchant et al, 2010) that led to best results in several competitions.

In the past three years most successful techniques use deep learning approaches (Koitka and Friedrich, 2016; Stefan et al, 2017) for most tasks. This holds true for almost all classification challenges but also more complex scenarios such as compound

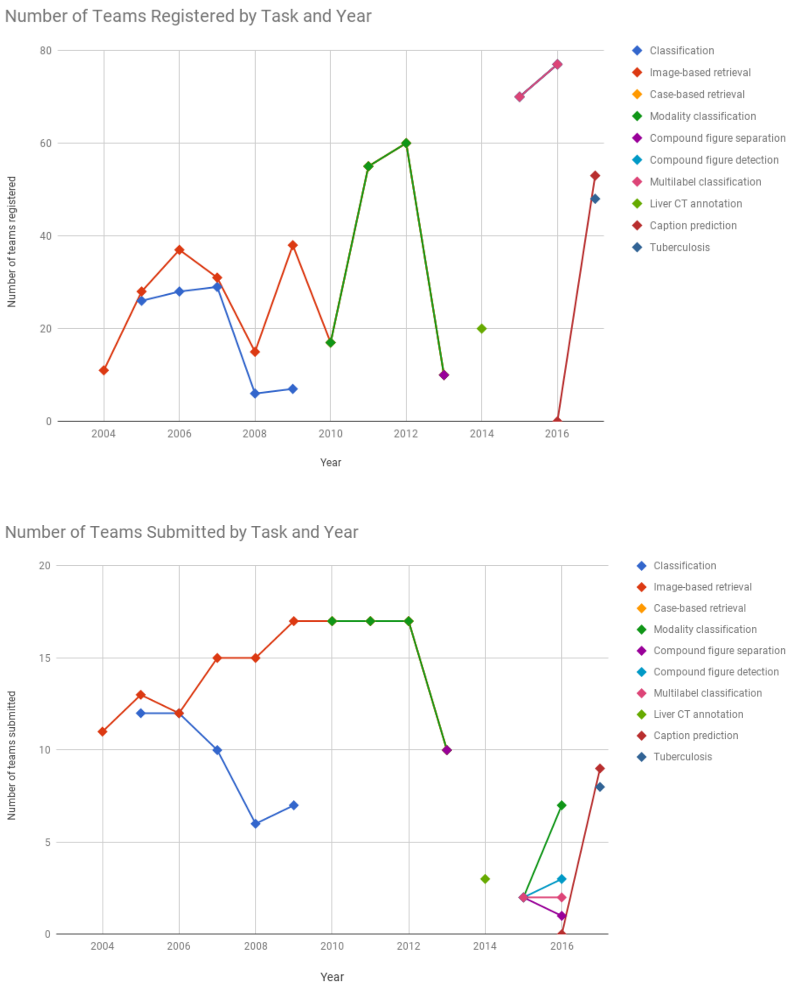


Fig. 4: Number of groups registered for ImageCLEF per task (figure at the top) and the number of groups that actually submitted runs (figure at the bottom) over all the years.

figure separation. Extraction of features from deep learning with classical classification techniques were also tested with success. There are several rather specific techniques that led to best results in focused tasks such as the tuberculosis task in 2017 (Dicente Cid et al, 2017). Here, a graph model was used that obtained best results in prediction multiple drug resistances. This can be attributed to the modeling of known knowledge on lung anatomy and distribution of disease, which would require a very large number of cases to learn the model with deep learning. Using more handcrafted features can model this existing knowledge.

3 Lessons Learned and Mistakes to Avoid

In (Müller et al, 2007) an early summary already gave several important lessons learned from running the first four years of the medical tasks in ImageCLEF. Since then, many things have changed with scientific challenges really becoming a standard tool in medical imaging and computer vision. Particularly the diversity of the medical tasks in ImageCLEF has increased massively over the years.

The main success factor for any scientific challenge is really to *create a community* around the task and engage participants in the entire process. This creates a positive energy and attracts other participants and particularly motivates to pursue and submit results in the end. Strong participation by peers also increases the number of groups actually submitting results. This number is often small and in the range of 20-30% of the groups that initially registered. It ensures that a task is not only run a single year but several years in a row. Tsikrika et al. (Tsikrika et al, 2011) show that for most of the tasks, there is a peak in terms of scholarly impact in their second or third year of operation, then followed by a slow stalling or even decline in impact if the tasks are not changed substantially. Running the same task for several years can lead to continuous improvement of the participating approaches. An important aspect is also to keep a continuous test set over the years to also measure absolute improvements of the techniques over time but this is often more difficult to realize.

Another important part that is linked to the community aspects is the general *communication* with participants. This is essential to keep participants or interested researchers updated on all details and the status of the competition at all times. The main entry point for all information in ImageCLEF is the web page that is regularly updated and contains all information on the tasks with details on data, task creation and performance measures, also of previous years. Results of the challenge are also published here. A registration system manages data access that requires the signature of an end user agreement. The registration system also allows to upload runs and all runs are automatically checked to be in the right format and only contain valid identifiers. This strongly reduces the work of organizers to check the submitted runs for mistakes, which was a common problem in the first years of ImageCLEF. A mailing list with all registered participants makes it possible to address all participants with targeted information, for example of deadline changes. As past participants can remain on the list this is also a prime means for announcing new tasks or task ideas that can be discussed with researchers. In recent years the communication strategy increasingly includes social media. ImageCLEF has a Facebook page[5] and a Twitter account[6] and these are also used to address participants. Part of this may be redundant but it makes it possible to reach all participants via a variety of channels. LinkedIn has also been used in recent years to advertise the tasks and broaden the participant base via focused groups in the area. In 2018, a new registration system based on the open source tool crowdAI[7] was implemented. This tool gives new possibilities, for example to not only have a final workshop where results are compared but a continuous leader board that is active also after the competition finishes and where groups can upload and compare their results in a continuous way. The use of EaaS approaches with code submission is also possible with such an infrastructure but currently not used by us.

Having a common *publication* that describes the data set, the creation of ground truth and that compares the results of all submitted results is another aspect that is important for reproducibility of the results and also for keeping the data accessible long term and having it used in a clear evaluation context. For this it is essential to have a description of the runs of the participants, so not only performance measures can be compared but also the techniques that lead to a specific performance. In the past it was often the case that best and worst approaches were using almost the same techniques but that small modifications had important effects on the outcome and for this reason a formal description of all techniques is essential.

Linked to the publications and an analysis of the results is the organization of a common *workshop*. At the workshop, participants can present the most interesting results of each task and can then compare the approaches and outcomes to find better ways to improve results in the future. This can foster collaborations between participants even if in the past only a few collaborations between research groups have evolved from the discussions in the meeting. The workshop has open discussion sections each year to plan tasks and also evaluate procedures for the future and thus integrate feedback into improving the tasks. This is linked to a community feeling among participants and can clearly improve motivation if handled well. It is important to transparently discuss all details, so the rankings are based on solid grounds.

To tackle *current research challenges* is also important, as universities which are the main participants of the tasks all depend on funding and this is usually assigned based on calls for topics that are currently hot research topics. If topics really are novel then a PhD student can for example engage in several years of work on such challenges in a efficient way, where they can compare results to others and rely on the same setting and data. Usually, challenges get harder each year, so the full potential of the techniques can really be tested over time.

No collection or setup for an evaluation campaign is free of errors and thus it is essential to have structures and manpower to *fix errors* and mistakes in the data and the evaluations quickly, as soon as participants report them. This creates confidence in the evaluation campaign and makes sure that meaningful results can be

[5] `https://www.facebook.com/profile.php?id=106932209431744`

[6] `https://twitter.com/imageclef?lang=en`

[7] `http.//www.crowdai.org/`

obtained in the end. The capacity to fix errors and run a professional campaign is also linked with obtaining good *funding* for such challenges. Most often, only research funding is available and infrastructures that create data sources, maintaining basic services for benchmarks and a physical infrastructure are harder to fund, even though scientific impact in terms of citations can be higher for data papers than for technical papers, as many researchers base their work on this. Without funding, a certain professionalism can be lost as all organizers engage in their free time as volunteers. With respect to ground truthing, whether manual annotations of the data or relevance judgments, it is important to have funding, even when relatively cheap options such as crowdsourcing are used.

An objective of ImageCLEF has always been to be *complementary* with other evaluation tasks, in other conferences (for example TRECvid) or also inside the CLEF labs, such as LifeCLEF and CLEFeHealth. Such a complementarity ensures a clear positioning of the tasks and thus also a good participation. There have also been suggestions to organize ImageCLEF with existing conferences in computer vision or machine learning, as most tasks at CLEF have been focusing rather on text analysis and retrieval. We have had collaborations with other conferences in the past but feel that CLEF is a good forum for multimodal interdisciplinary research.

4 Discussion and Conclusions

The chapter gives an overview of 15 years of scientific challenges in the medical ImageCLEF tasks. It is clear that no extremely detailed analysis can be given on all lessons learned and results obtained for 15 years in only a few pages. This text mainly focuses on overviews of how the data, the tasks and the techniques evolved over the years. Then, we highlight the lessons learned and several success factors that were identified via discussions among the organizers and also with participants.

With Open Science now gaining momentum in almost all fields related to data science many challenges have been organized at conferences and workshops. Many of the challenges are similar in nature or in the data used. With an increasing number there can be fewer participants in every single challenge, which reduces the impact of every single challenge. Professional platforms such as Kaggle[8] have also changed the field of scientific challenges but leading researchers to commercial challenges, where price money is available instead of publications at purely scientific challenges. The targets are in this case very different, so not so much on understanding the techniques but really on tuning existing techniques. There is clearly a large market for data science challenges and such complementary approaches will likely coexist in the future.

Whereas professional challenges with prize money often do not focus on documenting techniques of the runs submitted in detail and understanding the actual techniques they push towards optimal performance. Scientific challenges, some-

[8] `http://www.kaggle.com/`

times also called coopetitions (in between a competition and a cooperation), on the other hand aim at reproducible science that documents all experiments that were run and also concentrates on the interestingness of approaches and algorithms and not only pure performance. We feel that this contributes to better understanding techniques and having a long term optimization of approaches. Cheating in such scientific challenges seems less likely than when prize money is involved, even though it still needs to be checked that results are compared in a fair way.

There are clearly many next steps that can be taken for scientific challenges. It is important to keep a workshop where participants meet but also keeping past challenges and data open for new submissions is important, so best results can be tracked and compared over a longer period of time. Fostering more collaboration is one of our important objectives that has not been easy to reach. Maybe components based, for example, on Docker containers can be used in automatic work flows and help to make component sharing easier among researchers. With machine learning going increasingly towards deep learning it also becomes possible to explore large data sets with various levels of annotations, so for example, high level manual annotations but also noisy automatic annotations that could augment the training data, for example with silver corpuses (Krenn et al, 2016).

Acknowledgements We would like to thank the various funding organizations that have helped make ImageCLEF a reality (EU FP6 & FP7, SNF, RCSO, Google and others) and also all the volunteer researchers who helped organize the tasks. Another big thank you goes to the data providers that assured that medical data could be shared with the participants. A final thanks to all participants who work on the tasks and provide us with techniques to compare and with lively discussions at the ImageCLEF workshops.

References

Angelini M, Ferro N, Larsen B, Müller H, Santucci G, Silvello G, Tsikrika T (2014) Measuring and analyzing the scholarly impact of experimental evaluation initiatives. In: Italian Research Conference on Digital Libraries

Buyya R, Venugopal S (2005) A gentle introduction to grid computing and technologies. CSI Communications 29(1):9–19

Cleverdon C, Mills J, Keen M (1966) Factors determining the performance of indexing systems. Tech. rep., ASLIB Cranfield Research Project, Cranfield

Clinchant S, Csurka G, Ah-Pine J, Jacquet G, Perronnin F, Sánchez J, Minoukadeh K (2010) Xrce's participation in wikipedia retrieval, medical image modality classification and ad–hoc retrieval tasks of imageclef 2010. In: Working Notes of the 2010 CLEF Workshop

Clough P, Sanderson M (2004) The CLEF 2003 cross language image retrieval task. In: Proceedings of the Cross Language Evaluation Forum (CLEF 2003)

Clough P, Sanderson M (2013) Evaluating the performance of information retrieval systems using test collections. Information Research 18(2)

Clough P, Müller H, Sanderson M (2005) The CLEF 2004 cross–language image retrieval track. In: Peters C, Clough P, Gonzalo J, Jones GJF, Kluck M, Magnini B (eds) Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign, Springer, Bath, UK, Lecture Notes in Computer Science (LNCS), vol 3491, pp 597–613

Clough P, Müller H, Deselaers T, Grubinger M, Lehmann TM, Jensen J, Hersh W (2006) The CLEF 2005 cross–language image retrieval track. In: Cross Language Evaluation Forum (CLEF 2005), Springer, Lecture Notes in Computer Science (LNCS), pp 535–557

Clough P, Müller H, Sanderson M (2010) Seven Years of Image Retrieval Evaluation, Springer, pp 3–18

Depeursinge A, Müller H (2010) Fusion techniques for combining textual and visual information retrieval. In: Müller H, Clough P, Deselaers T, Caputo B (eds) ImageCLEF, The Springer International Series On Information Retrieval, vol 32, Springer Berlin Heidelberg, pp 95–114

Deselaers T, Weyand T, Keysers D, Macherey W, Ney H (2005) FIRE in ImageCLEF 2005: Combining content–based image retrieval with textual information retrieval. In: Working Notes of the CLEF Workshop, Vienna, Austria

Dicente Cid Y, Batmanghelich K, Müller H (2017) Textured graph-model of the lungs for tuberculosis type classification and drug resistance prediction: participation in ImageCLEF 2017. In: CLEF2017 Working Notes, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland, CEUR Workshop Proceedings

Dicente Cid Y, Kalinovsky A, Liauchuk V, Kovalev V, , Müller H (2017) Overview of ImageCLEFtuberculosis 2017 - predicting tuberculosis type and drug resistances. In: CLEF 2017 Labs Working Notes, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland, CEUR Workshop Proceedings

Eickhoff C, Schwall I, García Seco de Herrera A, Müller H (2017) Overview of ImageCLEFcaption 2017 - the image caption prediction and concept extraction tasks to understand biomedical images. In: CLEF2017 Working Notes, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland, CEUR Workshop Proceedings

Foncubierta-Rodríguez A, Müller H (2012) Ground Truth Generation in Medical Imaging: A Crowdsourcing Based Iterative Approach. In: Workshop on Crowdsourcing for Multimedia

Gollub T, Stein B, Burrows S, Hoppe D (2012) Tira: Configuring, executing, and disseminating information retrieval experiments. In: Database and expert systems applications (DEXA), 2012 23rd international workshop on, IEEE, pp 151–155

Hanbury A, Müller H, Langs G, Weber MA, Menze BH, Fernandez TS (2012) Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis. In: CLEF conference, Springer Lecture Notes in Computer Science

Hanbury A, Müller H, Balog K, Brodt T, Cormack GV, Eggel I, Gollub T, Hopfgartner F, Kalpathy-Cramer J, Kando N, Krithara A, Lin J, Mercer S, Potthast M (2015) Evaluation–as–a–service: Overview and outlook. ArXiv 1512.07454

Heimann T, Van Ginneken B, Styner M, Arzhaeva Y, Aurich V, Bauer C, Beck A, Becker C, Beichel R, Bekes G, et al (2009) Comparison and evaluation of methods for liver segmentation from CT datasets. Medical Imaging, IEEE Transactions on 28(8):1251–1265

García Seco de Herrera A, Kalpathy-Cramer J, Demner Fushman D, Antani S, Müller H (2013) Overview of the ImageCLEF 2013 medical tasks. In: Working Notes of CLEF 2013 (Cross Language Evaluation Forum)

García Seco de Herrera A, Foncubierta-Rodríguez A, Markonis D, Schaer R, Müller H (2014) Crowdsourcing for Medical Image Classification. In: Annual Congress SGMI 2014

García Seco de Herrera A, Müller H, Bromuri S (2015) Overview of the ImageCLEF 2015 medical classification task. In: Working Notes of CLEF 2015 (Cross Language Evaluation Forum)

García Seco de Herrera A, Schaer R, Antani S, Müller H (2016) Using crowdsourcing for multi-label biomedical compound figure annotation. In: MICCAI workshop Labels, Springer, Lecture Notes in Computer Science

García Seco de Herrera A, Schaer R, Bromuri S, Müller H (2016) Overview of the ImageCLEF 2016 medical task. In: Working Notes of CLEF 2016 (Cross Language Evaluation Forum)

Jimenez-del-Toro O, Hanbury A, Langs G, Foncubierta-Rodríguez A, Müller H (2015) Overview of the VISCERAL Retrieval Benchmark 2015. In: Multimodal Retrieval in the Medical Domain: First International Workshop, MRMD 2015, Vienna, Austria, March 29, 2015, Revised Selected Papers, Springer, Lecture Notes in Computer Science, vol 9059, pp 115–123

Jimenez-del-Toro O, Müller H, Krenn M, Gruenberg K, Taha AA, Winterstein M, Eggel I, Foncubierta-Rodríguez A, Goksel O, Jakab A, Kontokotsios G, Langs G, Menze B, Salas Fernandez T, Schaer R, Walleyo A, Weber MA, Dicente Cid Y, Gass T, Heinrich M, Jia F, Kahl F, Kechichian R, Mai D, Spanier AB, Vincent G, Wang C, Wyeth D, Hanbury A (2016) Cloud–based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL Anatomy Benchmarks. IEEE Transactions on Medical Imaging 35(11):2459–2475

Jones KS, van Rijsbergen C (1975) Report on the need for and provision of an ideal information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge

Kalpathy-Cramer J, Müller H, Bedrick S, Eggel I, García Seco de Herrera A, Tsikrika T (2011) The CLEF 2011 medical image retrieval and classification tasks. In: Working Notes of CLEF 2011 (Cross Language Evaluation Forum)

Kalpathy-Cramer J, García Seco de Herrera A, Demner-Fushman D, Antani S, Bedrick S, Müller H (2015) 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

Koitka S, Friedrich CM (2016) Traditional feature engineering and deep learning approaches at medical classification task of ImageCLEF 2016. In: CLEF2016 Working Notes, CEUR-WS.org, Évora, Portugal, CEUR Workshop Proceedings

Krenn M, Dorfer M, Jimenez-del-Toro O, Müller H, Menze B, Weber MA, Hanbury A, Langs G (2016) Creating a Large–Scale Silver Corpus from Multiple Algorithmic Segmentations, Springer International Publishing, pp 103–115

Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, Lanczi L, Gerstner E, Weber MA, Arbel T, Avants BB, Ayache N, Buendia P, Collins DL, Cordier N, Corso JJ, Criminisi A, Das T, Delingette H, Demiralp C, Durst CR, Dojat M, Doyle S, Festa J, Forbes F, Geremia E, Glocker B, Golland P, Guo X, Hamamci A, Iftekharuddin KM, Jena R, John NM, Konukoglu E, Lashkari D, Mariz JA, Meier R, Pereira S, Precup D, Price SJ, Raviv TR, Reza SMS, Ryan M, Sarikaya D, Schwartz L, Shin HC, Shotton J, Silva CA, Sousa N, Subbanna NK, Szekely G, Taylor TJ, Thomas OM, Tustison NJ, Unal G, Vasseur F, Wintermark M, Ye DH, Zhao L, Zhao B, Zikic D, Prastawa M, Reyes M, Van Leemput K (2015) The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). Medical Imaging, IEEE Transactions on 34(10):1993–2024

Müller H, Geissbuhler A, Ruch P (2005) ImageCLEF 2004: Combining image and multi–lingual search for medical image retrieval. In: Peters C, Clough P, Gonzalo J, Jones GJF, Kluck M, Magnini B (eds) Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign, Springer, Bath, UK, Lecture Notes in Computer Science (LNCS), vol 3491, pp 718–727

Müller H, Deselaers T, Lehmann T, Clough P, Kim E, Hersh W (2006) Overview of the ImageCLEFmed 2006 Medical Retrieval and Annotation Tasks. In: Nardi A, Peters C, Vicedo JL, Ferro N (eds) CLEF 2006 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, `http://ceur-ws.org/Vol-1172/`

Müller H, Boyer C, Gaudinat A, Hersh W, Geissbuhler A (2007) Analyzing web log files of the Health On the Net HONmedia search engine to define typical image search tasks for image retrieval evaluation. In: MedInfo 2007, Brisbane, Australia, IOS press, Studies in Health Technology and Informatics, vol 12, pp 1319–1323

Müller H, Deselaers T, Grubinger M, Clough P, Hanbury A, Hersh W (2007) Problems with running a successful multimedia retrieval benchmark. In: MUSCLE/ImageCLEF workshop 2007, Budapest, Hungary, pp 9–18

Müller H, Deselaers T, Kim E, Kalpathy-Cramer J, Deserno TM, Clough P, Hersh W (2008a) Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: CLEF 2007 Proceedings, Springer, Budapest, Hungary, Lecture Notes in Computer Science (LNCS), vol 5152, pp 473–491

Müller H, Kalpathy-Cramer J, Hersh W, Geissbuhler A (2008b) Using Medline queries to generate image retrieval tasks for benchmarking. In: Medical Informatics Europe (MIE2008), IOS press, Gothenburg, Sweden, pp 523–528

Müller H, Kalpathy-Cramer J, Kahn Jr CE, Hatt W, Bedrick S, Hersh W (2009) Overview of the ImageCLEFmed 2008 medical image retrieval task. In: Peters C, Giampiccolo D, Ferro N, Petras V, Gonzalo J, Peñas A, Deselaers T, Mandl T, Jones G, Kurimo M (eds) Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum, Aarhus, Denmark, Lecture Notes in Computer Science (LNCS), vol 5706, pp 500–510

Müller H, Clough P, Deselaers T, Caputo B (eds) (2010a) ImageCLEF – Experimental Evaluation in Visual Information Retrieval, The Springer International Series On Information Retrieval, vol 32. Springer, Berlin Heidelberg

Müller H, Kalpathy-Cramer J, Eggel I, Bedrick S, Reisetter J, Kahn Jr CE, Hersh W (2010b) Overview of the CLEF 2010 medical image retrieval track. In: Working Notes of CLEF 2010 (Cross Language Evaluation Forum)

Müller H, García Seco de Herrera A, Kalpathy-Cramer J, Demner Fushman D, Antani S, Eggel I (2012) Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In: Working Notes of CLEF 2012 (Cross Language Evaluation Forum)

Radhouani S, Kalpathy-Cramer J, Bedrick S, Bakke B, Hersh W (2009) Multimodal medical image retrieval improving precision at ImageCLEF 2009. In: Working Notes of the 2009 CLEF Workshop, Corfu, Greece

Rowe BR, Wood DW, Link AN, Simoni DA (2010) Economic impact assessment of NIST text retrieval conference (TREC) program. Technical report project number 0211875, National Institute of Standards and Technology

Stefan LD, Ionescu B, Müller H (2017) Generating captions for medical images with a deep learning multi-hypothesis approach: ImageCLEF 2017 caption task. In: CLEF2017 Working Notes, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland, CEUR Workshop Proceedings

Thornley CV, Johnson AC, Smeaton AF, Lee H (2011) The scholarly impact of TRECVid (2003–2009). Journal of the American Society for Information Science and Technology 62(4):613–627

Tommasi T, Caputo B, Welter P, Güld M, Deserno TM (2010) Overview of the clef 2009 medical image annotation track. In: Peters C, Caputo B, Gonzalo J, Jones G, Kalpathy-Cramer J, Müller H, Tsikrika T (eds) Multilingual Information Access Evaluation II. Multimedia Experiments, Lecture Notes in Computer Science, vol 6242, Springer Berlin / Heidelberg, pp 85–93

Tsikrika T, García Seco de Herrera A, Müller H (2011) Assessing the scholarly impact of ImageCLEF. In: CLEF 2011, Springer Lecture Notes in Computer Science (LNCS), pp 95–106

Tsikrika T, Larsen B, Müller H, Endrullis S, Rahm E (2013) The scholarly impact of CLEF (2000–2009). In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Springer, pp 1–12

Year	task type	resource	no images
2004	image retrieval	teaching files, CasImage	8,725
2005	image retrieval	CasImage, PEIR, MIR, PathoPic	50,000
	annotation	Radiographies of IRMA	9,000
2006	retrieval	CasImage, PEIR, MIR, PathoPic	50,000
	annotation	Radiographics of IRMA	11,000
2007	retrieval	myPACS, CORI added	66,636
	annotation	Radiograhies of IRMA	12,000
2008	retrieval	RSNA	66,000
	annotation	Radiograhies of IRMA	12,076
2009	retrieval	RSNA	74,902
	annotation	Radiograhies of IRMA	12,677
2010	retrieval	RSNA	77,506
	case retrieval	RSNA	77,506
	classification	RSNA modality classification	5,010
2011	image retrieval	PMC subset 1	231,000
	case retrieval	PMC subset 1	231,000
	classification	PMC subset 1 modality class.	2,000
2012	image retrieval	Pmc subset 2	300,000
	case retrieval	Pmc subset 2	300,000
	classification	PMC subset 2 modality class.	2,000
2013	image retrieval	PMC subset 2	300,000
	case retrieval	PMC subset 2	300,000
	classification	PMC subset 2 modality class.	5,483
	compound figure separation	PMC	2,967
2014	annotation	Liver CT annotation dataset	60
2015	compound figure detection	PMC subset 3	20,867
	compound figure separation	PMC subset 3 figure separation	6,784
	multi-label	PMC subset 3 multi-label classification	1,568
	classification	PMC subset 3 subfigure classification	6,776
	clustering	Medical Clustering	5,000
	annotation	Liver CT annotation dataset	60
2016	compound figure detection	PMC subset 4	24,456
	compound figure separation	PMC subset 4 figure sep.	8,397
	multi-label	PMC subset 4 multi-label classification	2,651
	classification	PMC subset 4 subfigure classification	10,942
	caption prediction	PMC subset caption prediction 1	20,000
2017	caption prediction	PMC subset caption prediction 2	184,614
	concept detection	PMC subset caption prediction 2	184,614
	classification	Tuberculosis dataset - MDR	444
	resistance detection	Tuberculosis dataset - TBT	801
2018	caption prediction	PMC subset caption prediction 3	232,305
	concept detection	PMC subset caption prediction 3	232,305
	classification	Tuberculosis dataset - MDR	1,513
	resistance detection	Tuberculosis dataset - TBT	495
	severity scoring	Tuberculosis dataset - SVR	279
	visual question answering	PMC subset VQA	2,866

Table 2: Overview of the data sets that were created over the years for the various tasks.

Year	task	registered	submitted
2004	image-based retrieval	[11]	11
2005	image-based retrieval	28	13
	classification	26	12
2006	image-based retrieval	37	12
	classification	28	12
2007	image-based retrieval	31	15
	classification	29	10
2008	image-based retrieval	[15]	15
	classification	[6]	6
2009	image-based retrieval	38	17
	classification	[7]	7
2010	image-based retrieval	[17]	17
	case-based retrieval	[17]	17
	modality classification	[17]	17
2011	image–based retrieval	55	17
	case–based retrieval	55	17
	modality classification	55	17
2012	image–based retrieval	60	17
	case–based retrieval	60	17
	modality classification	60	17
2013	image–based retrieval	[10]	10
	case–based retrieval	[10]	10
	modality classification	[10]	10
	compound figure separation	[10]	10
2014	liver CT annotation	20	3
2015	modality classification	70	2
	compound figure separation	70	2
	compound figure detection	70	2
	multi–label classification	70	2
	liver annotation	51	1
2016	modality classification	77	7
	compound figure separation	77	1
	compound figure detection	77	3
	multi–label classification	77	2
	caption prediction	[0]*	0
2017	caption prediction	53	9
	tuberculosis	48	8
2018	caption prediction	46	8
	tuberculosis	33	11
	visual question answering 28	48	5

Table 3: Overview of the participation in the tasks over the years, ”[]” denotes years when the exact numbers of registered users are not known (only the number of those submitting results) and ”*” highlights a task where in the end no group submitted results, which in combination with ”[]” means that nothing concrete can be said about participation. The participants list can also include the task organizer if the team registered.