Nicolai Spicher, Tim Wesemeyer and Thomas M. Deserno*

Crowdsourcing image segmentation for deep learning: integrated platform for citizen science, paid microtask, and gamification

https://doi.org/10.1515/bmt-2023-0148
Received April 11, 2023; accepted November 30, 2023;
published online December 26, 2023

Abstract

Objectives: Segmentation is crucial in medical imaging. Deep learning based on convolutional neural networks showed promising results. However, the absence of large-scale datasets and a high degree of inter- and intra-observer variations pose a bottleneck. Crowdsourcing might be an alternative, as many non-experts provide references. We aim to compare different types of crowdsourcing for medical image segmentation.
Methods: We develop a crowdsourcing platform that integrates citizen science (incentive: participating in the research), paid microtask (incentive: financial reward), and gamification (incentive: entertainment). For evaluation, we choose the use case of sclera segmentation in fundus images as a proof-of-concept and analyze the accuracy of crowdsourced masks and the generalization of learning models trained with crowdsourced masks.
Results: The developed platform is suited for the different types of crowdsourcing and offers an easy and intuitive way to implement crowdsourcing studies. Regarding the proof-of-concept study, citizen science, paid microtask, and gamification yield a median F-score of 82.2, 69.4, and 69.3 % compared to expert-labeled ground truth, respectively. Generating consensus masks improves the gamification masks (78.3 %). Despite the small training data (50 images), deep learning reaches median F-scores of 80.0, 73.5, and 76.5 % for citizen science, paid microtask, and gamification, respectively, indicating sufficient generalizability.
Conclusions: As the platform has proven useful, we aim to make it available as open-source software for other researchers.

Keywords: crowdsourcing; image segmentation; deep learning; platform

Introduction

Image segmentation is the process of detecting (interconnected) regions with similar properties. This involves localization of the region, its delineation, and if multiple objects are within the image, the assignment of labels. The output is a mask in which each pixel is assigned to a label or to the background. Segmentation plays a vital role in many applications in medical imaging [1, 2].

Recently, deep learning (DL) has shifted the segmentation paradigm [3, 4]. DL outperforms conventional machine learning (ML) using hand-crafted features and parameters based on domain knowledge, as it considers training data for automatically computing a model that describes the relationship between input (image) and output (mask) in an end-to-end fashion. For example, convolutional neural networks (CNN) such as the U-Net [5] are used in many medical imaging applications [6].

The accuracy of DL increases with the quantity and quality of training data. For supervised learning, the training images need additional ground truth annotations, which are tedious and time-consuming to generate. Annotated databases for everyday images are freely available but for medical imaging, large and carefully expert-labelled datasets are rare [7]. In addition, many of the available datasets, such as the ones used in scientific challenges, e.g. the MICCAI challenges, can rather be assumed to be “silver standards” as they do not require the same strict criteria as gold standards [8]. Ker et al. consider the lack of ground truth as the main bottleneck for DL-applications in medical imaging [9].

Domain experts bear another unsolved problem, as they yield inter- and intra-observer variability. Therefore, expert-trained models often cannot generalize to real-world data. To account for intra-observer variability, more experts can create more annotations, but this increases

***Corresponding author: Prof. Dr. Thomas M. Deserno**, FPSIE, FIAHSI, Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Lower Saxony, Germany, Phone: +49 531 391 2130, E-mail: thomas.deserno@plri.de. https://orcid.org/0000-0003-3492-4407
Nicolai Spicher and Tim Wesemeyer, Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Lower Saxony, Germany. https://orcid.org/0000-0002-2879-9948 (N. Spicher)

costs and efforts. Furthermore, a consensus must be created from multiple annotations. Algorithms like the simultaneous truth and performance level estimation (STAPLE) are useful [10]. Based on expectation-maximization, STAPLE computes a consensus mask and improves generalization [11].

In previous work, we investigated the influence of training data characteristics on segmentation accuracy [12]. We segmented the sclera in photographs of the human eye which is an important preprocessing step in biometrics [13], medical research [14], diagnostics [15], and clinical trials [16]. We analyzed the total number of trainings images vs. the number of masks per image [12]. According to our results, more images with fewer masks yield better results but a minimum of three masks is required to obtain adequate STAPLE results.

In 2006, the term "crowdsourcing" was coined with promising results in various fields of research [17]. There are multiple definitions of crowdsourcing in the literature, which usually involve a large group of individuals that are connected via the internet and perform an open task proposed by an imitator and the fulfillment of the task results in a benefit which might be monetary but might also be social recognition or self-esteem [18]. Motivation is a key factor for crowdsourcing and depending on the incentive, different methods can be defined:

- Citizen science motivates by participating in science [19],
- Paid microtask by receiving payment [20], and
- Gamification by entertainment [21].

Typical crowdsourcing collects or annotates data for research (e.g., image classification). Recently, crowdsourcing of image segmentation has been comprehensively reviewed [22]. For example, in 2021, Bafti et al. presented a crowdsourcing platform for the semi-automatic image segmentation in cell biology [23], and Marzahl et al. introduced a collaboration toolset for local image annotations [24]. In previous work, we build an internet platform for crowdsourcing image segmentation provided by citizen scientists [25]. However, these approaches do not compare different types of motivation. So far, the accuracy of paid microtask or gamification has not been directly compared to citizen science in the medical domain.

In this work, we present a novel platform for crowdsourcing that offers different crowdsourcing methods as no "out-of-the-box" software is available for this task. As a proof-of-concept study, we segment the sclera in photographs of the human eye. For each crowdsourcing method, we quantitatively analyze the accuracy of the acquired masks and associated costs and compare the different methods among each other. We aim to answer the following questions:

- *Platform: Is the platform suitable for integrating different crowdsourcing methods?*
- *Proof-of-concept-study: How accurate are the crowdsourced masks compared to expert delineations and how suitable are they for training of a DL model?*

Related work

Citizen science has a long history in medical imaging. Albarqouni et al. developed a web platform for mitosis detection in breast cancer histology images that they used for DL training [26]. Regarding image segmentation, Maier-Hein et al. used endoscopy data from laparoscopic surgeries [27], and Grote et al. used microscopy images annotated by third-year medical students [28]. All works report the feasibility of citizen science for medical image segmentation but differ in the reported quality.

The concept of gamification was introduced approximately a decade ago and found applications in education, advertising, and e-health [29]. To make non-game activities more entertaining, gamification adds elements such as [30, 31]: (i) points rewarding successfully finishing a task, (ii) leaderboards ranking gamers by their points, or (iii) badges visualizing important achievements. In addition, messages or voice communications enable direct interconnection between the gamers. In medical applications, for instance, Balducci et al. proposed games for skin lesion analysis in dermatology [32]; Ionescu et al. suggested a game for optimizing computer-aided detection in mammography [33], and Mavandadi et al. developed a game with red blood cell images for malaria diagnosis [34]. Arganda-Carreras et al. organized a challenge on electron microscopic images of the brain [35]. The medical data donors project (https://www.medicaldatadonors.org/) offers a variety of games for different medical images. Recently, major studios (Gearbox Studio Québec, CCP Games) integrated crowdsourcing games into their commercial video games [36, 37].

Paid microtask require a platform handling the financial rewards, such as Amazon's Mechanical Turk (MTurk) (https://www.mturk.com/). A requester defines a human intelligence task (HIT) associated with a description, qualification constraint, and payment. All HITs are offered to "paid crowdworkers", who freely decide which HIT to perform. Gurari et al. used a library of biological and biomedical images, which were segmented by MTurk paid crowdworkers [38]. Sharma et al. used the crowd for segmenting of chromosomes in microscopic cell spreads [39]. Heim et al. analyzed liver segmentation in abdominal CT [40] and Cheplygina et al. airway annotation in chest CT [41]. Regarding images of the eye, Mitry et al. performed different

studies with retinal images for detecting clinically relevant features that can also be detected by ophthalmologically naive individuals [42–44]. All authors report principle feasibility, but Heim et al., for instance, reported up to 30 % of unusable segmentations [40].

Recruitment and motivation is a major problem of all types of crowdsourcing [45]. However, only a few attempts have been made to combine incentives: Feyisetan et al. improved paid microtasks by gamification: they added game levels and a leaderboard [46]. Bowser et al. used mobile apps to recruit citizen scientists [47]. Tinati et al. integrated a real-time chat and activity feeds to further motivate the users [48].

Materials and methods

Requirements of the platform

We aim to develop a platform integrating three types of crowdsourcing methods for image segmentation, addressing (i) citizen scientists, (ii) paid crowdworkers, and (iii) gamers. The platform is intended to be task-oriented with a task consisting of a set of images, an informative text about the task, the number of desired masks per image, and one or multiple types of crowdsourcing methods that should be used to acquire the masks. The three different users groups should interact in different ways with the platform.

The admin should be able to define a task and its parameters and to import images which are then stored in a database. There needs to be a way to monitor or abort running tasks. After cancellation or completion of a task, the generated masks should be available for downloaded. Additionally, an interface to MTurk is required which allows to (i) see all running human intelligence tasks (HITs) on the MTurk platform, (ii) analyze their status (number of pending or performed segmentation), or (iii) cancel a running HIT.

A citizen scientist should be able to register at the platform and after login, a personal dashboard shall be offered. The dashboard should show the list of available tasks as well as graphical elements indicating performance, i.e., the activity in the past months and a user level, which is based on the number of created masks.

A paid crowdworker should not directly interact with the platform but only via the Amazon MTurk platform. For that, the platform needs to be able to transfers all parameters of a task to MTurk, making it accessible as a HIT on Amazon MTurk. A Gamer should also not directly interact with the platform with all data exchange taking place in the background between the game and the platform.

In addition, the platform requires several background services, e.g. for regular backups and for mask processing, e.g. the STAPLE algorithm to combine several masks into an aggregated one.

Crowd acquisition

Our previous work [12] indicates that for sclera segmentation, three to five masks with STAPLE yield adequate consensus quality. To generate four masks for each image, we acquire data from the crowd with different strategies:

- Citizen scientists: We invite volunteers (students, trainees, other university members) per e-mail to contribute to our research project voluntarily without offering financial reward. We ask each participant to generate a single mask for each image and randomly pick four masks for each image.
- Paid crowdworkers: For each image, we define a HIT with four assignments. Thereby, four workers generate a mask. We pay 0.10$ per mask to all workers of any qualification who fulfill the HIT within 2 min.
- Gamers: We invite volunteers (students, trainees, other university members) per e-mail and ask if they are interested in playing a new game. If they reply positive, we schedule an appointment with 4–6 volunteers and meet using a free video conference tool. In that meeting, we present the game to the gamers and ask them to play. The video conference system allows the gamers to communicate with each other. We organized four appointments with different gamers.

Deep learning architecture

We use acquired masks to train a UNet as UNets can cope with low numbers of training data [5]. For implementation, we use a Python API [49], which is based on PyTorch and related libraries (torch==1.8.1, torchvision==0.9.1, pillow==8.2.0, scipy==1.6.2).

We apply a VGG-16 pre-trained model, which was initially proposed by the Visual Geometry Group (VGG) of Oxford University [50] and adjust the learning rates using the Adam optimizer [51] and a binary cross entropy loss function. Initial learning rates are 10^{-4} and 10^{-6} for the decoder and encoder, respectively. We kept the number of epochs dynamically to account for early stopping with a maximum of 20 epochs. If the validation error does not decrease more than a threshold of 0.0008 for three epochs we lower the decoder learning rate by a factor of 10. If for another three epochs there is again no improvement, we stop the training. To increase generalizability, we train four randomly initialized models with random image order. During training, we store the entropy loss and the F-scores.

We scaled all images and masks to 640×427 pixels and applied data augmentation using a Python framework (albumentations==0.5.2) with global operators scaling, rotation, and color transform. We apply an 80 %/20 % training/validation split and run all experiments on an out-of-the-shelf computer with GPU (NVIDIA GTX1070 GPU, CUDA runtime: 10.1).

Evaluation methodology

To answer our research question in how far the acquired masks are feasible, we perform different evaluations. In order to compare the accuracy of the masks with the ground truth, we use the Boundary F-score. A mask (I) is compared to the corresponding Ground Truth mask (I_{GT}) using precision p and recall r combined into a single metric

$$\mathrm{F-score}(I, I_{GT}) = 2\frac{p(I, I_{GT})r(I, I_{GT})}{p(I, I_{GT}) + r(I, I_{GT})}$$

We used this metric as it is the *de facto* standard method in sclera segmentation and allows to compare our results to other works. First, we compute the F-score of all individual masks. Second, to account for intra-observer variability, we use STAPLE to generate consensus masks for each image and each crowdsourcing method and compute F-score of

the STAPLE masks. Third, we use the STAPLE masks directly for DL training. Subsequently, we predict the masks of an unseen test data and compute the F-score between the predicted masks and the ground truth.

Additionally, we analyze the time it takes to acquire the masks. As the gaming sessions needed to be organized, we exclude gamification from this analysis, hence i∈{cs,pm} denoting citizen science (cs) and paid microtask (pm). We store a time stamp when a citizen scientist or paid crowdworker saved a mask on the platform. Assuming that citizen scientists work on the images consecutively, we can assess the effort per mask. Contrarily, paid crowdworker work in parallel such that the individual efforts remain unclear to us. Furthermore, we are interested in the total time T_i to establish the reference set. We derive it from the waiting time W_i between task creation and the first mask being submitted and the duration D_i between the first and last mask being submitted

$$T_i = W_i + D_i$$

Dataset

We compose the dataset of 100 photographs of the human eye by randomly selecting 50 images from the public Sclera Blood Vessels, Periocular and Iris (SBVPI) dataset [13] and 50 images from our private dataset from previous work [14]. The high-resolution RGB images have 3,000 × 1,700 pixels and 2,992 × 2,000 pixels, respectively. The healthy subjects look into four different view directions (up, down, left, right). Further information on the complexity of sclera segmentation are reported in [14] which also gives information on the interobserver differences, underlining the complexity of the sclera segmentation task.

We randomly split the data into 50 images for training and 50 for testing. Crowdsourcing is used to generate masks for the 50 training images only.

Gamification

Inspired by the crowdsourcing game “Dr. Detective” [52], we developed the multiplayer game “Dr. Columbus” that features all typical components of a client-server architecture. The game concept is proposed in [53].

The client is available for Windows, and Linux and offers a graphical user interface (GUI). On first use, the gamer creates a personal account with associated score, rank, and badge. The server stores all other data (e.g. user information) and is connected to our crowdsourcing platform to receive the images and store the generated masks. Multiple gamers perform a game round, which lasts 200 s. At first, each gamer chooses a starting position on the sclera. Similar to the concept of “Snake”, the gamer needs to increase the territory without being cut by any opponent [53]. Figure 1 visualizes some stages. The remaining time is indicated on the upper left part of the GUI.

The gamers receive positive and negative points according to their area size within and outside the sclera, respectively. We use an expert annotation as Ground Truth: A true positive pixel covering the sclera increases the gamer score, a false positive pixel outside the sclera, reduces it. Based on the points, the system continuously updates the leadership board. We motivate the gamers with popups, in-game music, and sound effects [53]. A video demonstrating the key concepts is publicly-available: https://www.youtube.com/watch?v=tyB-yFzRvZM.

The mask generated by a gamer in one game round is incomplete. Therefore, we combine the masks of all gamers into a consensus mask, i.e., each pixel covered by a gamer in a round is assumed to be part of the sclera.

Results

Platform

Figure 2 displays the developed architecture of the platform: boxes represent logical units of the platform which are connected via interfaces that are visualized as arrows. Solid arrows indicate direct access of a user group to a functionality, e.g., citizen scientists access directly their personal dashboard via the browser. Dashed arrows indicate access in the background, e.g., all data exchange between the webinterface and the database interface are not visible to the users. The arrow heads indicate the type of data exchange. A one-way arrow represents a data exchange on request, e.g., the admin decides on request when to upload new images to the platform. Two-way arrows indicate a constant flow of data, e.g., between the database and the background services.

The admin controls the platform, which is composed of a web interface and a database. There are several background services, e.g., STAPLE and backup engines. In addition, we have an external interface to Amazon MTurk and a game server, which is based on our own previous work [53]. The entire platform uses free software only (Table 1) and can be accessed at https://welineation.plri.de.

The admin defines the task and its parameters. Images can be imported in popular formats (*.png/*.jpg) which

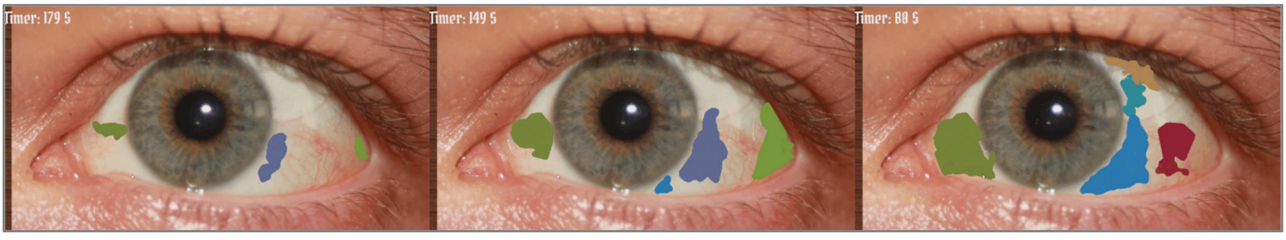

Figure 1: Progress during a single round of the game. Each colored area represents the progress of a single gamer.

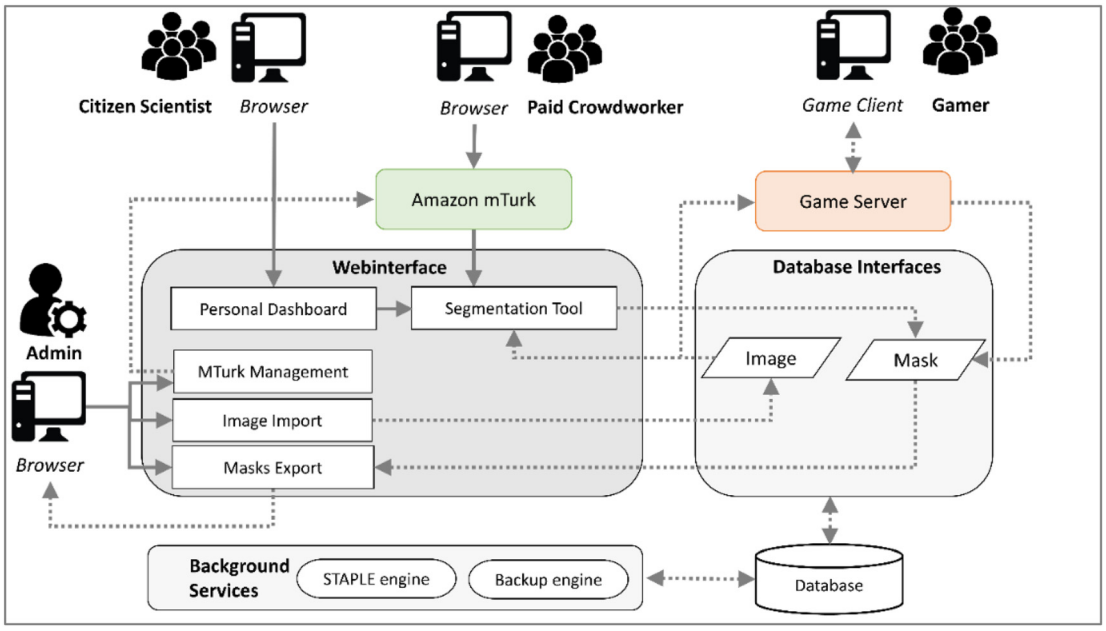

Figure 2: Platform architecture. For the sake of clarity, tasks which assign a number of images and a number of desired masks to a crowdsourcing group are not represented.

Table 1: Software used for platform development.

Component	Software library
Web interface frontend	Bootstrap material design, Javascript
Segmentation tool	PaperJS, Simplify.js, jQuery
Personal dashboard	charts.js, pace.js, progressbar.js, D3.js
Web interface backend	Django framework
Database interfaces	Django framework
Database	PostgreSQL
STAPLE engine	Python SciPy, NumPy, Pillow
Containerization and deployment	Docker, Kubernetes, Gitlab CI
Game client	Unity framework
Game server	Node.js

are then stored in the database. Running tasks can be monitored and aborted. After cancellation or completion of a task, the generated masks can be downloaded. Additionally, we provide the administrator an MTurk management interface, which allows to i) see all running HITs on the MTurk platform, ii) analyze their status (number of pending or performed segmentation), or iii) cancel a running HIT.

Figure 3 shows the personal dashboard of the citizen scientists. It shows graphical elements indicating performance, i.e., the activity in the past months and a user level, which is based on the number of created masks. Below, it shows the list of tasks.

If the citizen scientist starts a task for the first time, the informative text is shown. Then, the first image is presented. Citizen scientists can navigate through the images (Figure 4). For annotation, they can choose to click individual contour points or continuously move the mouse with pressed mouse button along the contour. The number of vertices is automatically minimized while contour is smoothed using Bezier interpolation. Citizen scientists can add, delete, or move vertices [54]. If they press “prev” or “next” or any menu option, the system automatically stores the vertices in the database as a mask of the corresponding images.

If the admin marks a task for being processed by paid crowdworkers, a background process makes the task accessible as a HIT on Amazon MTurk (Figure 5) using the provided API. If paid crowdworkers select a HIT, they are forwarded directly to the image segmentation tool (Figure 4, right part). If the segmentation is done, they directly return to the MTurk site, our system automatically approves the HIT without any sanity check, and MTurk transfers the payment to the worker’s account.

Comparison of crowdsourcing methods

For each crowdsourcing method, a task was generated to generate masks for the 50 training images. For citizen science and paid microtasks exactly 4 masks/image were acquired. For the crowdsourcing game, there were sometimes more masks generated as Gamers played more than 4 rounds per images. In that case, 4 masks were randomly selected.

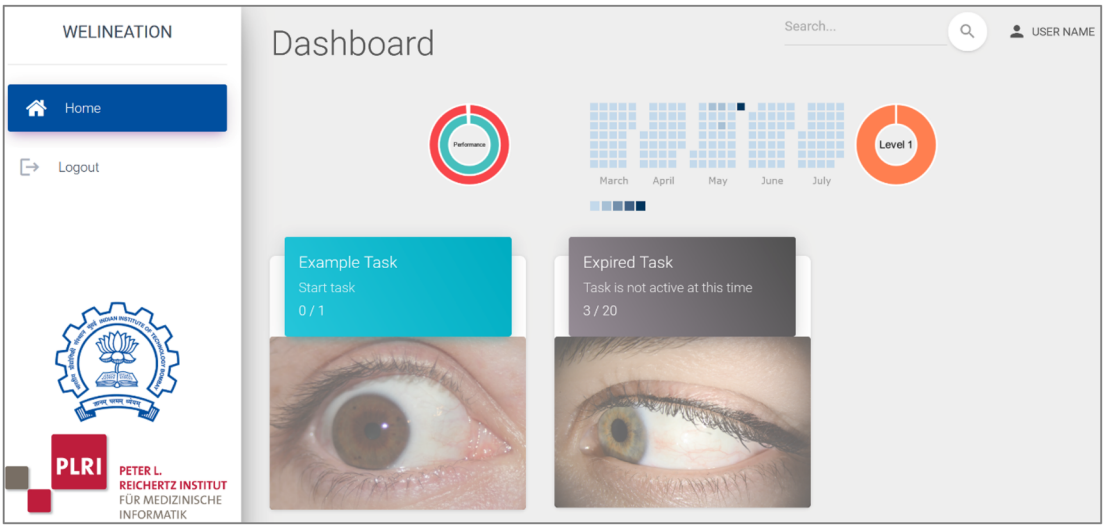

Figure 3: Personal dashboard.

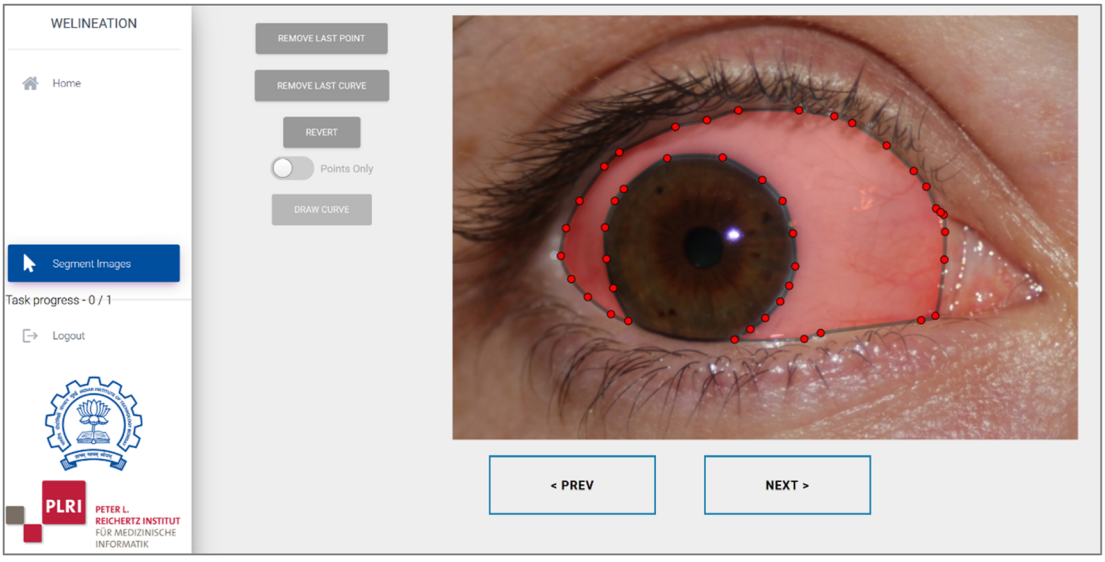

Figure 4: Segmentation tool.

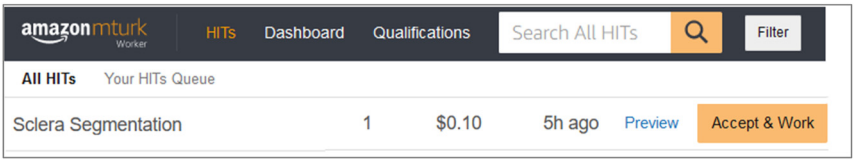

Figure 5: MTurk interface.

Unfortunately, we needed to exclude two images retrospectively from the training set as we observed that the ground truth was inaccurate (SBVP filenames: “11L_u_1.JPG”, “24L_u_3.JPG”).

Using raw masks, the median F-scores for citizen science, paid microtask, and gamification are 82.2, 69.4, 69.3 %, respectively. For STAPLE masks, they are 80.7, 54.3, 78.3 %, respectively (Figure 6). The minimum values for raw and STAPLE masks are 7 and 39.0 %, 1 and 12.6 %, and 27.5 and 30.0 % respectively.

In addition, we computed F-scores for individual images (Figure 7). For each image and technique, a boxplot

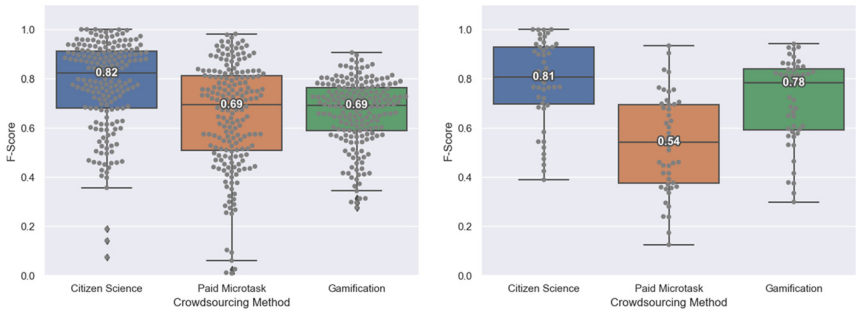

Figure 6: F-Scores of all masks (left) and STAPLE masks (right). Numbers indicate median values rounded to two decimal digits.

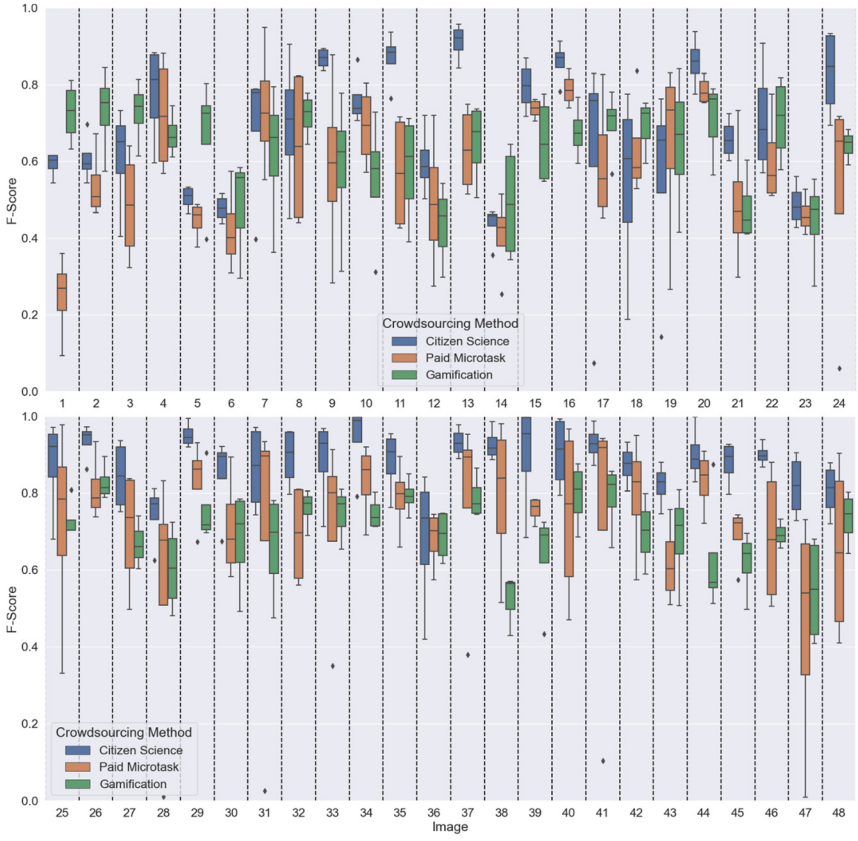

Figure 7: F-Scores of all masks for each individual image.

representing the four acquired masks is given. Citizen science, paid microtask, and gamification showed highest median values in 77.08 %, 6.25 %, and 16.6 7% of images, respectively. Outliers occurred in 25 %, 22.9 %, and 14.6 % of images, respectively.

We manually analyzed paid microtask images with particular low F-scores smaller than 20 %. We detected wrong segmentation as the major reasons, namely delineation of the whole eye (Figure 8, left), delineation of the iris (Figure 8, middle), and wrong usage of the tool (Figure 8, right).

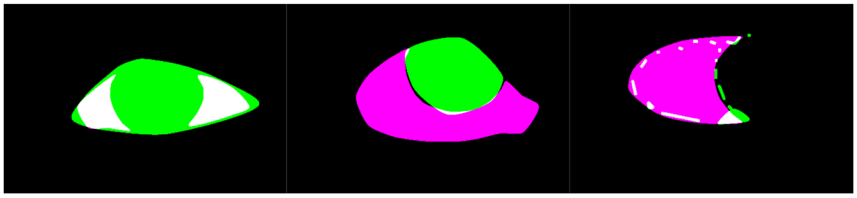

Figure 8: Masks acquired by paid microtask with particular low F-scores. White, green, and magenta pixels indicate true positive, false negative, and false positive pixel labelling, respectively.

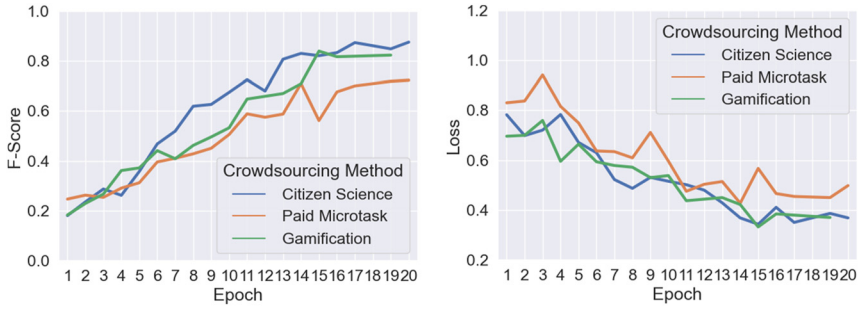

Figure 9: Mean training F-scores (left) and loss (right). For some epochs there are no values due to early stopping.

Deep learning performance

We trained four models for each incentive and mean F-scores and losses saturate after 15 epochs (Figure 9). The highest F-score is from citizen science and only slightly below 90 %. It is followed by gamification and paid microtask, with about 80 % and 70 %, respectively.

We applied the four models that were trained for each incentive technique on the test dataset. The predicted masks qualitatively reflect the performance from training (Figure 10) with citizen science performing best, followed by gamification and paid microtask. Overexposure and speckles occur on the eye lid at the outer regions of the images with these pixels being often falsely classified.

Quantitatively, the highest average F-scores of 80.0 % results from citizen science, followed by gamification and paid microtask with 76.5 % and 73.5 %, respectively (Figure 11). The four models trained with gamification data show the largest interquartile range deviation, followed by paid microtask and citizen science.

Time and effort analysis

Citizen scientists freely decided when to perform the task and waited W_{cs}=21:50 (hh:mm; median value) after receiving the invitation mail before submitting their first mask. After generation of a task on our platform, it took W_{pm}≈00:01 until the HIT was available on MTurk and paid crowdworkers started working in parallel.

Analyzing time stamps on the server shows D_{cs}≈06:11 for acquisition of the whole dataset. Citizen scientists deliver one mask with a median of 60 s (Figure 12) but make long breaks. Paid crowdworkers deliver a mask every 6 s and D_{pm}=00:48. Therefore, total times are

$$T_{cs} = W_{cs} + D_{cs} = 21:50 + 06:11 = 28:01$$

$$T_{pm} = W_{pm} + D_{pm} = 00:01 + 00:48 = 00:49$$

Figure 10: Sclera pixels (yellow) predicted by the U-Net trained with data from citizen science (left), gamification (center), and paid microtask (right).

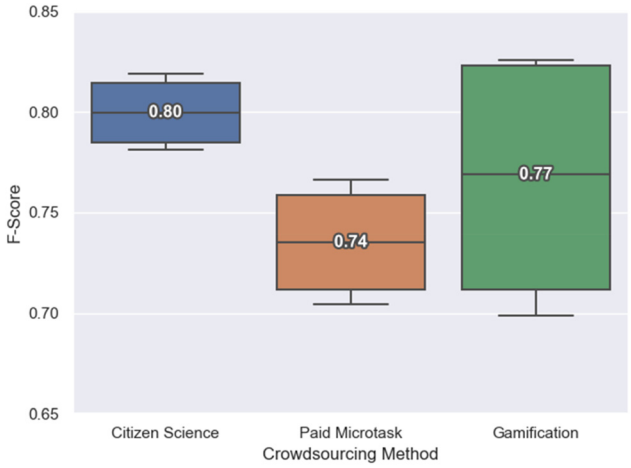

Figure 11: F-Scores of applying trained models to test dataset. Each boxplot represents the F-scores of four models.

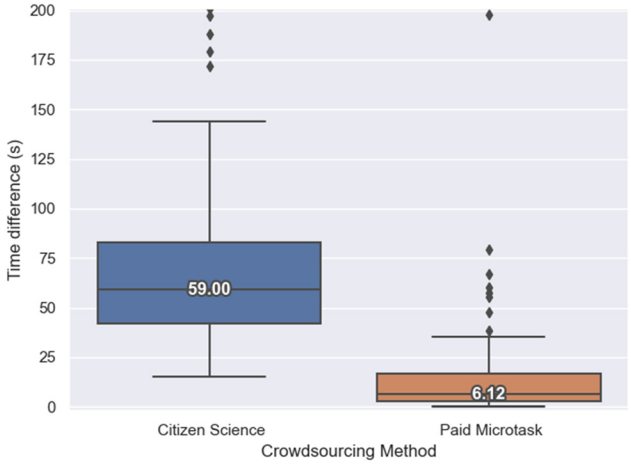

Figure 12: Time differences between masks provided by citizen scientists (left) and paid crowdworkers (right). Values are clamped at 200 s with 4.8 % of citizen science values being larger.

Discussion

In this work, we introduced a platform for crowdsourcing medical image segmentation that integrates citizen science, paid microtask, and gamification to generate ground truth for deep learning. As proof-of-concept, we compared the three crowdsourcing methods for acquiring sclera masks on RGB photographs of human eyes. We answer the first research question (*Is the platform suitable for integrating different crowdsourcing methods?)* with a positive answer. The platform proved useful for acquiring masks from each method. Due to all data being integrated in a single system, masks and metadata were all in a unified format, avoiding data conversions which facilitated the comparison of the different methods significantly.

Regarding the proof-of-concept study, masks of citizen scientists yield highest F-scores with many reaching values higher than 90 %. With gamification, only a single mask reaches such a high score but the number of outliers is reduced: all masks are higher than 20 %. Contrarily, paid microtask results in six outliers (F-scores<20 %). After applying STAPLE, the F-score of gamification is improved but paid microtasks are decreases. Apparently, masks with low scores gain too much weight, negatively impacting the consensus STAPLE masks.

To answer our second research question (*How accurate are crowdsourced masks as compared to expert-based delineations?)*: crowdsourcing delivers results from non-experts which are insufficient to directly replace the domain expert with individual masks being not acceptable as a training dataset. However, in other works it has been shown that several novices can in fact replace the expert in case enough data is available [54] so maybe for the problem at hand, scaling the experiments too larger number of masks acquired could improve F-scores to expert level.

Subsequently, we analyzed deep learning performance when being trained with masks from crowdsourcing. The models trained with citizen science data perform best, followed by gamification, and paid microtask. The worst model from citizen science outperforms the best model from paid microtask. Hence, we answer our third research question (*How good generalizes a DL model that is trained with crowdsourced masks?*): in general, as well as in our particular use case, an F-score of 80 % is insufficient for computer-aided medical diagnostics. However, further analysis with more than 50 training images is required as we expect to increase the performance of DL with more training data rather than better quality of the training data [12]. Moreover, the DL architecture was not finely tuned. For comparison, state-of-the-art methods report F-scores larger than 95 % on the SBVPI dataset [55]. In a recent challenge for sclera detection in photographs acquired using mobile phones, the best-performing group used a modified U-Net and reached 86.8 % [56]. While the reported results in our work are lower, we believe that they reflect meaningfully the effectiveness of the different crowdsourcing incentives.

With respect to the comparison of the different methods of crowdsourcing, our work has several limitations, as we primarily aimed to present the integrated platform architecture and a proof of concept study. For instance, we only used F-score for evaluation which has certain disadvantages as it is a composite measure [57] and other measures might result in slightly different outcomes. However, we used it because it is a typical measure in the field of sclera segmentation and intuitively reflects segmentation quality.

Another limitation is that our game needs a ground truth segmentation. In the future, we want to analyze in how far the game can be realized without any underlying annotation. Furthermore, the partial masks from gamification were simply added. If a gamer confuses the sclera with reflections, the entire mask is impacted negatively. In future work, more meaningful combinations of the partial masks can be applied, for example, as recently proposed by Petit et al. [58]. We will also employ graph-cut algorithms according to Balaji et al. [59].

Furthermore, we analyzed the time to acquire a set of reference data. As HITs do not require any planning time and the tasks are performed in parallel, this is a very fast option for ground truth generation. Our 50 images were processed in just 49 min. Regarding the lower quality of masks from a paid microtask, we observed that 7 % of the paid crowdworkers misunderstood the task: they delineated the iris or the whole eye instead of the sclera (Figure 9). 2.5 % of the masks appear to be intentionally wrong. The workers randomly clicked somewhere and finished the HIT. Evidently, this unintended behavior could also be explained by a non-optimal explanation and missing quality control. In a follow-up study, we target this issue by a more detailed explanation and a "training phase" which requires a certain quality of segmentations before the paid task begins.

Another avenue for future work is to narrow the worker's qualification or adapt payment. We offered US$ 0.10 for 2 min or less, which is more than US$ 3 per hour and equals similar HITs on MTurk. In Germany however, the hourly minimal wage by law is about US$ 12. To the best of our knowledge, there is no "payment vs. quality" analysis yet regarding the crowdsourcing of (medical) image segmentations. However, Moayedikia, Gaderi & Yeoh recently proposed algorithms to optimize the budget of HITs [60]. Furthermore, according to Heim et al. [61], we will add a quality control stage to our platform that checks a submitted mask before its acceptance and payment.

Our work has several limitations. We try to compare three different crowdsourcing methods using the acquisition of medical image masks as proof-of-concept study. By the nature of this approach, this results in a problem of comparing their pros and cons due to different tasks design and – as it was not possible to recruit the same participants for all three methods. Therefore, we see this work as a first step towards a comparison of the different methods but evidently, more work is required to increase the generalizability of the results. Moreover, the different methods had different time constraints which were necessary to enable a joyful game experience for the gamification approach and a natural time limit is required for a paid microtask needs to be defined in the mTurk system. Therefore, our results are specific for this use case and might be biased by that, therefore requiring more analysis.

Currently, we are turning the platform into open source. This will allow others to use the platform and contribute to its improvement. In future work we aim at conducting large-scaled experiments to analyze open parameters of paid microtask (e.g., payment, qualification) and gamification (e.g., combination of partial masks).

Conclusions

- In this paper, we introduced a unified platform for crowdsourcing medical image segmentation.
- Citizen science, paid microtask, and gamification are all suitable to obtain training data for CNNs we the proposed platform.
- Citizen science performs with best accuracy, paid microtask is fastest but also most cost intensive, and gamification is inexpensive and performs with accuracies in between the other approaches.
- All crowdsourcing methods yield silver standards only [8] but bridge missing training data in medical deep learning [9].

Acknowledgments: The authors thanks Marcel Fricke, Jan Gabow, Markus Gamperle, and Luca Heinrich of TU Braunschweig for support in software development.
Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: All authors state no conflict of interest.
Research funding: This work was performed in the scope of the FAIR4Health project (https://www.fair4health.eu/) and received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 824666.
Data availability: Not applicable.

Appendix

Examples of F-scores

The figure illustrates the accuracy of masks with F-Scores of 91.0 %, 76.6 %, and 37.6 %, respectively (Figure 13).

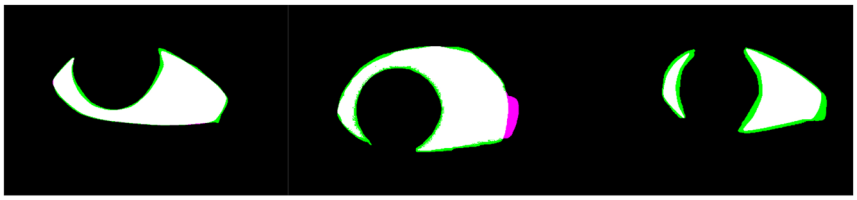

Figure 13: Masks accuracy. White, green, and magenta pixels indicate true positive, false negative, and false positive pixel labelling, respectively.

Details on STAPLE algorithm

We use the STAPLE algorithm proposed by Warwick et al. [11] to combine different masks into a consensus mask which is an estimate of the unknown true mask. STAPLE is a majority voting algorithm which is agnostic of the underlying image type and therefore does not include any kind of semantic information or segmentation mechanisms. It assumes that the masks were acquired independently from each other and estimates a sensitivity and the specificity value for each mask. Using these metrics, it uses the concept of iterative expectation-maximization by two subsequent steps: in the E-step, the expected value of the log likelihood function under the posterior distribution of the observed data is estimated. In the M-step, the expected value is maximized by finding appropriate sensitivity and the specificity values. Both steps are repeated until convergence and thereby enable the computation of a consensus mask and also an estimate of sensitivity and the specificity for each mask.

References

1. Hsu W, Baumgartner C, Deserno TM. Notable papers and new directions in sensors, signals, and imaging informatics. Yearb Med Inform 2021;30:87–95.
2. Hsu W, Baumgartner C, Deserno TM. Notable papers and trends from 2019 in sensors, signals, and imaging informatics. Yearb Med Inform 2020;29:139–44.
3. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
4. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
5. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing; 2015:234–41 pp.
6. Hsu W, Baumgartner C, Deserno T. Advancing artificial intelligence in sensors, signals, and imaging informatics. Yearb Med Inform 2019;28: 115–7.
7. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. Med Image Anal 2020;63:101693.
8. Lehmann TM. From plastic to gold: a unified classification scheme for reference standards in medical image processing. In: Sonka M, Fitzpatrick JM, editors. Medical Imaging 2002: image processing. San Diego, CA, USA: SPIE; 2002, vol 4684:1819–27 pp.
9. Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. IEEE Access 2018;6:9375–89.
10. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004;23:903–21.
11. Lucena O, Souza R, Rittner L, Frayne R, Lotufo R. Silver standard masks for data augmentation applied to deep-learning-based skull-stripping. In: Amini A, Acton S, editors. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). Washington, DC: IEEE; 2018:1114–7 pp.
12. Wesemeyer T, Jauer M-L, Deserno TM. Annotation quality vs. quantity for deep-learned medical image segmentation. In: Park BJ, Deserno TM, editors. Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications. Online Only: SPIE; 2021: 11601 p.
13. Vitek M, Rot P, Štruc V, Peer P. A comprehensive investigation into sclera biometrics: a novel dataset and performance study. Neural Comput Appl 2020;32:17941–55.
14. Sirazitdinova E, Gijs M, Bertens CJF, Berendschot TTJM, Nuijts RMMA, Deserno TM. Validation of computerized quantification of ocular redness. Trans Vis Sci Tech 2019;8:31.
15. Dogan S, Astvatsatourov A, Deserno TM, Bock F, Shah-Hosseini K, Michels A, et al. Objectifying the conjunctival provocation test: photography-based rating and digital analysis. Int Arch Allergy Immunol 2014;163:59–68.
16. Sárándi I, Claßen DP, Astvatsatourov A, Pfaar O, Klimek L, Mösges R, et al. Quantitative conjunctival provocation test for controlled clinical trials. Methods Inf Med 2014;53:238–44.
17. Ghezzi A, Gabelloni D, Martini A, Natalicchio A. Crowdsourcing: a review and suggestions for future research: crowdsourcing. Int J Manag Rev 2018;20:343–63.
18. Estellés-Arolas E, González-Ladrón-de-Guevara F. Towards an integrated crowdsourcing definition. J Inf Sci 2012;38:189–200.
19. Cohn JP. Citizen science: can volunteers do real research? Bioscience 2008;58:192–7.
20. Kaufmann N, Schule T, Veit D. More than fun and money. Worker motivation in crowdsourcing – a study on mechanical turk. In: Rajagopalan B, Goes P, editors. 17th Americas Conference on Information Systems (AMCIS 2011). Atlanta, GA: AISeL. 340 p.
21. Hamari J, Koivisto J, Sarsa H. Does gamification work? – a literature review of empirical studies on gamification. In: Sprague Jr RH, editor. 2014 47th Hawaii international conference on system sciences. Waikoloa, HI: IEEE; 2014:3025–34 pp.
22. Ørting SN, Doyle A, Van Hilten A, Hirth M, Inel O, Madan CR, et al. A survey of crowdsourcing in medical image analysis. Hum Comput J 2020;7:1–26.
23. Bafti SM, Ang CS, Hossain MM, Marcelli G, Alemany-Fornes M, Tsaousis AD. A crowdsourcing semi-automatic image segmentation platform for cell biology. Comput Biol Med 2021;130:104204.

24. Marzahl C, Aubreville M, Bertram CA, Maier J, Bergler C, Kröger C, et al. EXACT: a collaboration toolset for algorithm-aided annotation of images with annotation version control. Sci Rep 2021;11:4343.
25. Goel S, Sharma Y, Jauer M-L, Deserno TM. WeLineation: crowdsourcing delineations for reliable ground truth estimation. In: Park BJ, Deserno TM, editors. Medical imaging 2020: imaging informatics for healthcare, research, and applications. Houston, TX, USA: SPIE; 2020.
26. Albarqouni S, Baur C, Achilles F, Belagiannis V, Demirci S, Navab N. AggNet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Trans Med Imaging 2016;35:1313–21.
27. Maier-Hein L, Ross T, Gröhl J, Glocker B, Bodenstedt S, Stock C, et al. Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Cham: Springer International Publishing; 2016:616–23 pp.
28. Grote A, Schaadt NS, Forestier G, Wemmert C, Feuerhake F. Crowdsourcing of histological image labeling and object delineation by medical students. IEEE Trans Med Imaging 2019;38:1284–94.
29. Sardi L, Idri A, Fernández-Alemán JL. A systematic review of gamification in e-Health. J Biomed Inform 2017;71:31–48.
30. Morschheuser B, Hamari J, Koivisto J. Gamification in crowdsourcing: a review. In: Bui TX, Sprague Jr RH, editors. 2016 49th Hawaii International Conference on System Sciences (HICSS). Koloa, HI, USA: IEEE; 2016:4375–84 pp.
31. Deterding S, Dixon D, Khaled R, Nacke L. From game design elements to gamefulness: defining "gamification." In: Lugmayr A, Franssila H, Safran C, Hammouda I, editors. 15th International Academic MindTrek Conference on Envisioning Future Media Environments – MindTrek '11. Tampere, Finland: ACM Press; 2011:9 p.
32. Balducci F, Buono P. Building a qualified annotation dataset for skin lesion analysis trough gamification. In: Catarci T, Kent N, Mecella M, editors. 2018 international conference on advanced visual interfaces. Castiglione della Pescaia Grosseto, Italy: ACM; 2018:1–5 pp.
33. Ionescu GV, Harkness EF, Hulleman J, Astley SM. A citizen science approach to optimising computer aided detection (CAD) in mammography. In: Nishikawa RM, Samuelson FW, editors. Medical Imaging 2018: image perception, observer performance, and technology assessment [Internet]. Houston, United States: SPIE; 2018: 34 p. [cited 2021 Nov 5]. Available from: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10577/2293668/A-citizen-science-approach-to-optimising-computer-aided-detection-CAD/10.1117/12.2293668.full.
34. Mavandadi S, Dimitrov S, Feng S, Yu F, Sikora U, Yaglidere O, et al. Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study. PLoS One 2012;7:e37245.
35. Arganda-Carreras I, Turaga SC, Berger DR, Cireşan D, Giusti A, Gambardella LM, et al. Crowdsourcing the creation of image segmentation algorithms for connectomics. Front Neuroanat 2015;9. https://doi.org/10.3389/fnana.2015.00142.
36. Waldispühl J, Szantner A, Knight R, Caisse S, Pitchford R. Leveling up citizen science. Nat Biotechnol 2020;38:1124–6.
37. Sullivan DP, Winsnes CF, Åkesson L, Hjelmare M, Wiking M, Schutten R, et al. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. Nat Biotechnol 2018;36: 820–8.
38. Gurari D, Theriault D, Sameki M, Isenberg B, Pham TA, Purwada A, et al. How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms. In: 2015 IEEE winter conference on applications of computer vision. Waikoloa, HI, USA: IEEE; 2015:1169–76 pp.
39. Sharma M, Saha O, Sriraman A, Hebbalaguppe R, Vig L, Karande S. Crowdsourcing for chromosome segmentation and deep classification. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW). Honolulu, HI, USA: IEEE; 2017:786–93 pp.
40. Heim E, Roß T, Seitel A, März K, Stieltjes B, Eisenmann M, et al. Large-scale medical image annotation with crowd-powered algorithms. J Med Imag 2018;5:1.
41. Cheplygina V, Perez-Rovira A, Kuo W, Tiddens HAWM, de Bruijne M. Crowdsourcing airway annotations in chest computed tomography images. PLoS One 2021;16:e0249580.
42. Mitry D, Peto T, Hayat S, Morgan JE, Khaw K-T, Foster PJ. Crowdsourcing as a novel technique for retinal fundus photography classification: analysis of images in the epic norfolk cohort on behalf of the ukbiobank eye and vision consortium. PLoS One 2013;8:e71154.
43. Mitry D, Peto T, Hayat S, Blows P, Morgan J, Khaw K-T, et al. Crowdsourcing as a screening tool to detect clinical features of glaucomatous optic neuropathy from digital photography. PLoS One 2015;10:e0117401.
44. Mitry D, Zutis K, Dhillon B, Peto T, Hayat S, Khaw K-T, et al. The accuracy and reliability of crowdsource annotations of digital retinal images. Trans Vis Sci Tech 2016;5:6.
45. Liang H, Wang M-M, Wang J-J, Xue Y. How intrinsic motivation and extrinsic incentives affect task effort in crowdsourcing contests: a mediated moderation model. Comput Hum Behav 2018;81:168–76.
46. Feyisetan O, Simperl E, Van Kleek M, Shadbolt N. Improving paid microtasks through gamification and adaptive furtherance incentives. In: Gangemi A, Leonardi S, Panconesi A, editors. 24th International World Wide Web Conference. Florence Italy: International World Wide Web Conferences Steering Committee; 2015:333–43 pp.
47. Bowser A, Hansen D, He Y, Boston C, Reid M, Gunnell L, et al. Using gamification to inspire new citizen science volunteers. In: Nacke LE, Harrigan K, Randall N, editors. First international conference on gameful design, research, and applications. Toronto, Ontario, Canada: ACM; 2013:18–25 pp.
48. Tinati R, Luczak-Roesch M, Simperl E, Hall W. An investigation of player motivations in Eyewire, a gamified citizen science project. Comput Hum Behav 2017;73:527–40.
49. Yakubovskiy P. Segmentation Models Pytorch [Internet]. GitHub repository. GitHub; 2020. Available from: https://github.com/qubvel/segmentation_models.pytorch.
50. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:14091556 [cs] [Internet]. 2015 [cited 2021 Jul 15]; Available from: http://arxiv.org/abs/1409.1556.
51. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Bengio Y, LeCun Y, editors. 3rd international conference for learning representations. San Diego, CA, USA: Arxiv; 2017.
52. Dumitrache A, Aroyo L, Welty C, Sips R-J, Levas A. "Dr. Detective": combining gamication techniques and crowdsourcing to create a gold standard in medical text. In: Acosta M, Aroyo L, Bernstein A, Lehman J, Noy N, editors. 1st international conference on crowdsourcing the semantic web. Sydney, Australia: ACM; 2013:16–31 pp.
53. Jauer M-L, Spicher N, Deserno TM. Gamification concept for acquisition of medical image segmentation via crowdsourcing. In: Park BJ, Deserno TM, editors. Medical Imaging 2021: imaging informatics for healthcare, research, and applications. Online Only: SPIE; 2021:12 p.
54. Goel S, Sharma Y, Jauer M-L, Deserno TM. WeLineation: crowdsourcing delineations for reliable ground truth estimation. In: Proc SPIE Medical

Imaging 2020: imaging informatics for healthcare, research, and applications. Houston, Texas, United States: SPIE; 2020.
55. Rot P, Vitek M, Grm K, Emeršič Ž, Peer P, Štruc V. Deep sclera segmentation and recognition. In: Uhl A, Busch C, Marcel S, Veldhuis R, editors. Handbook of Vascular Biometrics. Cham: Springer International Publishing; 2020:395–432 pp.
56. Vitek M, Das A, Pourcenoux Y, Missler A, Paumier C, Das S, et al. SSBC 2020: sclera segmentation benchmarking competition in the mobile environment. In: Kakadiaris IA, Phillips J, Vatsa M, editors. 2020 IEEE International Joint Conference on Biometrics (IJCB). Houston, TX, USA: IEEE; 2020:1–10 pp.
57. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag 2009;45:427–37.
58. Petit O, Thome N, Soler L. Iterative confidence relabeling with deep ConvNets for organ segmentation with partial labels. Comput Med Imag Graph 2021;91:101938.
59. Balaji VR, Suganthi ST, Rajadevi R, Krishna Kumar V, Saravana Balaji B, Pandiyan S. Skin disease detection and segmentation using dynamic graph cut algorithm and classification through Naive Bayes classifier. Measurement 2020;163:107922.
60. Moayedikia A, Ghaderi H, Yeoh W. Optimizing microtask assignment on crowdsourcing platforms using Markov chain Monte Carlo. Decis Support Syst 2020;139:113404.
61. Heim E, Seitel A, Andrulis J, Isensee F, Stock C, Ross T, et al. Clickstream analysis for crowd-based object segmentation with confidence. IEEE Trans Pattern Anal Mach Intell 2018;40:2814–26.