


Please cite the Published Version

Eisenmann, M, Reinke, A, Weru, V, Tizabi, MD, Isensee, F, Adler, TJ, Ali, S, Andrearczyk, V, Aubreville, M, Baid, U, Bakas, S, Balu, N, Bano, S, Bernal, J, Bodenstedt, S, Casella, A, Cheplygina, V, Daum, M, De Bruijne, M, Depeursinge, A, Dorent, R, Egger, J, Ellis, DG, Engelhardt, S, Ganz, M, Ghatwary, N, Girard, G, Godau, P, Gupta, A, Hansen, L, Harada, K, Heinrich, M, Heller, N, Hering, A, Huauilmé, A, Jannin, P, Kavur, AE, Kodym, O, Kozubek, M, Li, J, Li, H, Ma, J, Martín-Isla, C, Menze, B, Noble, A, Oreiller, V, Padoy, N, Pati, S, Payette, K, Rädtsch, T, Rafael-Patiño, J, Bawa, VS, Speidel, S, Sudre, CH, Van Wijnjen, K, Wagner, M, Wei, D, Yamlahi, A, Yap, MH , Yuan, C, Zenk, M, Zia, A, Zimmerer, D, Aydogan, D, Bhattarai, B, Bloch, L, Brüngel, R, Cho, J, Choi, C, Dou, Q, Ezhov, I, Friedrich, CM, Fuller, C, Gaire, RR, Galdran, A, García Faura, A, Grammatikopoulou, M, Hong, S, Jahanifar, M, Jang, I, Kadkhodamohammadi, A, Kang, I, Kofler, F, Kondo, S, Kuijff, H, Li, M, Luu, M, Martinčić, T, Morais, P, Naser, MA, Oliveira, B, Owen, D, Pang, S, Park, J, Park, S, Plotka, S, Puybareau, E, Rajpoot, N, Ryu, K and Saeed, N (2023) Why is the winner the best? In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 17 June 2023 - 24 June 2023, Vancouver, Canada.

DOI: <https://doi.org/10.1109/CVPR52729.2023.01911>

Publisher: IEEE

Version: Accepted Version

Downloaded from: <https://e-space.mmu.ac.uk/635029/>

Usage rights:  In Copyright

Additional Information: © 2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

Why is the winner the best?

M. Eisenmann^{1, 2}, A. Reinke^{1, 2, 3}, V. Weru⁴, M. D. Tizabi^{1, 2}, F. Isensee^{5, 2}, T. J. Adler¹, S. Ali⁶, V. Andrearczyk^{7, 8}, M. Aubreville⁹, U. Baid^{10, 11, 12}, S. Bakas^{10, 11, 12}, N. Balu¹³, S. Bano¹⁴, J. Bernal¹⁵, S. Bodenstedt¹⁶, A. Casella¹⁷, V. Cheplygina¹⁸, M. Daum¹⁹, M. de Bruijne^{20, 21}, A. Depeursinge^{22, 8}, R. Dorent^{23, 24}, J. Egger²⁵, D. G. Ellis²⁶, S. Engelhardt²⁷, M. Ganz^{28, 21}, N. Ghatwary²⁹, G. Girard^{30, 31, 32}, P. Godau^{1, 2, 3, 33}, A. Gupta³⁴, L. Hansen³⁵, K. Harada³⁶, M. Heinrich³⁵, N. Heller³⁷, A. Hering^{38, 39}, A. Huault⁴⁰, P. Jannin⁴⁰, A. E. Kavur^{1, 2}, O. Kodym⁴¹, M. Kozubek⁴², J. Li²⁵, H. Li⁴³, J. Ma⁴⁴, C. Martín-Isla⁴⁵, B. Menze⁴⁶, A. Noble⁴⁷, V. Oreiller^{48, 8}, N. Padoy^{49, 50}, S. Pati^{51, 11, 12, 52}, K. Payette^{53, 54}, T. Rädtsch^{1, 2}, J. Rafael-Patiño^{31, 32}, V. Singh Bawa⁵⁵, S. Speidel^{16, 56}, C. H. Sudre^{57, 58, 59, 60}, K. van Wijnen²⁰, M. Wagner¹⁹, D. Wei⁶¹, A. Yamlahi¹, M. H. Yap⁶², C. Yuan¹³, M. Zenk^{5, 63}, A. Zia⁶⁴, D. Zimmerer^{5, 2}, D. Aydogan^{65, 66}, B. Bhattarai⁶⁷, L. Bloch^{68, 69, 70}, R. Brüngel^{68, 69, 70}, J. Cho⁷¹, C. Choi⁷², Q. Dou⁷³, I. Ezhov⁷⁴, C. M. Friedrich^{68, 69}, C. Fuller⁷⁵, R. R. Gaire⁷⁶, A. Galdran^{77, 78}, Á. García Faura⁷⁹, M. Grammatikopoulou⁸⁰, S. Hong⁸¹, M. Jahanifar⁸², I. Jang^{83, 84, 85}, A. Kadkhodamohammadi⁸⁰, I. Kang⁷¹, F. Kofler^{86, 87, 88, 89}, S. Kondo⁹⁰, H. Kuijff⁹¹, M. Li⁹², M. Luu⁹³, T. Martinčič⁷⁹, P. Morais⁹⁴, M. A. Naser⁷⁵, B. Oliveira^{94, 95, 96}, D. Owen⁸⁰, S. Pang⁹⁷, J. Park⁷¹, S. Park⁹³, S. Płotka^{98, 99, 100}, E. Puybareau¹⁰¹, N. Rajpoot⁸², K. Ryu¹⁰², N. Saeed¹⁰³, A. Shephard⁸², P. Shi¹⁰⁴, D. Štepec^{79, 105}, R. Subedi⁷⁶, G. Tochon¹⁰¹, H. R. Torres^{94, 95, 96}, H. Urien¹⁰⁶, J. L. Vilça⁹⁴, K. A. Wahid⁷⁵, H. Wang¹⁰⁷, J. Wang¹⁰⁷, L. Wang¹⁰⁷, X. Wang¹⁰⁸, B. Wiestler¹⁰⁹, M. Wodzinski^{110, 111}, F. Xia^{112, 113}, J. Xie¹¹⁴, Z. Xiong⁹², S. Yang¹¹⁵, Y. Yang¹⁰⁴, Z. Zhao¹¹³, K. Maier-Hein^{5, 2, 116}, P. F. Jäger^{117, 2}, A. Kopp-Schneider⁴, and L. Maier-Hein^{1, 2, 3, 63, 33}

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 multi-center 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 method 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.

Data from all ISBI and MICCAI 2021 competitions (n = 80)

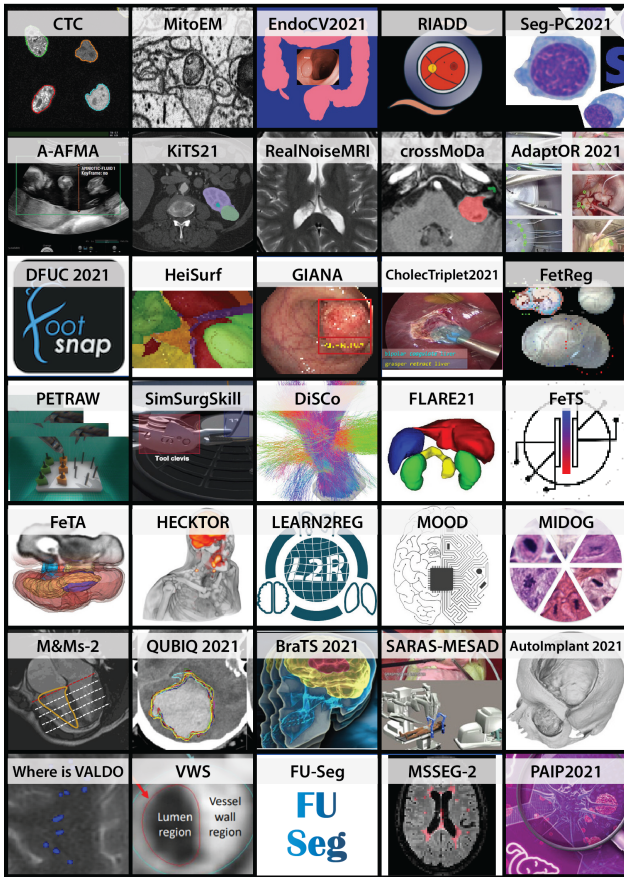


Figure 1. Overview of the IEEE ISBI 2021 and MICCAI 2021 challenges. Under the umbrella of 35 challenges (each represented by a teaser image and acronym), a total of 80 competitions with dedicated leaderboards were organized, as detailed in Suppl. A-C. We used data from participants, organizers, and 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 in-

vested 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 initiative and to bring us into contact with participants (if allowed by the challenge privacy policy) or distribute the survey link to them. We created an individual survey website for each challenge to be able to accommodate the individual challenge submission deadline. To avoid bias in survey responses, participants were asked to complete the survey before knowing their position in the final ranking. Out of a maximum of 168 questions, the survey only showed questions that were relevant to the specific situation. The responses and feedback from the IEEE ISBI 2021 respondents were used to refine the survey for MICCAI 2021, and are thus not included in the results presented in Sec. 3.1.

Where organizers were allowed to share the contact details of the participants (20 challenges), the survey was conducted in closed-access mode, meaning that the participants received individual links to the survey and, where necessary, reminders. Fifteen surveys were conducted in open-access mode, meaning that the organizers were tasked with sharing the link to the respective survey and sending reminders. In these cases, we were not informed about the number of challenge participants and could not relate the number of responses to the total number.

2.2. RQ2: Do current competitions generate scientific progress?

The focus of the organizer survey was on the findings of the respective competition, particularly regarding whether scientific progress was made and, if yes, in which areas it

was achieved and which open questions remain. To better put the respective competition into context, we also acquired general information on the associated competition(s).

2.3. RQ3: Which strategies characterize challenge winners?

The complexity of state-of-the-art neural network-based approaches, involving numerous and interdependent design parameters, comes with the risk of attributing the success in a competition to the wrong component of a system. To approach the question *Why is the winner the best?*, we linked the survey results of Sec. 2.1 to the final outcome of the competition and subsequently applied mixed model analyses. Given the large number of parameters relative to the number of competitions, we were aware that differences in parameters might not achieve statistical significance. In a second step, we therefore explicitly asked challenge winners for successful algorithm design choices and strategies in an additional survey.

Mixed model analysis To compensate for the hierarchical data structure resulting from clusters corresponding to specific competitions, a logistic mixed model was used. In a first step, a univariable analysis was performed, i.e., the effect of each variable on the ranking was investigated separately. To further account for potential interdependencies between variables, two multivariable analyses were added. In the first analysis, the goal was to investigate the strategies influencing the probability of being the winner, while the second analysis focused on evaluating the strategies influencing the probability of being ranked among the best 30%. For both analyses, a logistic mixed model was implemented. The winning strategies were included as fixed effects while the challenge identifier was included as a random effect. Additionally, some of the strategies were allowed to vary across challenges, specifically the total training time in computation hours, time spent on analyzing data and annotations, and time spent on analysis of failure cases. Variables with highly varying magnitudes were scaled before fitting the model. Statistical analysis was done in R Statistical Software [38] (v4.0.3, package: lme4 [7]).

Survey on winning strategies 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

(per z-slice approach) with postprocessing (18%). 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 ensembled 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 normal-

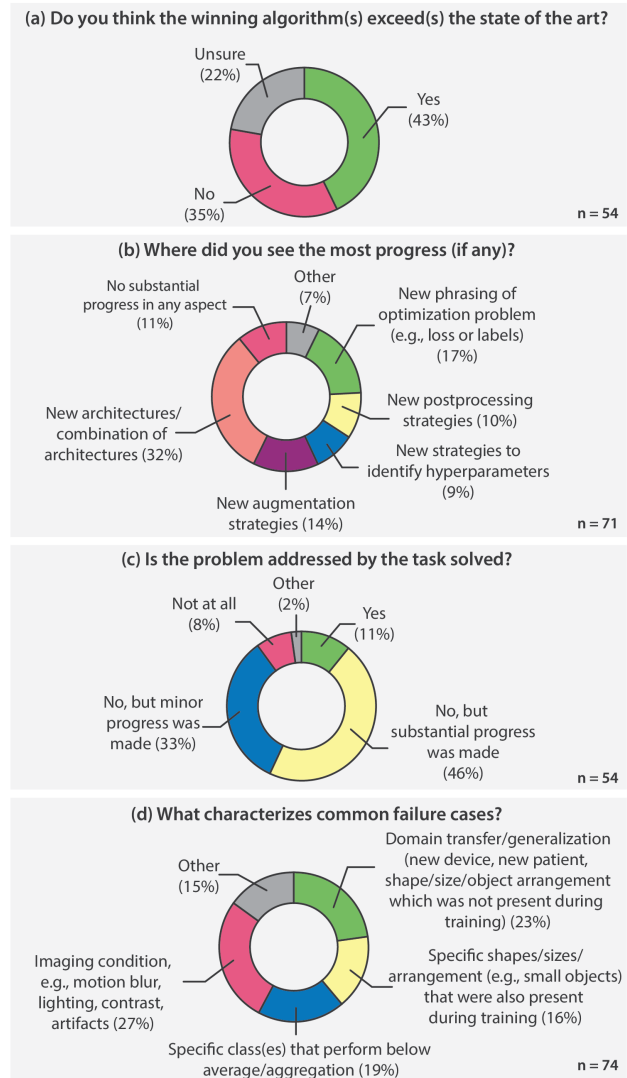


Figure 2. Key insights provided by the organizers of IEEE ISBI 2021 and MICCAI 2021 challenges.

ized 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 hyped 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

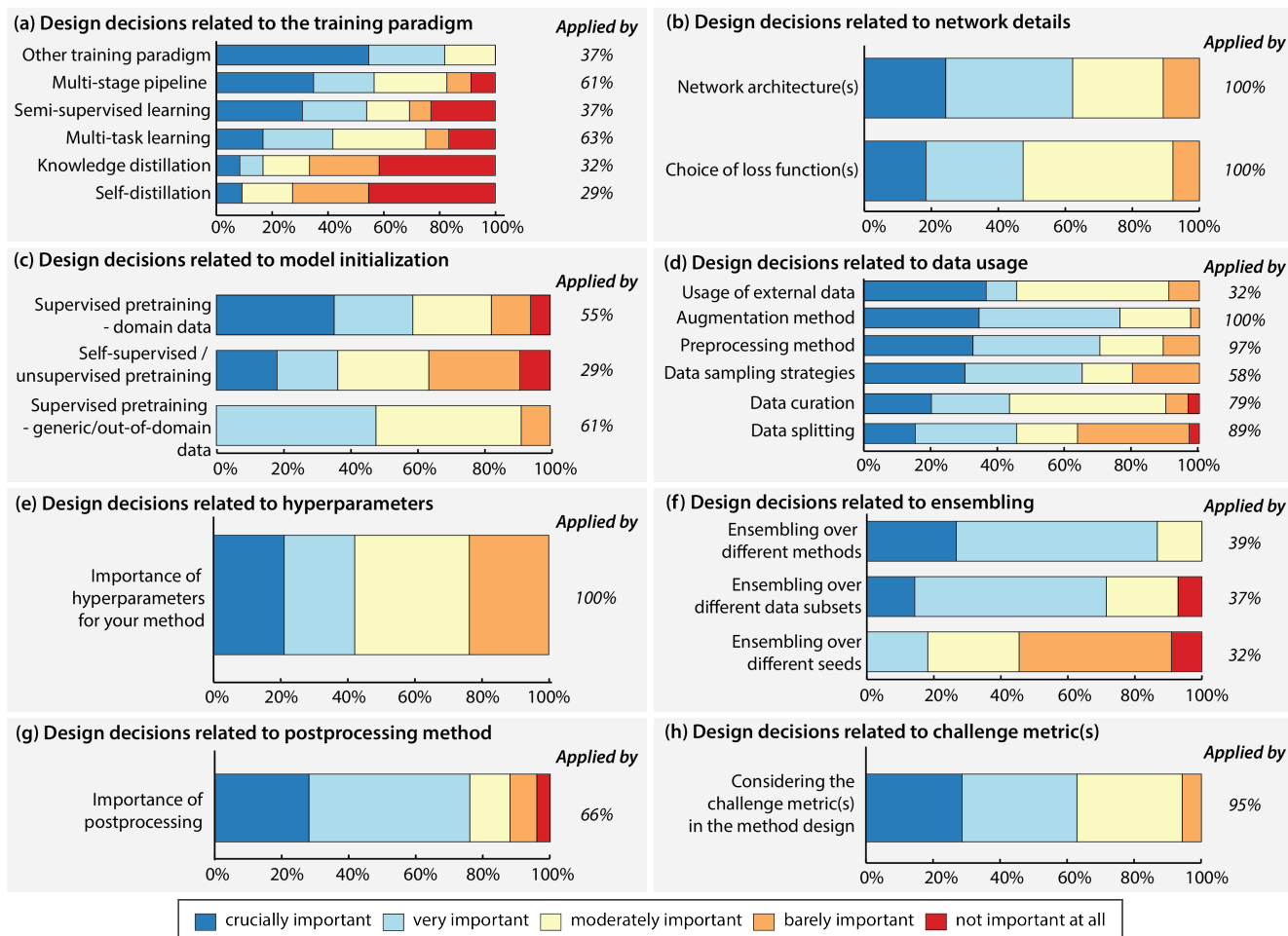


Figure 3. Importance of design decisions for the neural network-based winning submission of the respective IEEE ISBI 2021 and MICCAI 2021 competition rated by the (team) lead and ordered by percentage of highest vote (crucially important: dark blue). Voting was only conducted among those who used the 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 infor-*

mation 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

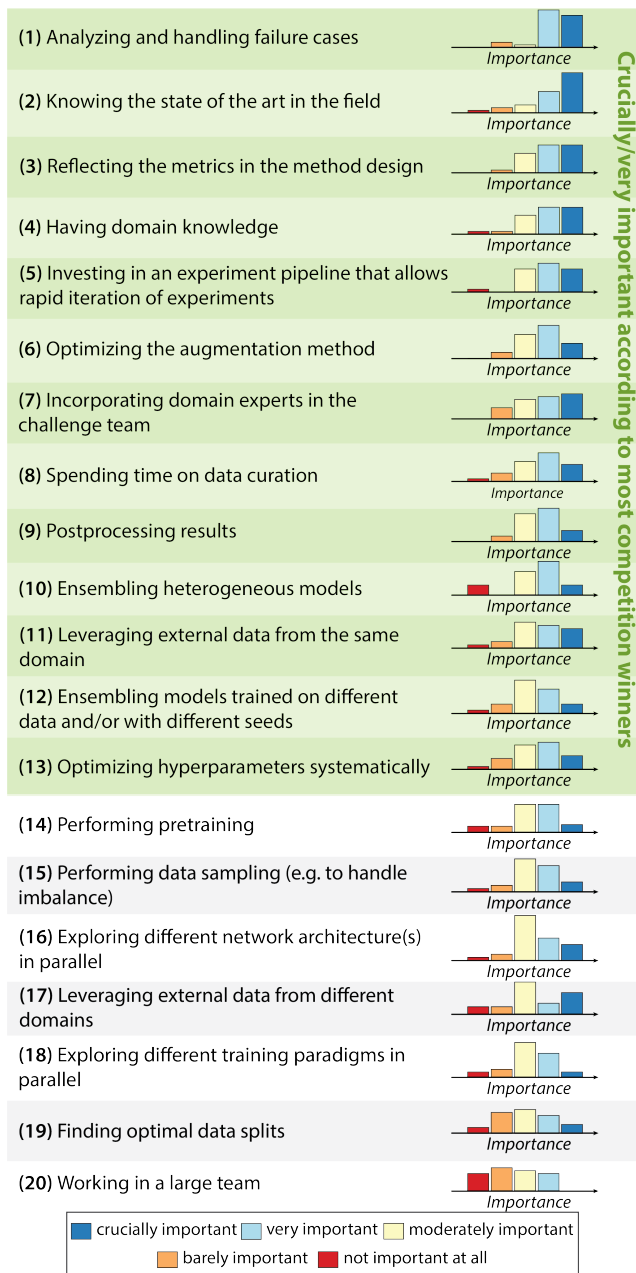


Figure 4. Strategies for winning a challenge according to winners of IEEE ISBI 2021 and MICCAI 2021 competitions, ordered by the sum of the “crucially important” (dark blue) and “very important” (light blue) categories. The distribution of importance (from left to right: not important at all, barely important, moderately important, very important, crucially important) is depicted for each strategy.

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; ⁹Technische Hochschule Ingolstadt, Ingolstadt, Germany; ¹⁰Center for Artificial Intelligence and Data Science for Integrated Diagnostics (AI²D) and Center for Biomedical Image 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, Heidel-

berg, 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 (LTS5), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland; ³³National Center for Tumor Diseases (NCT), Heidelberg, Germany; ³⁴SBILab, Department of ECE, IIIT-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; ⁴⁵Departament de Matemàtiques & Informàtica, 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 (AI²D) 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 Brookes 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 Am-

sterdam, 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

- [1] A-AFMA. A-AFMA - Grand Challenge. <https://a-afma.grand-challenge.org/>. Accessed: 2022-11-11. **3**
- [2] Sharib Ali, Noha Ghatwary, Debesh Jha, Ece Isik-Polat, Gorkem Polat, Chen Yang, Wuyang Li, Adrian Galdran, Miguel-Ángel González Ballester, Vajira Thambawita, et al. Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. *arXiv preprint arXiv:2202.12031*, 2022. **3**
- [3] Vincent Andrearczyk, Valentin Oreiller, Martin Vallières, Mathieu Hatt, Catherine Cheze-Le Rest, Dimitris Visvikis, Mario Jreige, Hesham Elhalawani, Sarah Boughdad, John O. Prior, and Adrien Depeursinge. HEad and neCK TumOR segmentation and outcome prediction in PET/CT images, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4573155>. **3**
- [4] Marc Aubreville, Christof Bertram, Mitko Veta, Robert Klopffleisch, Nikolas Stathonikos, Katharina Breininger, Natalie ter Hoeve, Francesco Ciompi, and Andreas Maier. Mitosis domain generalization challenge, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4573978>. **3**
- [5] Spyridon Bakas, Christopher Carr, Adam Flanders, Jayashree Kalpathy-Cramer, John Mongan, Bjoern Menze, and Luciano M Prevedello. RSNA-MICCAI Brain Tumor Segmentation (BraTS) Challenge 2021, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4575162>. **3**
- [6] Spyridon Bakas, Micah Sheller, Sarthak Pati, Brandon Edwards, G. Anthony Reina, Ujjwal Baid, Yong Chen, Rus-

- sell (Taki) Shinohara, Jason Martin, Bjoern Menze, and other. Federated tumor segmentation, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4573128>. 3
- [7] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. 3
- [8] Frédéric Cervenansky, Olivier Commowick, François Cotton, Michel Dojat, and Gilles Edan). Multiple sclerosis new lesions segmentation challenge, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4575409>. 3
- [9] Nicolas Chenouard, Ihor Smal, Fabrice De Chaumont, Martin Maška, Ivo F. Sbalzarini, Yuanhao Gong, Janick Cardinale, Craig Carthel, Stefano Coraluppi, Mark Winter, et al. Objective comparison of particle tracking methods. *Nature Methods*, 11(3):281–289, 2014. 2
- [10] Jinwook Choi, Kyoungbun Lee, Won-Ki Jeong, and Se Young Chun. PAIP2021: Perineural Invasion in Multiple Organ Cancer (Colon, Prostate, and Pancreatobiliary tract), Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4575424>. 3
- [11] Michael Coughlan, Patricia Cronin, and Frances Ryan. Survey research: Process and limitations. *International Journal of Therapy and Rehabilitation*, 16(1):9–15, 2009. 8
- [12] CTC. 6th Cell Tracking Challenge Edition at ISBI 2021 – Cell Tracking Challenge. <http://celltrackingchallenge.net/ctc-vi/>. Accessed: 2022-11-11. 3
- [13] Fabio Cuzzolin, Vivek Singh Bawa, Inna Skarga-Bandurova, Mohamed Mohamed, Jackson Ravindran Charles, Elettra Oleari, Alice Leporini, Carmela Landolfo, Armando Stabile, Francesco Setti, and Riccardo Muradore. SARAS challenge on Multi-domain Endoscopic Surgeon Action Detection, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4575197>. 3
- [14] Reuben Dorent, Aaron Kujawa, Jonathan Shapey, Samuel Joutard, Jorge Cardoso, Marc Modat, Nicola Rieke, Ben Glocker, Spyridon Bakas, and Tom Vercauteren. Cross-Modality Domain Adaptation for Medical Image Segmentation, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4573119>. 3
- [15] Matthias Eisenmann, Annika Reinke, Vivienne Weru, Minu D. 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. 4
- [16] Sandy Engelhardt, Anirban Mukhopadhyay, Raffaele De Simone, Lalith Sharan, Antonia Stern, Julian Brand, and Henry Krumb. Deep Generative Model Challenge for Domain Adaptation in Surgery 2021, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4646979>. 3
- [17] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nature Methods*, 16(1):67–70, Jan. 2019. 5
- [18] Melanie Ganz and Hannah Eichhorn. Brain MRI reconstruction challenge with realistic noise, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4572640>. 3
- [19] Gabriel Girard, Emmanuel Caruyer, Jonathan Rafael-Patino, Marco Pizzolato, Raphaël Truffet, and Jean-Philippe Thiran. Diffusion-simulated connectivity challenge, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4733450>. 3
- [20] Mark J. Gooding, Annamarie J. Smith, Maira Tariq, Paul Aljabar, Devis Peressutti, Judith van der Stoep, Bart Reymen, Daisy Emans, Djoya Hattu, Judith van Loon, et al. Comparative evaluation of autocontouring in clinical practice: a practical method using the turing test. *Medical Physics*, 45(11):5105–5115, 2018. 7
- [21] Mattias Heinrich, Adrian Dalca, Lasse Hansen, Alessa Hering, Bennett Landman, Keelin Murphy, and Bram van Ginneken. Learn2reg - the challenge (2021), Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4573968>. 3
- [22] Nicholas Heller, Nikolaos Papanikolopoulos, and Christopher Weight. 2021 Kidney and Kidney Tumor Segmentation Challenge, Mar. 2020. Zenodo. <https://doi.org/10.5281/zenodo.4674397>. 3
- [23] Kaggle Inc. Kaggle: Your Machine Learning and Data Science Community. <https://www.kaggle.com/>. Accessed: 2022-11-11. 2
- [24] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, Feb. 2021. 5
- [25] Florian Kofler, Ivan Ezhov, Fabian Isensee, Christoph Berger, Maximilian Korner, Johannes Paetzold, Hongwei Li, Suprosanna Shit, Richard McKinley, Spyridon Bakas, et al. Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient. *arXiv preprint arXiv:2103.06205v1*, 2021. 7
- [26] Jianning Li, Oldřich Kodym, David G. Ellis, Michal Španl, Michele R. Aizenberg, Victor Alves, Gord von Campe, and Jan Egger. Towards the Automatization of Cranial Implant Design in Cranioplasty: 2nd MICCAI Challenge on Automatic Cranial Implant Design, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4577269>. 3
- [27] Jun Ma, Song Gu, Yao Zhang, Xingle An, Cheng Zhu, Congcong Wang, Cheng Ge, Liwen Zou, Qiongjie Zhu, Guoqiang Dong, Jian He, and Xiaoping Yang. Fast and Low GPU Memory Abdominal Organ Segmentation in CT, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4596561>. 3
- [28] Klaus H. Maier-Hein, Peter F. Neher, Jean-Christophe Houde, Marc-Alexandre Côté, Eleftherios Garyfallidis, Jidan Zhong, Maxime Chamberland, Fang-Cheng Yeh, Ying-Chia Lin, Qing Ji, et al. The challenge of mapping the human connectome based on diffusion tractography. *Nature Communications*, 8(1):1–13, 2017. 2
- [29] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, et al. Why rankings of biomedical image analysis competitions

- should be interpreted with care. *Nature Communications*, 9(1):5217, 2018. 2, 8
- [30] Lena Maier-Hein, Annika Reinke, Evangelia Christodoulou, Ben Glocker, Patrick Godau, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. *arXiv preprint arXiv:2206.01653*, 2022. 7
- [31] Lena Maier-Hein, Annika Reinke, Michal Kozubek, Anne L. Martel, Tal Arbel, Matthias Eisenmann, Allan Hanbury, Pierre Jannin, Henning Müller, Sinan Onogur, Julio Saez-Rodriguez, Bram van Ginneken, Annette Kopp-Schneider, and Bennett A. Landman. BIAS: Transparent reporting of biomedical image analysis challenges. *Medical Image Analysis*, 66:101796, Dec. 2020. 2
- [32] Carlos Martín-Isla, José F. Rodríguez Palomares, Andrea Guala, Sergio Escalera, and Karim Lekadir. Multi-Disease, Multi-View & Multi-Center Right Ventricular Segmentation in Cardiac MRI (M&Ms-2), Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4573984>. 3
- [33] Bjoern Menze, Leo Joskowicz, Spyridon Bakas, Andras Jakab, Ender Konukoglu, Anton Becker, Amber Simpson, and Richard Do. Quantification of Uncertainties in Biomedical Image Quantification 2021, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4575204>. 3
- [34] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014. 2
- [35] MitoEM. MitoEM - Grand Challenge. <https://mitoem.grand-challenge.org/>. Accessed: 2022-11-11. 3
- [36] Kelly Payette, Priscille de Dumast, Andras Jakab, Meritxell Bach Cuadra, Lana Vasung, Roxane Licandro, Bjoern Menze, and Hongwei Li. Fetal Brain Tissue Annotation and Segmentation Challenge, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4573144>. 3
- [37] Jens Petersen, Gregor Köhler, Paul Jäger, Peter Full, David Zimmerer, Klaus Maier-Hein, Tobias Roß, Tim Adler, Annika Reinke, and Lena Maier-Hein. Medical Out-of-Distribution Analysis Challenge 2021, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4573948>. 3
- [38] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. 3
- [39] Annika Reinke, Matthias Eisenmann, Minu D. Tizabi, Carole H. Sudre, Tim Rädtsch, Michela Antonelli, Tal Arbel, Spyridon Bakas, M. Jorge Cardoso, Veronika Cheplygina, et al. Common limitations of image processing metrics: A picture story. *arXiv preprint arXiv:2104.05642*, 2021. 7
- [40] RIADD. RIADD (ISBI-2021) - Grand Challenge. <https://riadd.grand-challenge.org/>. Accessed: 2022-11-11. 3
- [41] Daniel Sage, Hagai Kirshner, Thomas Pengo, Nico Sturman, Junhong Min, Suliana Manley, and Michael Unser. Quantitative evaluation of software packages for single-molecule localization microscopy. *Nature Methods*, 12(8):717–724, 2015. 2
- [42] SegPC-2021. SegPC-2021 - Grand Challenge. <https://segpc-2021.grand-challenge.org/SegPC-2021/>. Accessed: 2022-11-11. 3
- [43] Stefanie Speidel, Lena Maier-Hein, Danail Stoyanov, Sebastian Bodenstedt, Martin Wagner, Beat Müller, Jonathan Chen, Benjamin Müller, Franziska Mathis-Ullrich, Paul Scheikl, et al. Endoscopic vision challenge 2021, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4572973>. 3
- [44] Carole Sudre, Kimberlin van Wijnen, Marius Groot, Florian Dubost, and Marleen de Bruijne. Vascular lesions detection, Mar. 2020. Zenodo. <https://doi.org/10.5281/zenodo.4600654>. 3
- [45] Christoph Tauchert, Peter Buxmann, and Jannis Lambinus. Crowdsourcing data science: A qualitative analysis of organizations’ usage of kaggle competitions. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020. 7
- [46] Vladimír Ulman, Martin Maška, Klas EG Magnusson, Olaf Ronneberger, Carsten Haubold, Nathalie Harder, Pavel Matula, Petr Matula, David Svoboda, Miroslav Radojevic, et al. An objective comparison of cell-tracking algorithms. *Nature Methods*, 14(12):1141–1152, 2017. 2
- [47] Femke Vaassen, Colien Hazelaar, Ana Vaniqui, Mark Gooding, Brent van der Heyden, Richard Canters, and Wouter van Elmpt. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology*, 13:1–6, 2020. 7
- [48] Chuanbo Wang, Behrouz Rostami, Jeffrey Niezgoda, Sandeep Gopalakrishnan, and Zeyun Yu. Foot ulcer segmentation challenge 2021, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4575314>. 3
- [49] Manuel Wiesenfarth, Annika Reinke, Bennett A. Landman, Matthias Eisenmann, Laura Aguilera Saiz, M. Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. Methods and open-source toolkit for analyzing and visualizing challenge results. *Scientific Reports*, 11(1):1–15, Jan. 2021. 2
- [50] Moi Hoon Yap, Neil Reeves, Andrew Boulton, Satyan Rajbhandari, David Armstrong, Arun G. Maiya, Bijan Najafi, Eibe Frank, and Justina Wu. Diabetic foot ulcers grand challenge 2021, Mar. 2020. Zenodo. <https://doi.org/10.5281/zenodo.4646982>. 3
- [51] Chun Yuan, Li Chen, Niranjana Balu, Mahmud Mossa-Basha, Jenq-Neng, David Saloner, and Peter Douglas. Carotid vessel wall segmentation challenge, Mar. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4575301>. 3