[Citation needed] Data usage and citation practices in medical imaging conferences

Théo Sourget[1,2] TSOU@ITU.DK
Ahmet Akkoç[1,3]
Stinna Winther[1]
Christine Lyngbye Galsgaard[1]
Amelia Jiménez-Sánchez[1]
Dovile Juodelyte[1]
Caroline Petitjean[2]
Veronika Cheplygina[1] VECH@ITU.DK
[1] *IT University of Copenhagen, Denmark*
[2] *University of Rouen, France*
[3] *ZiteLab ApS, Denmark*

Abstract

Medical imaging papers often focus on methodology, but the quality of the algorithms and the validity of the conclusions are highly dependent on the datasets used. As creating datasets requires a lot of effort, researchers often use publicly available datasets, there is however no adopted standard for citing the datasets used in scientific papers, leading to difficulty in tracking dataset usage. In this work, we present two open-source tools we created that could help with the detection of dataset usage, a pipeline[1] using OpenAlex and full-text analysis, and a PDF annotation software[2] used in our study to manually label the presence of datasets. We applied both tools on a study of the usage of 20 publicly available medical datasets in papers from MICCAI and MIDL. We compute the proportion and the evolution between 2013 and 2023 of 3 types of presence in a paper: cited, mentioned in the full text, cited and mentioned. Our findings demonstrate the concentration of the usage of a limited set of datasets. We also highlight different citing practices, making the automation of tracking difficult.

Keywords: Bibliometrics, citations, datasets, medical imaging, data re-use, annotation

1. Introduction

While the increased usage of open data is a positive development, we hypothesize it might introduce a shift in the targeted applications. For example, (Varoquaux and Cheplygina, 2022) show that since the Kaggle lung cancer challenge in early 2017 (Buckeye et al., 2017), there has been a disproportionate increase in machine learning papers on lung cancer, while many of the proposed solutions do not differ in practice. A similar concentration on fewer datasets has also been found in machine learning (Koch et al., 2021). Another medical imaging example is the segmentation of cardiac ventricles, addressed with multiple competitions (Bernard et al., 2018b; Suinesiaputra et al., 2012; Petitjean et al., 2015; Campello et al.,

1. https://github.com/TheoSourget/Public_Medical_Datasets_References
2. https://github.com/TheoSourget/pdf_annotator

2021). The latest competition achieved highly accurate results and commercially available software exists (Wu et al., 2024), yet the application still remains popular for evaluating novel algorithms. Moreover, while the availability of a public dataset is a positive step towards getting a problem addressed by the community, the choice of a single dataset for evaluation also results in an overestimation of performances leading to a gap when applied on a different one (Wu et al., 2021).

There is a need to analyse research within a field to understand the progress being made, but next to surveys focused either on methods (Litjens et al., 2017; Cheplygina et al., 2019; Budd et al., 2021) or on datasets within a specific application (Daneshjou et al., 2021; Wen et al., 2022), we find few studies on understanding *dataset use* beyond their initial release in the field. We believe this is in part due to identifying dataset usage, as datasets may be used without corresponding citations, and vice versa. Our contributions, aiming to achieve this identification of dataset usage are as follows: **(1)** We present a fully automated pipeline for quantifying dataset usage based on the analysis of references and the paper full text. **(2)** We present an open-source annotation tool which allows for validation of the method, and can aid in tracking dataset usage in research papers. **(3)** We apply both tools to study the usage of several popular segmentation and classification datasets and their usage in MICCAI and MIDL conference papers between 2013 and 2023. **(4)** We discuss the limitations of our study and tools, display additional practices we found during our study, and provide recommendations to ease the tracking of datasets.

2. Related Work

Meta-research papers in medical imaging often focus on **methods**, for example surveys on deep learning (Litjens et al., 2017), different types of supervision (Cheplygina et al., 2019), human-in-the-loop methods (Budd et al., 2021) and so forth. As a by-product of annotating and categorizing papers, some surveys also provide lists of commonly used datasets (Çallı et al., 2021b).

More recently some **dataset**-focused reviews started to emerge, in particular for dermatology (Daneshjou et al., 2021; Wen et al., 2022) and ophtalmology (Khan et al., 2021). These reviews focus on the type of data that is available, and find various biases in the patient populations, and/or that metadata about the patient demographics is missing. However these papers do not examine dataset use.

Perhaps at the intersection of datasets and methods, there is work focusing on challenges (Eisenmann et al., 2022) which review participation in medical image competitions at MICCAI and ISBI. Such competitions are often seen as one of the drivers of publicly available datasets, but the impact of these datasets beyond these competitions is not known.

The closest to our work are studies that examine dataset usage in other published works. (Koch et al., 2021) analyse dataset usage on PapersWithCode across various applications of machine learning, and find that the diversity of datasets used is decreasing. Within medical imaging, Heller et al (Heller et al., 2019) examined the role of publicly available data in MICCAI papers between 2014 and 2018, and found among others that over 20% of papers using public data did not cite the dataset. Simkó et al (Simko et al., 2022) examined reproducibility in MIDL papers between 2018 and 2022 and found that papers using public datasets are becoming more common but without proper citations or links.

3. Quantifying medical image dataset use

We propose tools to evaluate the presence of datasets using the following definitions: a dataset is **cited** if its paper is present in the reference section, **mentioned** if its name, aliases or URL are in the abstract or in a section of the paper associated with the method or results (i.e., not in a related work or discussion sections), in a table or figure captions or in a footnote, showing an actual **usage**. We show our pipeline in Figure 1. There are four main components: user input about the list of venues and datasets to track, the open citation index tool OpenAlex (Priem et al., 2022), the full texts of the papers and GROBID, a tool to extract information from scholarly documents. We used OpenAlex because it has an official freely accessible API aggregating and standardizing information from multiple sources such as arXiv, Crossref and Pubmed, and we aimed to create a generalizable process that could be complemented with other tools. We compared it with OpenCitations to evaluate their completeness using the citations returned for a set of cardiac segmentation datasets. We found that 97% of the citations returned by OpenCitations were also returned by OpenAlex while only 84% returned by OpenAlex were returned by OpenCitations; thus we chose OpenAlex for our pipeline.

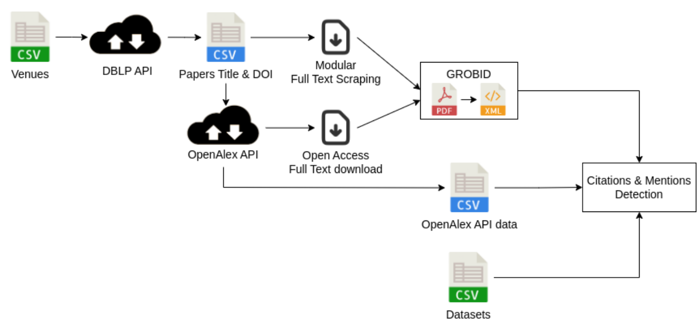


Figure 1: Pipeline to detect dataset presence and usage. Green CSV represents user input, blue CSV represents extracted data

First, we ask the user to specify the list of venues and datasets to track. This includes a dataset name, any aliases referring to the same datasets, and the titles and DOIs of papers associated with these datasets.

We use the venue list to fetch the list of papers from these venues using the DBLP API and store the papers' titles and DOIs. We then use the paper DOI (or title if the DOI is not available) to query the OpenAlex API to get the following: (i) list of referenced papers, (ii) list of words in the abstract, and (iii) open access link to the paper's full text, an example of data from OpenAlex is shown in Appendix D. We then try to fetch the paper's full text. If this step fails, we complement this step with a custom scraping tool. This step can be easily replaced for different venues, as long as the paper PDFs are stored in the same folder, In this step some data cleanup may be necessary to remove duplicates created from the combination of PDF extracted with OpenAlex and with the scraping tool. We then convert the PDF to an XML file using GROBID, a library applying a cascade of machine learning models that first segment the document in different structures like header, full text or reference and then specific models tuned for each type of structure to extract content. This allows to detect different paper sections, while keeping information about figures, tables and footnotes, which were lacking in alternative tools such as PyPDF, an example of the data obtained with GROBID is shown in Appendix E.

We use the dataset list to detect their citations and mentions. We detect citations in two ways: based on the dataset's DOI converted to an OpenAlex ID, and based on matching the dataset paper titles to the references sections of the full text. We detect dataset mentions by searching for the dataset's name, aliases or URL in certain fields of the full text. Examples of different types of citations and mentions can be seen in Appendix C.

Finally, we assign each dataset presence to one of the following types: only cited, only mentioned, and both cited and mentioned.

4. Annotation tool for paper PDFs

We also present our PDF annotation tool made with Streamlit, a Python library to easily create web apps. We used it to verify our detection process and therefore we designed it to fulfil two needs: having multiple users annotate the same project easily and being able to handle a large number of PDF files. While it was used to annotate datasets' presence in scientific papers, it can be extended to any PDF annotation task.

Once the software is installed on a local server, a user can create an annotation project by uploading the PDFs and choosing up to two initial sets of labels. In our study, the first set of labels corresponds to the list of datasets to detect and the second set is the list of locations a mention could be classified into (E.g. Abstract, Introduction, Method). While the second set of labels is fixed, the first one is not and new values can be added at any point during the labelling.

We also wanted to ease the annotation by multiple users. At the creation of the project, the owner can upload a file containing the division of the papers into different groups. This way, users can find the papers they were assigned to by selecting the right group on the annotation page. Finally, when the annotations are downloaded from the server, a file per person is obtained allowing more data processing afterwards.

5. A case study on publicly available medical datasets

5.1. Data selection

We apply our tools to a set of 20 publicly available medical datasets for both classification and segmentation of various organs shown in Appendix A. We initially tried a systematic procedure of identifying datasets via Google datasets and OpenAlex. However, this resulted in many poorly documented datasets (particularly on COVID-19) which did not have distinct names, and of which we could not trace whether they were in part duplicated from other datasets. Therefore, we selected datasets based on a combination of prior knowledge of the authors and consulting recent surveys in medical imaging which provided a table or list of datasets (Hesamian et al., 2019; El Jurdi et al., 2021; Niyas et al., 2022; Qureshi et al., 2022; Çallı et al., 2021a; Guan and Liu, 2021). In order to obtain enough data to analyse, we took the following aspects into account for the selection: presence of a paper linked to the dataset available in OpenAlex, year of publication, having some citations in OpenAlex, and having a unique name or acronym to make the detection process more reliable. We chose to analyse papers from two major conferences about medical image analysis, MICCAI and MIDL, so that the papers are more likely to contain the presence of such datasets.

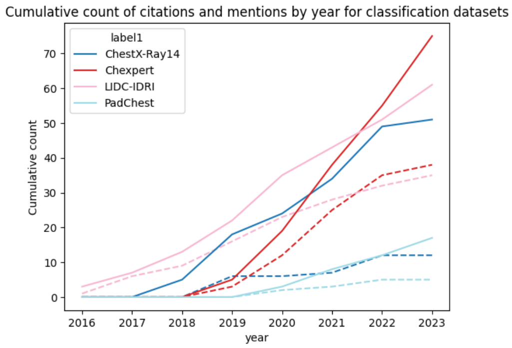
We identified 4835 papers in total (4569 from MICCAI and 266 from MIDL), however, 44 were discarded as we could not obtain information on the content of the paper or the list of references. We categorize the remaining papers in three groups, where for each group we slightly adjusted our processing due to the missing data:

- Group 1 with n=2327 papers either have all all information, or only miss the abstract from OpenAlex. In this case, we analyze the abstract from the full text of the paper.
- Group 2 with n=2237 where full text is not available, but we can still detect dataset mentions using the OpenAlex abstract. This is an important limitation, as the abstract does not always mention the datasets used. All the papers in this group are from MICCAI, showing the usefulness of the modular part to obtain the full texts. Unlike MICCAI papers, the structure of the PMLR website and the complete open access of PDFs made possible the development of the scraping tool for all MIDL full texts while they were not accessible from OpenAlex.
- Group 3 with n=227 papers which do not have references in OpenAlex. We therefore detect citations only with our simpler matching of dataset papers' titles with Grobid, which may result in less accurate detection. A majority of papers from this group are from MIDL as the information for papers from this conference is almost absent from OpenAlex. This shows that the download and analysis of the full text is a crucial and needed aspect of our method.

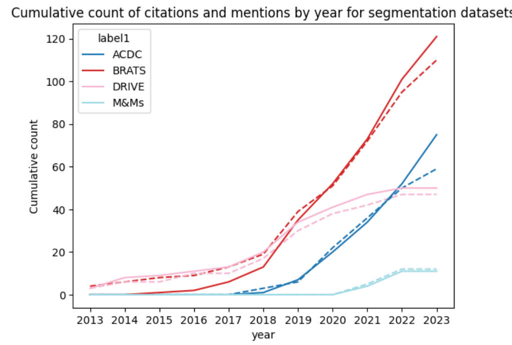
5.2. Concentration of research on few datasets

Although we considered the number of citations in OpenAlex to make the first selection of datasets, some datasets had very low numbers of citations and mentions in MICCAI and MIDL. We only present in Figures 2 and 3 results for eight datasets with the highest usage or that exemplify one of our conclusions, a more complete version including all the datasets can be found in Appendix B. This result may highlight the focus on some particular datasets also shown in (Koch et al., 2021) when using publicly available data, especially for datasets for the same task (cardiac segmentation and chest classification) as ACDC and M&Ms. This is also visible in Figure 2 with the large gap between the count of citations and mentions for BRATS and the rest of the most present datasets.

Note the difference in growth between the datasets, which might suggest a snowball effect where popular datasets become even more popular. This seems to be the case for BRATS, ACDC or Chexpert which have a very strong growth in citations and mentions. For other datasets like LIDC-IDRI or DRIVE, the number of citations and mentions is more gradual and even stagnates for DRIVE. Multiple factors can impact the popularity of a dataset, one of the most straightforward is the year of publication but while Chexpert and PadChest have been released at the same time, the second is almost absent from our list of papers. Therefore, other aspects such as how the dataset is updated or has a competition been organized with the dataset could be an explanation for such differences.



(a)



(b)

Figure 2: Cumulative counts per year of dataset citations (full line) and mentions (dashed line) for classification datasets (a) and segmentation datasets (b).

5.3. Lack of citation standards leads to difficulty in tracking usage

A dataset's citation doesn't necessarily imply actual usage and not all used datasets are cited in the references section. We further analyze this difference between mention and citations with Figure 3 in which we assign each presence of a dataset in a paper to one of the categories described in Section 3. We found out that even if there is variability in the groups' proportion for each subset, we can observe that almost every subset has more than 25% of datasets being only cited and around 10% being only mentioned. We considered papers from the "Only Cited" group as not using the dataset while citing it in the introduction or related works, mostly for general statements about machine learning usage in the medical sector. However, 132 papers out of 233 miss the full text and therefore only the abstract is used to detect the mention, a fraction of these papers could therefore mention the dataset and use it but the lack of information prevents our tool from detecting it. On the other hand, the "Only Mentioned" group mostly represents papers that are using a dataset without citing the associated paper. These two groups display the lack of standards to indicate the usage of a dataset such that it can be easily tracked. It also supports our approach to analyze part of the full text to determine such a usage.

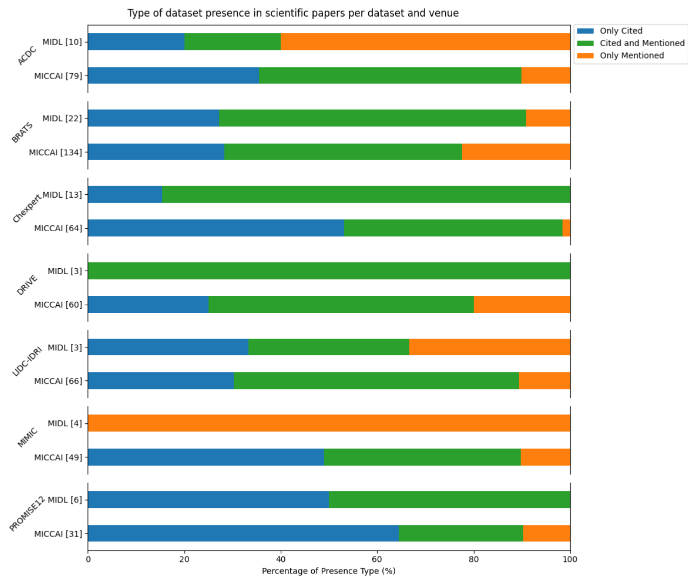


Figure 3: Type of presence per dataset and venue. The number in [] indicates the total number of papers for this subset. The "Only Cited" group in blue represents papers that cite a dataset without having a mention detected and therefore may not use it. The "Only Mentioned" group in orange represent the bad citation practice as the usage would not be detected by tools tracking the citations. The "Cited and Mentioned" group in green represent the best practice.

6. Discussion

We presented two open-source tools for the detection of dataset usage in scientific papers and applied them to a case study on publicly available medical datasets. We show that papers in major medical conferences tend to use a limited set of datasets, especially for papers addressing the same task. We also found that the lack of citation standards for dataset usage makes tracking such usage difficult, in particular due to (i) papers citing a dataset's paper without mentioning it in particular sections, indicating a non-usage, and (ii) papers mentioning a dataset without citing its paper, which classical bibliometric tools like OpenAlex can not detect.

Our study is limited to a set of datasets and venues manually selected and may therefore be biased by this selection. We also did not try to disambiguate between different datasets versions (for example different years of BRATS or datasets with similar names) due to already having difficulties with identifying these more-easily-identifiable datasets, it could however be valuable information to not overestimate the usage of a dataset or distinguish various tasks present across different version of a dataset. Doing a study on more datasets, venues and tasks would strengthen the outcome of our work. While datasets can be cited but not an associated paper, OpenAlex only keeps track of citations to papers. It is an important limitation and therefore a more precise matching of citations using GROBID could be a solution to track citations without a paper like it can be for Kaggle datasets.

Our method relies on regex matching and their location, it makes our tool usable to other data easier as only some information needs to be changed. We did not use text classification methods based on deep learning, such as fine-tuning a model pre-trained on scientific data like (Beltagy et al., 2019). While this could result in better performances, it implies a fine-tuning for every new set of datasets, reducing the applicability of our tools to new settings.

While doing this study we had some anecdotal findings that we do not report on in the paper, but which we feel may warrant further study.

- We saw the number of citations a paper has doubled, from 10 to 20, in 2019. This is likely because until 2019 MICCAI used to include citations in the 8-page limit. Relaxing such page restrictions may incentivize authors to not omit dataset citations.
- We found many instances of papers associated with Github repositories that were promising the code to be available upon acceptance, but never actually did this.
- We found cases where a "backup" of a dataset on Kaggle was cited as if it were the original source. The dataset was often stripped of its metadata and license information, and it was not clear whether the data was exactly the same or a derivative of the original, for a longer discussion please see (Jiménez-Sánchez et al., 2024).
- We discovered that ACDC is a popular name, as it can refer to the Automated Cardiac Diagnosis Challenge (Bernard et al., 2018a) but also to the Adverse Conditions Dataset with images of streets (Sakaridis et al., 2021) or to the Automatic Cancer Detection and Classification in Whole-slide Lung Histopathology challenge (Li et al., 2018).

We believe that better knowledge and therefore easier access to dataset usage information are needed. In addition to giving due credit to the creators of the dataset, it can raise awareness of the overuse of a particular dataset for a task, which could have a negative impact on real performance, but also an over-representation of a task in regards of real clinical needs. Working towards the adoption of a standard for indicating the usage of a dataset seems to be an essential step to achieve this objective. As examples, NeuroImage has a specific section on data availability at the end of each manuscript, and in 2023, MICCAI added the obligation to declare "the data origin, data license, and (when appropriate) ethics application number for any public or private data used in the preparation of the paper". While such requirements will not solve all the issues at hand, we believe that including a "Data availability" section could be an easy solution to put in place that would pave the way towards more systematic ways of determining the usage of a dataset. There are of course still many unanswered questions as to how exactly we want to implement this, for example what to do in cases of derivative datasets, synthetic data, and so forth, which we hope we can discuss together as a community.

Acknowledgments

Danish Data Science Academy DDSA-V-2022-004 and DFF - Inge Lehmann 1134-00017B

References

Samuel G. Armato, Henkjan Huisman, Karen Drukker, Lubomir Hadjiiski, Justin S. Kirby, Nicholas Petrick, George Redmond, Maryellen L. Giger, Kenny Cha, Artem Mamonov, Jayashree Kalpathy-Cramer, and Keyvan Farahani. PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *Journal of Medical Imaging*, 5(4):044501, 2018. doi: 10.1117/1.JMI.5.4.044501. URL https://doi.org/10.1117/1.JMI.5.4.044501.

Samuel G Armato, 3rd, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, Ella A Kazerooni, Heber MacMahon, Edwin J R Van Beeke, David Yankelevitz, Alberto M Biancardi, Peyton H Bland, Matthew S Brown, Roger M Engelmann, Gary E Laderach, Daniel Max, Richard C Pais, David P Y Qing, Rachael Y Roberts, Amanda R Smith, Adam Starkey, Poonam Batrah, Philip Caligiuri, Ali Farooqi, Gregory W Gladish, C Matilda Jude, Reginald F Munden, Iva Petkovska, Leslie E Quint, Lawrence H Schwartz, Baskaran Sundaram, Lori E Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Casteele, Sangeeta Gupte, Maha Sallamm, Michael D Heath, Michael H Kuhn, Ekta Dharaiya, Richard Burns, David S Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, and Barbara Y Croft. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med. Phys.*, 38(2): 915–931, February 2011.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text, 2019.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohe, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jager, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Isgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018a.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11): 2514–2525, 2018b.

Nicholas Bien, Pranav Rajpurkar, Robyn L. Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F. Amanatullah, Christopher F. Beaulieu, Geoffrey M. Riley, Russell J. Stewart, Francis G. Blankenberg, David B. Larson, Ricky H. Jones, Curtis P. Langlotz, Andrew Y. Ng, and Matthew P. Lungren. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of mrnet. *PLOS Medicine*, 15(11):1–19, 11 2018. doi: 10.1371/journal.pmed.1002699. URL https://doi.org/10.1371/journal.pmed.1002699.

Esther E. Bron, Marion Smits, Wiesje M. van der Flier, Hugo Vrenken, Frederik Barkhof, Philip Scheltens, Janne M. Papma, Rebecca M.E. Steketee, Carolina Méndez Orellana, Rozanna Meijboom, Madalena Pinto, Joana R. Meireles, Carolina Garrett, António J. Bastos-Leite, Ahmed Abdulkadir, Olaf Ronneberger, Nicola Amoroso, Roberto Bellotti, David Cárdenas-Peña, Andrés M. Álvarez Meza, Chester V. Dolph, Khan M. Iftekharuddin, Simon F. Eskildsen, Pierrick Coupé, Vladimir S. Fonov, Katja Franke, Christian Gaser, Christian Ledig, Ricardo Guerrero, Tong Tong, Katherine R. Gray, Elaheh Moradi, Jussi Tohka, Alexandre Routier, Stanley Durrleman, Alessia Sarica, Giuseppe Di Fatta, Francesco Sensi, Andrea Chincarini, Garry M. Smith, Zhivko V. Stoyanov, Lauge Sørensen, Mads Nielsen, Sabina Tangaro, Paolo Inglese, Christian Wachinger, Martin Reuter, John C. van Swieten, Wiro J. Niessen, and Stefan Klein. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: The caddementia challenge. *NeuroImage*, 111:562–579, 2015. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2015.01.048. URL https://www.sciencedirect.com/science/article/pii/S1053811915000737.

AJ Buckeye, Jacob Kriss, Josette BoozAllen, Josh Sullivan, Meghan O'Connell, and Will Cukierski Nilofer. Data science bowl 2017, 2017. URL https://kaggle.com/competitions/data-science-bowl-2017.

Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71: 102062, 2021.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020.

Victor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, Mario Parreno, Alberto Albiol, Fanwei Kong, Shawn C. Shadden, Jorge Corral Acero, Vaanathi Sundaresan, Mina Saber, Mustafa Elattar, Hongwei Li, Bjoern Menze, Firas Khader, Christoph Haarburger, Cian M. Scannell, Mitko Veta, Adam Carscadden, Kumaradevan Punithakumar, Xiao Liu, Sotirios A. Tsaftaris, Xiaoqiong Huang, Xin Yang, Lei Li, Xiahai Zhuang, David Vilades, Martin L. Descalzo, Andrea Guala, Lucia La Mura, Matthias G. Friedrich, Ria Garg, Julie Lebel, Filipe Henriques, Mahir Karakas, Ersin Cavus, Steffen E. Petersen, Sergio Escalera, Santi Segui, Jose F. Rodriguez-Palomares, and Karim Lekadir. Multicentre, multi-vendor and multi-disease cardiac segmentation: The m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021.

Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.

Roxana Daneshjou, Mary P Smith, Mary D Sun, Veronica Rotemberg, and James Zou. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA dermatology*, 157(11):1362–1369, 2021.

Matthias Eisenmann, Annika Reinke, Vivienn Weru, Minu Dietlinde Tizabi, Fabian Isensee, Tim J Adler, Patrick Godau, Veronika Cheplygina, Michal Kozubek, Sharib Ali, et al. Biomedical image analysis competitions: The state of current participation practice. *arXiv preprint arXiv:2212.08568*, 2022.

Rosana El Jurdi, Caroline Petitjean, Paul Honeine, Veronika Cheplygina, and Fahed Abdallah. High-level prior-based loss functions for medical image segmentation: A survey. *Computer Vision and Image Understanding*, 210:103248, 2021.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

Nicholas Heller, Jack Rickman, Christopher Weight, and Nikolaos Papanikolopoulos. The role of publicly available data in miccai papers from 2014 to 2018. In *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention: International Workshops, LABELS 2019, HAL-MICCAI 2019, and CuRIOUS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings 4*, pages 70–77. Springer, 2019.

Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32:582–596, 2019.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

Amelia Jiménez-Sánchez, Natalia-Rozalia Avlona, Dovile Juodelyte, Théo Sourget, Caroline Vang-Larsen, Hubert Dariusz Zając, and Veronika Cheplygina. Towards actionability for open medical imaging datasets: lessons from community-contributed platforms for data management and stewardship. *arXiv preprint arXiv:2402.06353*, 2024.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Saad M Khan, Xiaoxuan Liu, Siddharth Nath, Edward Korot, Livia Faes, Siegfried K Wagner, Pearse A Keane, Neil J Sebire, Matthew J Burton, and Alastair K Denniston. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health*, 3(1):e51–e66, 2021.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. *arXiv preprint arXiv:2112.01716*, 2021.

Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data*, 4(1):170177, December 2017.

Zhang Li, Zheyu Hu, Jiaolong Xu, Tao Tan, Hui Chen, Zhi Duan, Ping Liu, Jun Tang, Guoping Cai, Quchang Ouyang, Yuling Tang, Geert Litjens, and Qiang Li. Computer-aided diagnosis of lung carcinoma using deep learning - a pilot study, 2018.

Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, Robin Strand, Filip Malmberg, Yangming Ou, Christos Davatzikos, Matthias Kirschner, Florian Jung, Jing Yuan, Wu Qiu, Qinquan Gao, Philip "Eddie" Edwards, Bianca Maan, Ferdinand van der Heijden, Soumya Ghose, Jhimli Mitra, Jason Dowling, Dean Barratt, Henkjan Huisman, and Anant Madabhushi. Evaluation of prostate segmentation algorithms for mri: The promise12 challenge. *Medical Image Analysis*, 18(2):359–373, 2014. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2013.12.002. URL https://www.sciencedirect.com/science/article/pii/S1361841513001734.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, Quirine F Manson, Nikolas Stathonikos, Alexi Baidoshvili, Paul van Diest, Carla Wauters, Marcory van Dijk, and Jeroen van der Laak. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*, 7(6):giy065, 05 2018. ISSN 2047-217X. doi: 10.1093/gigascience/giy065. URL https://doi.org/10.1093/gigascience/giy065.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M.

John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. doi: 10.1109/TMI.2014.2377694.

Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2022.

S Niyas, SJ Pawan, M Anand Kumar, and Jeny Rajan. Medical image segmentation with 3d convolutional neural networks: A survey. *Neurocomputing*, 493:397–413, 2022.

Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in brief*, 32:106221, 2020.

Caroline Petitjean, Maria A. Zuluaga, Wenjia Bai, Jean-Nicolas Dacher, Damien Grosgeorge, Jérôme Caudron, Su Ruan, Ismail Ben Ayed, M. Jorge Cardoso, Hsiang-Chou Chen, Daniel Jimenez-Carretero, Maria J. Ledesma-Carbayo, Christos Davatzikos, Jimit Doshi, Guray Erus, Oskar M.O. Maier, Cyrus M.S. Nambakhsh, Yangming Ou, Sébastien Ourselin, Chun-Wei Peng, Nicholas S. Peters, Terry M. Peters, Martin Rajchl, Daniel Rueckert, Andres Santos, Wenzhe Shi, Ching-Wei Wang, Haiyan Wang, and Jing Yuan. Right ventricle segmentation from cardiac MRI: A collation study. *Medical Image Analysis*, 19(1):187–202, 2015.

Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022.

Imran Qureshi, Junhua Yan, Qaisar Abbas, Kashif Shaheed, Awais Bin Riaz, Abdul Wahid, Muhammad Waseem Jan Khan, and Piotr Szczuko. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Information Fusion*, 2022.

Perry Radau, Yingli Lu, Kim Connelly, Gideon Paul, Alexander J Dick, and Graham A Wright. Evaluation framework for algorithms segmenting short axis cardiac MRI. *The MIDAS Journal*, July 2009.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

Attila Simko, Anders Garpebring, Joakim Jonsson, Tufve Nyholm, and Tommy Löfstedt. Reproducibility of the methods in medical imaging with deep learning. *arXiv preprint arXiv:2210.11146*, 2022.

J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, and B. van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004. doi: 10.1109/TMI.2004.825627.

Avan Suinesiaputra, Brett R. Cowan, J. Paul Finn, Carissa G. Fonseca, Alan H. Kadish, Daniel C. Lee, Pau Medrano-Gracia, Simon K. Warfield, Wenchao Tao, and Alistair A. Young. Left ventricular segmentation challenge from cardiac MRI: A collation study. In *Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges*, pages 88–97. Springer Berlin Heidelberg, 2012.

Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *Nature Digital Medicine*, 5(1):1–8, 2022.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Computer Vision and Pattern Recognition*, pages 2097–2106, 2017.

David Wen, Saad M Khan, Antonio Ji Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, Carlos de Blas Perez, Alastair K Denniston, Xiaoxuan Liu, et al. Characteristics of publicly available skin cancer image datasets: a systematic review. *The Lancet Digital Health*, 4(1):e64–e74, 2022.

Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E. Ho, and James Zou. How medical ai devices are evaluated: limitations and recommendations from an analysis of fda approvals. *Nature Medicine*, 27(4):582–584, April 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-01312-x. URL http://dx.doi.org/10.1038/s41591-021-01312-x.

Kevin Wu, Eric Wu, Brandon Theodorou, Weixin Liang, Christina Mack, Lucas Glass, Jimeng Sun, and James Zou. Characterizing the clinical adoption of medical ai devices through u.s. insurance claims. *NEJM AI*, 1(1):AIoa2300030, 2024. doi: 10.1056/AIoa2300030. URL https://ai.nejm.org/doi/abs/10.1056/AIoa2300030.

Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G. van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72: 102125, 2021a. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2021.102125. URL https://www.sciencedirect.com/science/article/pii/S1361841521001717.

Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G. van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72: 102125, 2021b. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2021.102125. URL https://www.sciencedirect.com/science/article/pii/S1361841521001717.

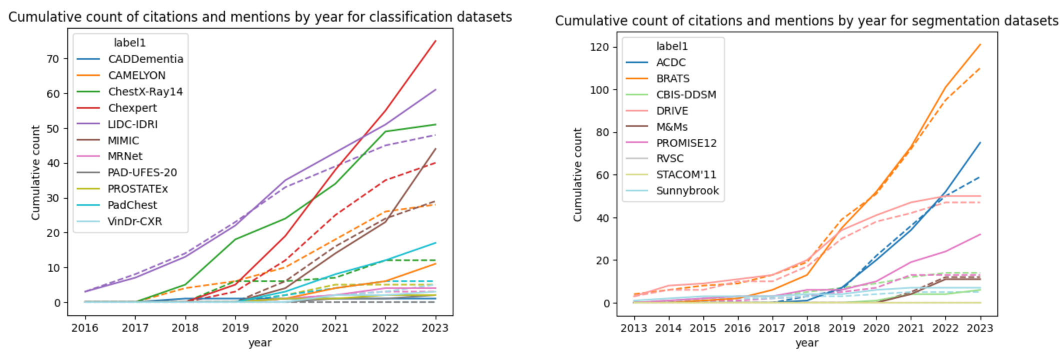
Appendix A. List of selected datasets

Table 1: Summary of selected datasets

Dataset	Organ	Published	Modality
Segmentation datasets			
ACDC (Bernard et al., 2018a)	Cardiac	2017	MRI
M&Ms (Campello et al., 2021)	Cardiac	2021	MRI
RVSC (Petitjean et al., 2015)	Cardiac	2015	MRI
STACOM'11 (Suinesiaputra et al., 2012)	Cardiac	2011	MRI
Sunnybrook (Radau et al., 2009)	Cardiac	2009	MRI
BRATS (Menze et al., 2015)	Brain	2014	MR
DRIVE (Staal et al., 2004)	Eye	2004	Fundus
CBIS-DDSM (Lee et al., 2017)	Breast	2017	Mammography
PROMISE12 (Litjens et al., 2014)	Prostate	2014	MR
Classification datasets			
ChestX-Ray14 (Wang et al., 2017)	Chest	2017	X-rays
Chexpert (Irvin et al., 2019)	Chest	2019	X-rays
LIDC-IDRI (Armato et al., 2011)	Chest	2011	CT
MIMIC (Johnson et al., 2019)	Chest	2002	X-rays
PadChest (Bustos et al., 2020)	Chest	2019	X-rays
VinDr-CXR (Nguyen et al., 2022)	Chest	2020	X-rays
CADDementia (Bron et al., 2015)	Brain	2015	MRI
CAMELYON (Litjens et al., 2018)	Breast	2018	whole-slide images
MRNet (Bien et al., 2018)	Knee	2018	MRI
PAD-UFES-20 (Pacheco et al., 2020)	Skin	2020	Phone picture
PROSTATEx (Armato et al., 2018)	Prostate	2018	mpMRI

Appendix B. Figures with original set of datasets

The following figures are the same as for Figures 2 and 3 without removing the datasets we considered not having enough matching. The non-presence of a dataset in one of the figures means that no paper contained a matching for this dataset.



(a) (b)

Figure 4: Cumulative counts per year of dataset citations (full line) and mentions (dashed line) for classification datasets (a) and segmentation datasets (b).

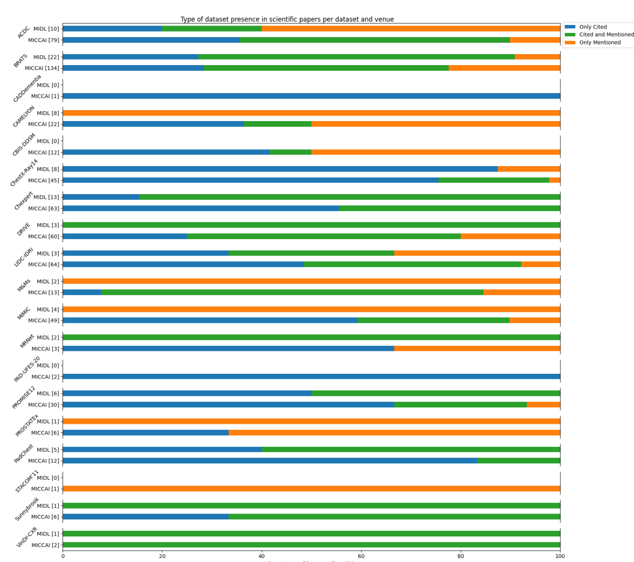


Figure 5: Type of presence per dataset and venue. The number in [] indicates the total number of papers for this subset.

Appendix C. Example of dataset presence

C.1. Citations

2015; Zhou et al., 2016). WSL has been applied in a wide range of medical-imaging applications, including the detection of lung disease in chest X-ray images (Wang et al., 2017; Yao et al., 2018; Tang et al., 2018; Ma et al., 2019; Liu et al., 2019; Guan et al., 2018), diagnosis of injuries from

Figure 6: Citation of a dataset without mention (Wang et al., 2017, ChestX-Ray8) in the background section for a demonstration of previous use

1. ISLES challenge (2017), http://www.isles-challenge.org/ISLES2017/
2. BRATS challenge (2018), https://www.med.upenn.edu/sbia/brats2018.html
3. ISLES challenge (2018), http://www.isles-challenge.org/ISLES2018/

Figure 7: Citation with a link to the datasets and not to the paper as indicated in BRATS guideline

C.2. Mentions

In text:

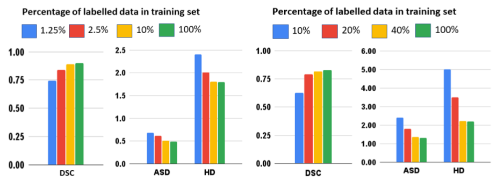
Diseased MRI Data. For additional experiments, two brain disorder MRI datasets are used. First, we train the model with BRATS 2018 [2,11] dataset for brain tumor MRI generation, using 210 subjects in the training dataset labeled

Figure 8: Mention and citation to the papers of BRATS, following guidelines from the challenge

We also use the low grade glioma cases from the multimodal *Brain Tumor Segmentation* (BRATS) 2015 challenge[2]. This data contains 54 volumes imaged in

Figure 9: Mention of BRATS without a proper citation but only a footnote with a link to the data

In figures and tables:



(a) ACDC (b) MMWHS

Fig. 2: Quantitative segmentation performance of our proposed method using different percentage of labelled data of ACDC and MMWHS.

Figure 10: ACDC mentioned in a figure's caption

Table 1: Quantitative segmentation results of 2D U-Net, 3D U-Net, Residual DSM and Distill DSM on BRATS 2020 dataset. ET represents Enhancing Tumor, WT represents Whole Tumor and TC represents Tumor Core

	Class	2D U-Net	Residual DSM	3D U-Net	Distill DSM(Ours)
Parameters		1,082,211	1,082,211	4,288,208	1,216,266
Flops per voxel		38,662	38,735	58,709	39,456
Wall time per voxel(s)		7.9498e-7	8.1726e-7	8.6517e-7	8.2641e-7
Dice	ET	0.712	0.732	0.704	0.753
	WT	0.861	0.867	0.879	0.873
	TC	0.687	0.704	0.796	0.742
Sensitivity	ET	0.714	0.707	0.687	0.761
	WT	0.859	0.835	0.898	0.841
	TC	0.660	0.693	0.779	0.726
Specificity	ET	0.9997	0.99978	0.99975	0.99969
	WT	0.99903	0.99939	0.99896	0.99944
	TC	0.99975	0.99986	0.99958	0.9997
Hausdorff95	ET	35.20	29.21	43.27	30.52
	WT	6.52	8.42	11.46	5.98
	TC	27.39	34.85	18.84	32.87

Figure 11: BRATS mentioned in a table's caption

In footnotes:

[4] Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning
[5] Automated Cardiac Diagnosis Challenge (ACDC), MICCAI Challenge 2017

Figure 12: ACDC dataset's name mentioned in a footnote

[2] https://camelyon16.grand-challenge.org

Figure 13: CAMELYON dataset mentioned in a footnote with the URL

Appendix D. Example of data from OpenAlex


```
▼ abstract_inverted_index:
  ▼ Primary:
      0:    0
  ▼ angle:
      0:    1
      1:    42
  ▼ closure:
      0:    2
  ▼ glaucoma:
      0:    3
  ▼ (PACG):
      0:    4
  ▼ is:
      0:    5
      1:    19
      2:    37
      3:    93
      4:    194
  ▼ the:
      0:    6
      1:    29
      2:    33
      3:    41
      4:    44
      5:    63
      6:    71
      7:    82
      8:    89
      9:    97
      10:   113
```


Figure 14: Example of abstract obtained from OpenAlex

id: "https://openalex.org/W3096981517"
doi: "https://doi.org/10.1007/978-3-030-59722-1_70"
title: "A Macro-Micro Weakly-Supervised Framework for AS-OCT Tissue Segmentation"
display_name: "A Macro-Micro Weakly-Supervised Framework for AS-OCT Tissue Segmentation"
publication_year: 2020
publication_date: "2020-01-01"
ids: {…}
language: "en"
primary_location: {…}
type: "book-chapter"
type_crossref: "book-chapter"
indexed_in: […]
open_access:
    is_oa: true
    oa_status: "green"
    oa_url: "https://arxiv.org/pdf/2007.10007"
    any_repository_has_fulltext: true

Figure 15: Example of link to full text PDF obtained from OpenAlex

referenced_works_count: 15
referenced_works:
    0: "https://openalex.org/W1977623353"
    1: "https://openalex.org/W2160605010"
    2: "https://openalex.org/W2165394378"
    3: "https://openalex.org/W2302594701"
    4: "https://openalex.org/W2331515946"
    5: "https://openalex.org/W2566969499"
    6: "https://openalex.org/W2613836512"
    7: "https://openalex.org/W2799738340"
    8: "https://openalex.org/W2904884925"
    9: "https://openalex.org/W2918552801"
    10: "https://openalex.org/W2963198662"
    11: "https://openalex.org/W2963573435"
    12: "https://openalex.org/W2964309882"
    13: "https://openalex.org/W2979907638"
    14: "https://openalex.org/W3102538446"

Figure 16: Example of list of citations in a paper obtained from OpenAlex

Appendix E. Example of data from GROBID


```xml
−<TEI xml:space="preserve" xsi:schemaLocation="http://www.tei-c.org/ns/1.0 https://raw.githubusercontent.com/kermitt2/grobid/master/grobid-home/schemas/xsd/Grobid.xsd">
 −<teiHeader xml:lang="en">
   +<fileDesc></fileDesc>
   +<encodingDesc></encodingDesc>
   −<profileDesc>
     −<textClass>
       −<keywords>
           <term>Primary angle-closure glaucoma</term>
           <term>Weakly-supervised learning</term>
           <term>Segmentation</term>
           <term>AS-OCT</term>
         </keywords>
       </textClass>
     −<abstract>
       −<div>
         −<p>
             Primary angle closure glaucoma (PACG) is the leading cause of irreversible blindness among Asian people. Early detection of PACG is essential, so as to provide timely treatment and
             minimize the vision loss. In the clinical practice, PACG is diagnosed by analyzing the angle between the cornea and iris with anterior segment optical coherence tomography (AS-
             OCT). The rapid development of deep learning technologies provides the feasibility of building a computeraided system for the fast and accurate segmentation of cornea and iris
             tissues. However, the application of deep learning methods in the medical imaging field is still restricted by the lack of enough fully-annotated samples. In this paper, we propose a
             novel framework to segment the target tissues accurately for the AS-OCT images, by using the combination of weakly-annotated images (majority) and fully-annotated images
             (minority). The proposed framework consists of two models which provide reliable guidance for each other. In addition, uncertainty guided strategies are adopted to increase the
             accuracy and stability of the guidance. Detailed experiments on the publicly available AGE dataset demonstrate that the proposed framework outperforms the state-of-theart semi-/
             weakly-supervised methods and has a comparable performance as the fully-supervised method. Therefore, the proposed method is demonstrated to be effective in exploiting
             information contained in the weakly-annotated images and has the capability to substantively relieve the annotation workload.
           </p>
         </div>
       </abstract>
     </profileDesc>
   </teiHeader>
```


Figure 17: Example of a header with abstract obtained after GROBID conversion to XML


```xml
 −<text xml:lang="en">
   −<body>
     −<div>
         <head n="1">Introduction</head>
       −<p>
           Glaucoma is the leading cause of irreversible vision loss world-widely that is predicted to affect more than 100 million people by year 2040
           <ref type="bibr" target="#b18">[19]</ref>
           . Primary angle closure glaucoma (PACG), as a major subtype of glaucoma, develops when the angle between the iris and cornea is closed or narrowed, resulting in the blockage of
           drainage canals and sudden rise in intraocular pressure
           <ref type="bibr" target="#b15">[16]</ref>
           . In the clinical practice, the anterior segment optical coherence technology (AS-OCT)
           <ref type="bibr" target="#b13">[14]</ref>
           is widely utilized to obtain both quantitative and qualitative information on the anatomical structures of cornea and iris for the PACG diagnosis
           <ref type="bibr" target="#b5">[6,</ref>
           <ref type="bibr" target="#b9">10,</ref>
           <ref type="bibr" target="#b10">11,</ref>
           <ref type="bibr" target="#b11">12]</ref>
           . However, manual analysis of each image is laborious and requires professional knowledge. Although the rapid development of deep learning technologies reveals the feasibility of fully
           automatic anatomical structure segmentation with high accuracy
           <ref type="bibr" target="#b4">[5]</ref>
           , it still requires a large quantity of images with pixel-wise annotations for the related structures, which is time-consuming and expertisedemanding.
         </p>
       +<p></p>
       +<p></p>
       +<p></p>
       +<p></p>
       </div>
     −<div>
         <head n="2">Method</head>
```


Figure 18: Example of full text body obtained after GROBID conversion to XML

```
−<div type="references">
  −<listBibl>
    −<biblStruct xml:id="b0">
      −<analytic>
        −<title level="a" type="main">
           Pyramid Network with Online Hard Example Mining for Accurate Left Atrium Segmentation
         </title>
        −<author>
          −<persName>
             <forename type="first">Cheng</forename>
             <surname>Bian</surname>
           </persName>
         </author>
        +<author></author>
        +<author></author>
        +<author></author>
        +<author></author>
        +<author></author>
        +<author></author>
        +<author></author>
         <idno type="DOI">10.1007/978-3-030-12029-0_26</idno>
       </analytic>
      +<monogr></monogr>
     </biblStruct>
```


Figure 19: Example of a citation obtained after GROBID conversion to XML