Why is the winner the best?


1Division of Intelligent Medical Systems, German Cancer Research Center (DKFZ), Heidelberg, Germany
2Helmholtz Imaging, German Cancer Research Center (DKFZ), Heidelberg, Germany
Full affiliations given in Sec. 5. Acknowledgments/funding information given in Suppl. G.

Abstract

International benchmarking competitions have become fundamental for the comparative performance assessment of image analysis methods. However, little attention has been given to investigating what can be learnt from these competitions. Do they really generate scientific progress? What are common and successful participation strategies? What makes a solution superior to a competing method? To address this gap in the literature, we performed a multicenter study with all 80 competitions that were conducted in the scope of IEEE ISBI 2021 and MICCAI 2021. Statistical analyses performed based on comprehensive descriptions of the submitted algorithms linked to their rank as well as the underlying participation strategies revealed common characteristics of winning solutions. These typically include the use of multi-task learning (63%) and/or multi-stage pipelines (61%), and a focus on augmentation (100%), image preprocessing (97%), data curation (79%), and post-processing (66%). The “typical” lead of a winning team is a computer scientist with a doctoral degree, five years of experience in biomedical image analysis, and four years of experience in deep learning. Two core general development strategies stood out for highly-ranked teams: the reflection of the metrics in the model design and the focus on analyzing and handling failure cases. According to the organizers, 43% of the winning algorithms exceeded the state of the art but only 11% completely solved the respective domain problem. The insights of our study could help researchers (1) improve algorithm development strategies when approaching new problems, and (2) focus on open research questions revealed by this work.
Figure 1. Overview of the IEEE ISBI 2021 and MICCAI 2021 challenges. Under the umbrella of 35 challenges (each represented by a teaser image and acronym), a total of 80 competitions with dedicated leaderboards were organized, as detailed in Suppl. A-C. We used data from participants, organizers, and winners to address the key research questions of this contribution: (RQ1) What is common practice in challenge participation?, (RQ2) Do current competitions generate scientific progress?, and (RQ3) Which strategies characterize challenge winners?

1. Introduction

Validation of biomedical image analysis algorithms is typically conducted through so-called challenges – large international benchmarking competitions that compare algorithm performance on datasets addressing specific problems. Recent years have not only seen an increase in the complexity of the machine learning (ML) models used to solve the tasks, but also a substantial increase in the scientific impact of challenges, with results often being published in prestigious journals (e.g., [9, 28, 34, 41, 46]), and winners receiving tremendous attention in terms of citations and (sometimes) high monetary compensation [23]. However, despite this impact, little effort has so far been invested in investigating what can be learnt from a challenge. Firstly, we identified a notable gap in literature regarding insights into current common practices in challenges as well as studies that critically analyze whether challenges actually generate scientific progress. Secondly, while recent work has addressed the problem of deriving meaningful conclusions from challenges [29, 49], it still remains largely unclear what makes winners the best and hence what constitutes a good strategy for approaching a new challenge or problem. The specific questions are manifold, e.g., Which specific training paradigms are used in current winning solutions?, What are the most successful strategies for achieving generalization?, Is it beneficial to involve domain experts or to work in a large team?. While ablation studies on the effects of ML model component removal could be used to address some questions, they suffer from the major drawback of only providing insights into submitted solutions, but not into underlying strategies. Furthermore, they typically only allow for investigating few aspects of a solution, and come at the cost of a substantial carbon footprint.

To overcome these issues, we chose an approach that allowed us to systematically assess all of the aforementioned questions related to biomedical image analysis competitions within one cohesive study. To this end, members of the Helmholtz Imaging Incubator (HI) and of the Medical Image Computing and Computer Assisted Intervention (MICCAI) Special Interest Group on biomedical image analysis challenges designed a series of comprehensive international surveys that were issued to participants, organizers, and winners of competitions conducted within the IEEE International Symposium on Biomedical Imaging (ISBI) 2021 and the International Conference on MICCAI 2021. By collaborating with the organizers of all 80 competitions (100%, see overview in Suppl. A-C), we were able to link algorithmic design decisions and challenge participation strategies to the outcome captured in rankings. Based on the study data, we explicitly addressed three research questions: (RQ1) What is common practice in challenge participation?, (RQ2) Do current competitions generate scientific progress?, and (RQ3) Which strategies characterize challenge winners?

2. Methods

According to the Biomedical Image Analysis ChallengeS (BIAS) Enhancing the QUAlity and Transparency Of health Research (EQUATOR) guideline on biomedical challenges [31], a biomedical image analysis challenge is defined as an “[...] open competition on a specific scientific problem in the field of biomedical image analysis. A challenge may encompass multiple competitions related to multiple tasks, whose participating teams may differ and for which separate rankings/leaderboards/results are generated.”. As the term challenge task is uncommon in the ML
community, we will use the term competition instead. The term challenge will be reserved for the collection of tasks that are performed under the umbrella of one dedicated organization, represented by an acronym (Fig. 1). For our analyses, we targeted three main groups that are relevant in the context of challenges, namely (1) challenge participants, (2) challenge organizers, and (3) challenge winners. The following sections present the methodology developed to address the corresponding research questions RQ1-RQ3.

2.1. RQ1: What is common practice in challenge participation?

To investigate current common practice in biomedical image analysis challenge participation, we designed a survey that was addressed to challenge participants and structured in five parts covering: (1) general information on the team and the tackled task(s), (2) information on expertise and environment, (3) strategy for the challenge, (4) algorithm characteristics, and (5) miscellaneous information (details provided in Sec. 3).

The organizers of all IEEE ISBI 2021 challenges (30 competitions across 6 challenges [1,2,12,35,40,42]), and all MICCAI 2021 challenges (50 competitions across 29 challenges [3–6, 8, 10, 13, 14, 16, 18, 19, 21, 22, 26, 27, 32, 33, 36, 37, 43, 44, 48, 50, 51]) were invited to participate in the survey that was addressed to challenge participants and structured in five parts covering: (1) general information on the team and the tackled task(s), (2) information on expertise and environment, (3) strategy for the challenge, (4) algorithm characteristics, and (5) miscellaneous information (details provided in Sec. 3).

The survey of competition winners consisted of three main parts targeting the design decisions related to the winning submission, general recommended strategies for winning a competition, and the profile of a winner, respectively. In the first part, we asked the winners about the importance of various design decisions for their submitted method. These comprised design decisions related to (1) the training paradigm, such as the usage of multi-task learning or semi-supervised learning, (2) network details, such as the choice of loss function(s), (3) model initialization, specifically pretraining, (4) data usage, covering aspects like data...
curation, augmentation, data splitting, and sampling, (5) hyperparameters, (6) ensembling, (7) postprocessing, and (8) metrics (see Fig. 3). For each of these design decisions, winners specified their method (e.g., whether they performed pretraining and, if so, based on which data) and rated the importance of this design choice for winning the challenge. We further explicitly asked what distinguished the winning solution from competing solutions and what were key factors for success.

The second part of the survey investigated general successful strategies (independent of the specific challenge). To this end, several authors of this paper who had already won multiple challenges compiled the list of strategies (Fig. 4). The winners were asked to rate the importance of each strategy and further complement the list.

Finally, the third part of the survey covered questions on the profile of a challenge winner (Fig. 2). This was particularly relevant for those winners that had not taken part in the original survey of Sec. 2.1.

3. Results

Based on the positive responses of all organizers from all IEEE ISBI 2021 (n = 30) and MICCAI 2021 (n = 50) competitions, a total of 80 competitions conducted across 35 challenges were included in this study (Fig. 1). These covered a wide range of problems related to semantic segmentation, instance segmentation, image-level classification, tracking, object detection, registration, and pipeline evaluation.

3.1. Common practice in challenge participation

A median (min/max) of 72% (11%/100%) of the challenge participants took part in the survey, according to the closed-access surveys. Overall, we received 292 completed survey forms, of which 249 met our inclusion criteria (i.e., second version of the survey refined for MICCAI 2021, survey completed by a lead developer, no duplicate responses from the same team). Detailed responses to all aspects of the survey (including interquartile ranges (IQR) and min/max values of all parameters) are provided in a white paper [15]. This section summarizes a selection of answers. The profile of a winner is depicted in Suppl. D.

Infrastructure and strategies Knowledge exchange was the most important incentive for participation (mentioned by 70%; respondents were allowed to pick multiple answers), followed by the possibility to compare their own method to others (65%), having access to data (52%), being part of an upcoming challenge publication (50%), and winning a challenge (42%). The awards/prize money was important to only 16% of the respondents. Regarding the computing infrastructure, only 25% of all respondents thought that their infrastructure was a bottleneck. The vast majority of respondents used a Graphics Processing Unit (GPU) cluster. The total training time of all models trained during method development including failure models was estimated to be a median of 267 GPU hours, while the training time of the final submission was estimated to be a median of 24 GPU hours. The most popular frameworks were PyTorch for method implementation (76%), NumPy for analyzing data (37%), and NumPy for analyzing annotations/reference data (27%).

The most common approach to development (42%) consisted of going through related literature and building upon/modifying existing work. The majority (51%) estimated the edited lines of code of the final solution to be in the order of magnitude of $10^3$. A median of 80 working hours was spent on method development in total. The respondents reported more human-driven decisions (median of 60%), e.g., parameter setting based on expertise, than empirical decisions (median of 40%), e.g., automated hyperparameter tuning via grid search. 94% of the respondents used a deep learning-based approach. For those approaches, most time (up to three picks allowed) was spent on selecting one or multiple existing architectures that best matched the task (45%), configuring the data augmentation (33%), configuring the template architecture (e.g., How deep? How many stages/pooling layers?) (28%), exploring existing loss functions (25%), and ensembling (22%).

The survey revealed that almost one third of the respondents did not have enough time for development. A majority thereof (65%) felt that more time in the scale of weeks would have been beneficial (months: 18%, days: 14%).

Algorithm characteristics Among the deep learning-based approaches, only 9% actively used additional data, i.e., data not provided for the respective challenge, in their final solution (note that this does not include the usage of already pretrained models). One reason may be that some challenges (24%) explicitly do not allow the usage of external data. Of those that did leverage external data, the majority used public biomedical data for the same type of task (40%), private biomedical data for the same type of task (25%), or public biomedical data for a different type of task (15%). Non-biomedical data was only used in 5% of the cases. If additional data was used, it was used for pretraining (55%) and/or co-training (50%).

Data augmentation was applied by 85% of the respondents. The most common augmentations were random horizontal flip (77%), rotation (74%), random vertical flip (62%), contrast (49%), scale (48%), crop (44%), resize crop (35%), noise (34%), elastic deformation (26%), color jitter (19%), and shearing (15%). 43% of the respondents reported that the data samples were too large to be processed at once (e.g., due to GPU memory constraints). This issue was mainly solved by patch-based training (cropping) (69%), downsampling to a lower resolution (37%), and/or solving 3D analysis tasks as a series of 2D analysis tasks.
The most common loss functions were Cross-Entropy (CE) Loss (39%), combined CE and Dice Loss (32%), and Dice Loss (26%). 29% of the respondents used early stopping, 12% used warmup. Internal evaluation via a single train:val:test split was performed by more than half of the respondents (52%). K-fold cross-validation on the training set was performed by 37%. 6% did not perform any internal evaluation. 48% of the respondents applied postprocessing steps.

The final solution of 50% of the respondents was a single model trained on all available data. An ensemble of multiple identical models, each trained on the full training set but with a different initialization (random seed), was proposed by 6%. 21% proposed an ensemble of multiple identical models, each trained on a randomly drawn subset of the training set (regardless of whether the same seed was used or not). 9% reported having ensemble multiple different models and trained each on the whole training set (different seeds). 8% ensembled multiple different models, each trained on a randomly drawn subset of the training set (regardless of whether the same seed was used or not). If multiple models were used, the final solution was composed of a median of 5 models.

### 3.2. Key insights related to scientific progress generated by challenges

According to the responses of challenge organizers (n = 54), 43% of the winning algorithms exceeded the state of the art (Fig. 2). While substantial (47%) or minor (32%) progress was made in most competitions, the underlying problem was regarded as solved in only 11% of the competitions. Most progress was seen in new architectures/combination of architectures (32%), the phrasing of the optimization problem (e.g., new losses) (17%), and new augmentation strategies (14%). Failure cases were mainly attributed to specific imaging conditions (e.g., image blur) (27%), generalization issues (23%), and specific classes that perform particularly poorly (19%).

According to the responses from several organizers, the trend of simple algorithms (e.g., U-Net [17]/nnU-Net [24]) outperforming complex ones continued. As a prominent feature in 2021, many competitions provided additional information that is not usually available, such as the identifier of the hospital for domain generalization, multiple expert segmentations to represent label uncertainty, or k-space data in reconstruction problems. However, the participants were not able to leverage the additional data for better performance. The same holds true for temporal data in video analysis, although organizers hypothesize that frame-based analysis is not sufficient.

Several organizers also reported a lack of heterogeneity in methods. Often, submitted methods performed similarly (e.g., differing only in the fourth decimal digit in normalized scores). On a positive note, some competitions that had been run for multiple years observed a drastic improvement compared to previous years, sometimes even surpassing human performance. Regarding computational aspects, in one case the winning method surpassed the existing state-of-the-art method, achieving a 19 times faster inference speed and reduction of the GPU memory consumption by 60% while yielding comparable accuracy.

According to our study, generalization remains a major issue. One challenge, which mimicked “in-the-wild” deployment, found that models failed to generalize in 3 out of 21 testing institutions. Similarly, performance in rare classes was reported as a core issue in several competitions. This is a problem of high clinical relevance as diseases often correspond to a rare class. A related problem is the fact that...
the detection of multiple conditions in a multi-label setting still remains challenging. Finally, some organizers reported the failure of metrics to reflect the biomedical domain interest. Along these lines, pixel-level performance was sometimes reported to be substantial while instance-/case-level performance, which is typically biomedically more relevant, was not improved substantially.

Cheating was observed in 4% of the cases. It was related to an excessive number of submissions of similar methods with different user accounts or the attempt to retrieve the test set from the submission platform. In these cases, participants were excluded from the competition, the rankings and/or the publication.

3.3. Key insights related to winning strategies

When comparing winners to other participants, several differences stood out. Firstly, winners were more determined to win a challenge (64% vs. 40%). The majority of winning lead developers have a doctorate degree (41%) while the majority of non-winning lead developers have a master’s degree (47%) as their highest degree. Furthermore, while only 66% of other participants felt that there was enough development time, 86% of the winners agreed with this statement. Winners spent 120 hours (e.g., on method development, analyzing data and annotations) before deciding to submit, compared to 56 hours for other participants, and decided to submit a week earlier (3 vs. 2 weeks prior to submission). Notably, winners spent twice as much time on failure analysis (10% of median working hours dedicated to method development vs. 5%). Compared to non-winners, winners used ensembling based on random seeds, data splits, and heterogeneous models (see Fig. 3(f)) 5.6 times, 1.7 times, and 2.5 times as much.

According to univariable mixed model analysis, eight parameters were found to provide statistically significant differences between winners and non-winners ($p < 0.05$): (1) Number of team members who were developers/engineers, (2) time invested before planning to submit results, (3) time spent in data preprocessing/augmentation, (4) use of professionally managed GPU cluster, (5) approach used for method development, (6) architecture type, (7) taking metrics used to evaluate the challenge into account while searching for hyperparameters, and (8) augmentations used. Note, however, that when multiple independent tests are performed, 5% can be expected to be identified as significant purely by chance when testing at 5% significance level. Correcting for this so-called multiplicity of testing, we did not obtain statistically significant differences. Multivariable model analysis based on a selection of variables identified by image analysis experts revealed the willingness to win the challenge as the only parameter with $p < 0.05$ when comparing winners to non-winners (64% vs. 40%). Analogously, the parameter of taking metrics used to evaluate the challenge into account while searching for hyperparameters was identified in the best 30% vs. the rest analysis. It is worth mentioning in this context that despite the high response rate of 72%, the number of winners covered by the survey presented in Sec. 2.1 was only 22. The resulting low power of identifying important contributors to winning challenges may well be the reason for the absence of statistical significance. We therefore additionally asked competition winners after the results announcement for key design decisions and strategies. The responses (n = 38) cover 67% and 62% of the IEEE ISBI 2021 and MICCAI 2021 challenges respectively, and are summarized in Fig. 3 and Fig. 4.

As detailed in Fig. 3, the most applied training pipelines were multi-task designs (63%) and multi-stage pipelines (61%). If multi-stage pipelines were applied, the importance of this strategy for winning the challenge was rated crucial. Pretraining was mainly performed in a supervised fashion using in-domain data (55%) or generic data (e.g., ImageNet) (61%). The usage of in-domain data, however, was found to be much more important. As mentioned above, it should be noted that many competitions do not allow for the usage of external data (24% according to the survey presented in Sec. 2.2). The most commonly applied design decisions related to data usage were preprocessing (97%), augmentation (100%), data splitting (beyond the splits provided by the competition, e.g., for cross-validation) (89%), data curation (e.g., cleaning of annotations) (79%), and data sampling (58%). One aspect that stood out when asking winners for key factors for success (free text) was the setting up of a good internal validation strategy, including the careful selection of a baseline model and appropriate validation tests.

With respect to general strategies (Fig. 4), the strategies of analyzing and handling failure cases, knowing the state of the art, and reflecting the metrics in the method design were rated most highly. Further recommended strategies in free-text answers were heterogeneous and comprise (1) inclusion of non-deep learning approaches in a model ensemble, (2) explicit determination of a time management strategy, (3) test-time augmentation, and (4) preferring matured architectures over brand-new hype'd machine learning methods.

4. Discussion

The presented study represents, to the best of our knowledge, the first systematic and large-scale examination of biomedical image analysis competitions with a focus on what the scientific community can learn from them. Based on comprehensive surveys and statistical analyses for a total of 80 competitions within the scope of two major conferences in the field, it provides unprecedented insights into common practice among challenge participants, progress
Figure 3. Importance of design decisions for the neural network-based winning submission of the respective IEEE ISBI 2021 and MICCAI 2021 competition rated by the (team) lead and ordered by percentage of highest vote (crucially important: dark blue). Voting was only conducted among those who used the respective design. “Applied by” indicates the percentage of respondents using the respective design.

generated by competitions, open issues, as well as key winning strategies.

A new insight with respect to common participation practice (RQ1) was that knowledge exchange is the primary participation incentive. This will most likely differ on platforms like Kaggle, in which prize money and achieving a high rank are expected to be substantially more important [45]. To our surprise, only a small portion of participants perceived the limiting computing power as a bottleneck. Similarly surprisingly, k-fold cross-validation on the training set as well as ensembling was only performed by a minority of participants.

The competitions clearly led to substantial scientific progress according to the organizers (RQ2). Notably, however, only a small fraction of image analysis problems addressed by current competitions can be regarded as solved (Suppl. E). Open research questions identified as part of this work include: (1) How can we better integrate meta information in neural network solutions?, (2) How can we effectively leverage temporal information in biomedical video analysis?, (3) How can we achieve generalization across devices, protocols, and sites?, (4) How can we arrive at performance metrics that better address the biomedical domain interest? The latter is particularly interesting in light of the fact that the reflection of metrics in the challenge design was identified as a key strategy for winning a challenge. In line with recent literature [20, 25, 39, 47], it implies that common efforts are focused largely on an overfitting to the current metrics rather than solving the underlying domain problem. Current initiatives are already addressing this issue [30], but our results imply that challenge organizers should focus more on ensuring that the actual biomedical needs are reflected in the design of their competition.

Our work revealed particularly successful algorithm design choices (Fig. 3) and general strategies for winning a competition (Fig. 4) (RQ3). In the spirit of reporting
negative results, we also included the results of the mixed model analysis despite the lack of statistical significance after correction for multiplicity of testing. Given the relatively small dataset (results from 80 competitions) compared to the number of parameters that we extracted from algorithm designs and strategies (> 100), we hypothesize that the lack of statistical significance can largely be attributed to small sample size.

A limitation of our study could be seen in the fact that we only covered IEEE ISBI and MICCAI challenges of one specific year. Prior work, however, revealed that the competitions performed in the scope of these conferences cover the majority of all biomedical image analysis competitions [29]. Further limitations can be regarded as general limitations when working with surveys [11] and include the uncertainty of self-reported data and the potential bias resulting from the preselection of categorical variables. Finally, it is not straightforward to address the heterogeneity of challenges with a single questionnaire. For example, using an in-domain similar dataset may not always be feasible due to the sparsity of public biomedical datasets. Similarly, a researcher may regard ensembling as a general key strategy but may not have had the computing power to train and optimize multiple models working with video, 3D, or 4D data. To compensate for this effect in the design of the surveys presented in Sec. 2.1 and Sec. 2.2, we additionally asked winners for general recommended strategies (Fig. 4). The discrepancy between general recommendation and feasibility is reflected in the answers. For example, most winners recommend the integration of biologists/clinicians in a team but did not do so themselves.

Despite the discussed limitations, our findings have the potential to impact a plethora of stakeholders in challenges. First, biomedical image analysis researchers and developers can “stand on the shoulders of giants” (the competition winners) to improve algorithm development strategies when approaching a new problem. Second, future challenge organizers can adapt their designs carefully to the open issues revealed by this work. This would include a focus on case/instance level rather than pixel/voxel level to reflect biomedical needs, metrics that reflect biomedical needs (see below), as well as dataset designs that allow for improving the capabilities of algorithms to perform well on rare classes and to generalize across domains. Given that the vast majority of participants perceived limited time and not computing power as a bottleneck, challenge timelines should be critically questioned. Finally, the wider community can benefit from the open research questions we identified (Suppl. F).

In conclusion, we performed the first systematic analysis of biomedical image analysis competitions, which revealed a plurality of novel insights with respect to participation, organization, and winning. Our work could pave the way for (1) developers to improve algorithm development strategies when approaching new problems, and (2) the scientific community to channel its activities into open issues revealed by this work.
5. Full affiliation list

1Division of Intelligent Medical Systems, German Cancer Research Center (DKFZ), Heidelberg, Germany; 2Helmholtz Imaging, German Cancer Research Center (DKFZ), Heidelberg, Germany; 3Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany; 4Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany; 5Division of Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany; 6School of Computing, Faculty of Engineering and Physical Sciences, University of Leeds, Leeds, UK; 7Institute of Informatics, School of Management, HES-SO Valais-Wallis University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland; 8Department of Nuclear Medicine and Molecular Imaging, Lausanne University Hospital, Lausanne, Switzerland; 9Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany; 10Center for Artificial Intelligence and Data Science for Integrated Diagnostics (A²FD) and Center for Biomedical Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA; 11Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; 12Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; 13Department of Radiology, University of Washington, Seattle, WA, USA; 14Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) and Department of Computer Science, University College London, London, UK; 15Universitat Autònoma de Barcelona & Computer Vision Center, Barcelona, Spain; 16Division of Translational Surgical Oncology, National Center for Tumor Diseases (NCT/UCC) Dresden, Dresden, Germany; 17Department of Advanced Robotics, Istituto Italiano di Tecnologia, Italy and Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy; 18IT University of Copenhagen, Copenhagen, Denmark; 19Department of General, Visceral and Transplantation Surgery, Heidelberg University Hospital, Heidelberg, Germany; 20Biomedical Imaging Group Rotterdam, Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands; 21Department of Computer Science, University of Copenhagen, Copenhagen, Denmark; 22Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland; 23Harvard Medical School, Brigham and Women’s Hospital, Boston, MA, USA; 24School of Biomedical Engineering and Imaging Sciences, King’s College London, London, UK; 25Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen (AöR), Essen, Germany; 26University of Nebraska Medical Center, Omaha, NE, USA; 27Department of Internal Medicine III, Heidelberg University Hospital, Heidelberg, Germany; 28Neurobiology Research Unit, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark; 29Arab Academy of Science and Technology, Cairo, Egypt; 30CIBM Center for Biomedical Imaging, Lausanne, Switzerland; 31Radiology Department, Centre Hospitalier Universitaire Vaudois (CHUV) and University of Lausanne (UNIL), Lausanne, Switzerland; 32Signal Processing Laboratory (LTSS), École Polytechnique Fédérante de Lausanne (EPFL), Lausanne, Switzerland; 33National Center for Tumor Diseases (NCT), Heidelberg, Germany; 34SBILab, Department of ECE, IIT-Delhi, Delhi, India; 35University of Lübeck, Lübeck, Germany; 36Mechanical Engineering, School of Engineering, The University of Tokyo, Tokyo, Japan; 37University of Minnesota, Department of Computer Science & Engineering, Minneapolis, MN, USA; 38Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, The Netherlands; 39Fraunhofer MEVIS, Lübeck, Germany; 40Univ Rennes, INSERM, LTSI - UMR 1099, F35000, Rennes, France; 41Brno University of Technology, Brno, Czech Republic; 42Centre for Biomedical Image Analysis, Masaryk University, Brno, Czech Republic; 43Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland; 44Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada; 45Département de Mathématiques & Informatique, Universitat de Barcelona, Barcelona, Spain; 46Biomedical Image Analysis & Machine Learning, Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland; 47Department of Engineering Science, University of Oxford, Oxford, UK; 48Institute of Informatics, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland; 49ICube, University of Strasbourg, CNRS, Strasbourg, France; 50IHU Strasbourg, Strasbourg, France; 51Center For Artificial Intelligence And Data Science For Integrated Diagnostics (A²FD) and Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA; 52Department of Informatics, Technical University of Munich, Munich, Germany; 53Center for MR Research, University Children’s Hospital Zurich, University of Zurich, Zurich, Switzerland; 54Neuroscience Center Zurich, University of Zurich, Zurich, Switzerland; 55Visual Artificial Intelligence Laboratory (VAIL), Oxford Brooks University, Oxford, UK; 56Centre for Tactile Internet with Human-in-the-Loop (CeTI), TU Dresden, Dresden, Germany; 57MRC Unit for Lifelong Health and Ageing, University College London, London, UK; 58Centre for Medical Image Computing, University College London, London, UK; 59School of Biomedical Engineering & Imaging Sciences, King’s College London, London, UK; 60Dementia Research Centre, University College London, London, UK; 61Computer Science, Boston College, Boston, USA; 62Department of Computing and Mathemat-
ics, Manchester Metropolitan University, Manchester, UK; 63Medical Faculty, Heidelberg University, Heidelberg, Germany; 64Intuitive Surgical, Inc., Sunnyvale, CA, USA; 65A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio, Finland; 66Department of Neuroscience and Biomedical Engineering, Aalto University School of Science, Espoo, Finland; 67University of Aberdeen, Aberdeen, UK; 68Department of Computer Science, University of Applied Sciences and Arts Dortmund, Dortmund, Germany; 69Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Essen, Germany; 70Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen, Essen, Germany; 71School of Computing, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea; 72Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA; 73The Chinese University of Hong Kong, Hong Kong; 74Department of Computer Science, Technical University of Munich, Munich, Germany; 75The University of Texas MD Anderson Cancer Center, Houston, TX, USA; 76Nepal Applied Mathematics and Informatics Institute for Research (NAAMII), Lalitpur, Nepal; 77Universidad Pompeu Fabra, Barcelona, Spain; 78University of Adelaide, Australia, Australia; 79XLAB d.o.o., Ljubljana, Slovenia; 80Touch Surgery, Medtronic, London, UK; 81CJ AI Center, Seoul, Republic of Korea; 82Tissue Image Analytics Centre, Department of Computer Science, University of Warwick, Coventry, UK; 83Hankuk University of Foreign Studies, Yongin, Republic of Korea; 84Massachusetts General Hospital, Boston, MA, USA; 85Harvard Medical School, Boston, MA, USA; 86Helmholtz AI, Helmholtz Zentrum München, Munich, Germany; 87Department of Informatics, Technical University Munich, Munich, Germany; 88TranslATUM - Central Institute for Translational Cancer Research, Technical University of Munich, Munich, Germany; 89Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany; 90Muroran Institute of Technology, Hokkaido, Japan; 91UMC Utrecht, Utrecht, The Netherlands; 92University of Science and Technology of China, Hefei, China; 93Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea; 942Ai, School of Technology, IPCA, Barcelos, Portugal; 95Algoritmi Center, School of Engineering, University of Minho, Guimarães, Portugal; 96Life and Health Sciences Research Institute, School of Medicine, University of Minho, Braga, Portugal; 97Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA; 98Sano Centre for Computational Medicine, Cracow, Poland; 99Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands; 100Department of Biomedical Engineering and Physics, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands; 101LRE, EPITA, Paris, France; 102Artificial Intelligence and Robotics Institute, Korea Institute of Science and Technology, Seoul, Republic of Korea; 103Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE; 104Harbin Institute of Technology, Shenzhen, China; 105University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia; 106ISEP, Paris, France; 107Department of Computer Science at School of Informatics, Xiamen University, Xiamen, China; 108College of Computer Science, Sichuan University, Chengdu, China; 109Department of Neuroradiology, Technical University of Munich, Munich, Germany; 110AGH UST, Department of Measurement and Electronics, Kraków, Poland; 111University of Applied Sciences and Arts Western Switzerland (HES-SO), Sierre, Switzerland; 112Data Science and Learning Division, Argonne National Laboratory, Lemont, IL, USA; 113University of Chicago, Chicago, IL, USA; 114Shaanxi Normal University, Xi’an, China; 115AI Lab, Tencent, Shenzhen, China; 116Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany; 117Interactive Machine Learning Group, German Cancer Research Center (DKFZ), Heidelberg, Germany

References

[6] Spyridon Bakas, Micah Sheller, Sarthak Pati, Brandon Edwards, G. Anthony Reina, Ujjwal Baid, Yong Chen, Russ...


should be interpreted with care. *Nature Communications*, 9(1):5217, 2018. 2, 8


