Health Informatics Journal

Original Article

Health Informatics Journal
2020, Vol. 26(4) 2446–2469
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1460458220907435
journals.sagepub.com/home/jhi
SAGE

Crowdsourcing human-based computation for medical image analysis: A systematic literature review

Nataša Petrović
University of Belgrade, Serbia

Gabriel Moyà-Alcover, Javier Varona and Antoni Jaume-i-Capó
Universitat de les Illes Balears, Spain

Abstract

Computer-assisted algorithms for the analysis of medical images require human interactions to achieve satisfying results. Human-based computation and crowdsourcing offer a solution to this problem. We performed a systematic literature review of studies on crowdsourcing human-based computation for medical image analysis based on the guidelines proposed by Kitchenham and Charters. We identified 43 studies relevant to the objective of this research. We determined three primary purposes and problems that crowdsourcing human-based computation systems can solve. We found that the users provided five information types. We compared systems that use pre-, post-evaluation and quality control methods to select and filter the user inputs. We analyzed the metrics used for the evaluation of the crowdsourcing human-based computation system performance. Finally, we identified the most popular crowdsourcing human-based computation platforms with their advantages and disadvantages.Crowdsourcing human-based computation systems can successfully solve medical image analysis problems. However, the application of crowdsourcing human-based computation systems in this research area is still limited and more studies should be conducted to obtain generalizable results. We provided guidelines to practitioners and researchers based on the results obtained in this research.

Keywords

crowdsourcing, human-based computation, medical image analysis

Corresponding author:
Antoni Jaume-i-Capó, Unitat de Gràfics i Visió per Computador i Inteligència Artificial, Departament de Ciències Matemàtiques i Informàtica, Universitat de les Illes Balears, Edifici Anselm Turmeda, Palma 07122, Spain.
Email: antoni.jaume@uib.es

Introduction

Medical images analysis is essential in diagnostics.[1,2] The most common task of medical image analysis is image segmentation. Image segmentation is the process that separates the desired object from its background. Typically, it is the foundation for all subsequent steps, such as the classification of segmented objects. In most cases, it can directly impact a patient in further therapy planning. Manual image segmentation can be monotonous and time consuming even for highly qualified experts. As an alternative to manual image processing, computer-aided algorithms for image segmentation are being developed.[3–5] Computer-aided detection/diagnosis (CAD) offers support with decision making to medical experts.[6] In this case, a specialist uses the program's output to make the final decision in diagnosis. In recent years, deep learning methods have been used for medical image analysis.[7,8] One of the primary advantages of automating medical image is the reduction in time and cost. However, these algorithms are not sufficiently precise and, in most cases, manual interaction is required to validate data and achieve satisfying results.

Interactive segmentation refers to methods that are based on an iterative process, in which the user is crucial for managing and correcting the segmentation results. Users can perform different tasks (human intelligence tasks—HITs) during the interaction.[9,10] They can evaluate the results of the algorithm that provides feedback and improve the future outcomes of the automatic system.[11] HITs can have a significant role in the development and result validation of the automatic algorithm, as that algorithm is an essential part of CAD systems. This is especially because machine learning methods are used in such systems and they usually need a great amount of ground truth data to be trained properly.[12] The task of initializing the parameters for the algorithm to obtain precise results for the current problem exists.[13,14]

Human-based computation (HbC) and crowdsourcing systems can be more time-saving and cost-efficient regarding time and money than systems that require experts' knowledge at a comparable quality.[15] HbC is a developing area of research using human intelligence to solve problems that are beyond the capabilities of the current artificial intelligence models. Many definitions of the term HbC exist, but the first to define it was Von Ahn[16] in his dissertation, "Human Computation"—". . . a paradigm for utilizing human processing power to solve problems that computers cannot yet solve." As Quinn and Banderson[17] stated, the principal constituents of HbC are the following: (a) the problems fit the general paradigm of computation, and as such might someday be solvable by computers; and (b) the human participation is directed by the computational system or process.

HbC systems can consist of several human computers, or they can distribute tasks to thousands of humans. The number of humans depends on the computational problem to be solved. For some specific problems, fewer highly trained individuals are required; however, in other cases, the tasks can be distributed to a large number of untrained humans. The limitations of HbC are how to reach large numbers of users and how to train the individuals effectively to ensure an acceptable processing time of the problem and the minimum confidence of the results.

To distribute the work to a group of people, a relatively new technique called crowdsourcing is used. Typically, it is used through online frameworks, for solving expensive and challenging computational problems. The term was initially created in 2006 by combining the words "crowd" and "outsourcing" to describe "the act of taking a job traditionally performed by a designated agent (typically an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call."[18] Crowdsourcing can be divided into two primary groups—those that solve numerous but straightforward microtasks, and others that solve megatasks.[19] The primary challenge in designing HbC and crowdsourcing systems is the motivation of the participants. People are motivated in different ways to solve a particular problem. One approach is to pay the workers, using a crowdsourcing platform, such as Amazon Mechanical Turk (AMT) or Figure Eight (FE, the former name was CrowdFlower).[20] The other motivation is the sense of performing a fulfilling job, such as solving DNA-based problems.[21] Meanwhile, the workers can be coerced in performing the job, for example, the ReCAPTCHA,[22] where to gain access to the system, a distorted image of two words is presented and only one word is known by the system. If the user correctly types the word known by the system, the user is granted access and the system considers that the unknown word has been correctly spelled. Thus, the system determines that the user is a human and also forces the digitization of unknown scanned words. Furthermore, the contributors can enjoy the participation, for example, by playing serious games.[23,24]

The gamification of crowdsourcing provides greater motivation to the participants.[25] Games with a purpose (GWAP) present microtasks in the context of simple, typically web-based games. Participants earn points and advance through the levels as well as in other games while contributing to a more important goal. To win, the players must solve real-world problems with high quality and in a large quantity. The first GWAP called the ESP[26] was developed in 2003 with the general purpose of image labeling. After its great success, GWAPs have been actively developed in various areas of research.

We herein present the findings from the systematic literature review on crowdsourcing HbC for medical image analysis. To our knowledge, none have systematically reviewed this topic. There is a systematic review of crowdsourcing in health but it is not focused only on medical image analysis but all of the available health disciplines and methods.[27] The review focused on medical image analysis is required to provide an overview of the rapidly developing field of crowdsourcing HbC and its use for complex problems such as image segmentation.

Our aim is to determine and compare the methods of crowdsourcing HbC for medical image analysis, to conduct further research in this area. In addition, we present the guidelines and recommendations for researchers and practitioners interested in this area.

This article is organized as follows. In section “Method,” we describe the method used for our systematic literature review; this involves producing and following the rules of a protocol. Section “Results and discussion” presents the results of our synthesis of the literature, including the geographical spread and publication details. Here, we report the results of our quality assessment and research questions (RQs). Subsequently, we discuss our key findings. Section “Limitations of the review” presents some limitations of this study. In section “Guidelines for practitioners and scientists,” we provide some recommendations for further research. The last part is section “Conclusion,” in which we present our conclusions.

Method

This study has been undertaken as a systematic literature review based on the guidelines proposed by Kitchenham and Charters.[28] The review protocol is developed, and it includes six stages: RQs definition, search strategy design, study selection, quality assessment, data extraction, and data synthesis.

In the first stage, a set of RQs are defined. In the second stage, the search strategy is designed to obtain the studies relevant to the RQs. This stage involves the definition of search terms and the selection of literature resources. In the third stage, the study selection criteria are defined to identify the relevant studies that can contribute to addressing the RQs. Next, the relevant studies are assessed with defined quality assessment (QA) questions. In the data extraction stage, a data extraction form is defined. Finally, in the final stage, the proper methodologies for synthesizing the extracted data based on the RQs are determined.

Research questions

This systematic literature review (SLR) aims to summarize and clarify the empirical evidence of HbC for medical image analysis. Hence, eight RQs were defined as follows:

RQ1. What is the purpose of crowdsourcing HbC in medical image analysis?

RQ2. What kind of medical image analysis problem is solved?

We identified the image analysis problems and explored the methods that are used for solving them. This question aims to understand all of the challenges that HbC can solve in medical imaging analysis.

RQ3. What methods are used for the evaluation of the crowdsourcing worker's confidence?

RQ4. What are the types of information provided by the user?

We related the types of problems with the types of provided information to differentiate between them more clearly.

RQ5. What metrics are used for the evaluation of the obtained results?

The identification of general metrics provides us with a possible comparison between different systems that are solving a similar problem.

RQ6. What are the results of the comparison between crowdsourcing HbC and other types of approaches?

We focused on the comparison between human computation and automatic systems. Furthermore, the crowdsourcing HbC system performance and the experts' performance are compared. The third possible type is the system that combines automatic methods with human intelligence (fusion system), with which we also conducted the comparison.

RQ7. What HbC or crowdsourcing platform is used?

We explored platforms' similarities and differences, as well as their advantages and disadvantages.

RQ8. How many workers are required to obtain satisfying results, and at what cost and time frame?

Search strategy

The search process includes defining query strings (search terms) and steps conducted to select the papers of interest, as shown in Figure 1. These parts are detailed as follows:

(a) Query strings.

Query strings are defined by deriving important terms from the RQs. These terms are "human-based computation," "crowdsourcing," "medical image analysis," and "serious games." After a brief search, we found a few papers and extracted new significant keywords from them. These keywords are "GWAP," "labeling game," "web-based interaction," "gamesourcing," "image segmentation," "image annotation," and "image classification." By combining the keywords with the original query, we formed the final query string, as follows:

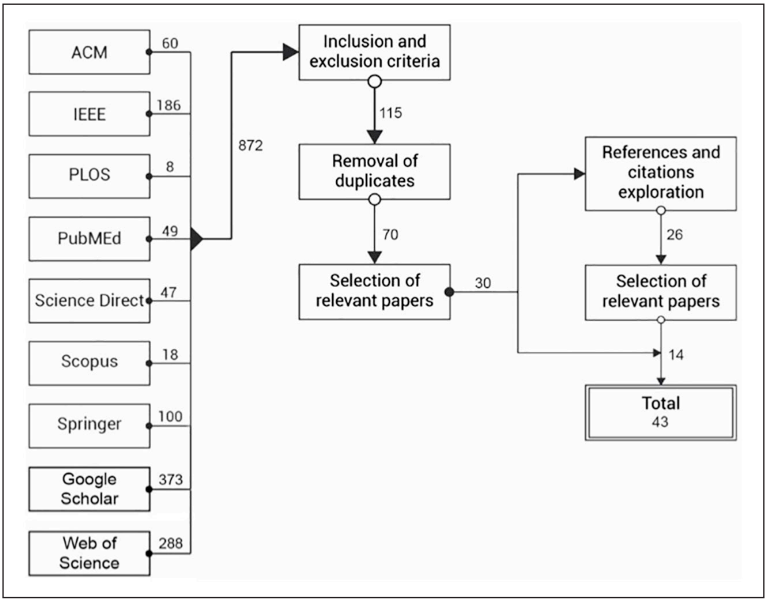

Figure 1. Research process.

(HbC OR crowdsourcing OR GWAP OR serious games OR labeling game OR web-based interaction OR gamesourcing) AND (medical image segmentation OR medical image analysis OR medical image annotation OR medical image classification).

Using this query we performed a general search in every selected database.

(b) Sources selection.

The search has been performed on digital databases as follows: ACM Digital Library, IEEE Xplore Digital Library, PLOS, PubMed, ScienceDirect, Scopus, SpringerLink, Web of Science, and Google Scholar. These web search engines were chosen as they are the most popular academic libraries and are a reliable means of obtaining all the relevant papers. PLOS and PubMed were chosen as they are reliable sources of medical papers.

Inclusion and exclusion criteria

Inclusion criteria. All studies were retrieved with the search string. They were published in journals or conferences and written in English. The most critical inclusion criterion was that the study tries to solve the image analysis problem through crowdsourcing HbC tasks.

Exclusion criteria. We excluded the studies that did not focus on crowdsourcing HbC. Furthermore, we excluded those that did not focus on medical image analysis. We have not found any systematic reviews; however, if we had, we would have excluded them.

Reference and citation exploration

We performed an additional measure to ensure that all of the available published studies have been obtained. This was achieved by the exploration of references and citations of selected studies, as shown in Figure 1. This step included all the papers that had not been found in the initial search, but were related to the topic of the review. When new papers were acquired, the same selection process was conducted as with the initial set of papers.

QA

Kitchenham and Charters[28] suggested performing QA for each included study. However, a standardized definition of the study quality is not available. QA questions are defined to assess the rigor, credibility, and relevance of the selected studies as defined by the Critical Appraisal Skills Programme (CASP).[29] We chose this set of QA questions because it reflects our intention to evaluate the selected papers. Questions QA1, QA3, QA4, and QA5 are obtained from the CASP qualitative checklist, while the others are derived from Dybå and Dingsøyr.[30] Each question has three optional answers: "Yes," "Partly," and "No." They are scored as follows: Yes = 1, Partly = 0.5, and No = 0. The score quality of each study is computed by adding the scores of the answers to the following QA questions:

QA1. Is there a clear statement about the aims and objectives of the research?

This question focuses on the goal of the research and its relevance. We focused on the explanation of the importance of the study topic.

QA2. Is there an adequate description of the context in which the research was performed?

We wanted to verify the clarity of the stated context of the research. We assessed the explanation of the problem and investigated methods that were used to address it.

QA3. Is the research design appropriate to address the aims of the research?

This question explores the justification of the research design. We assessed whether the authors had discussed how they decided which methods to use, and why these were the best options.

QA4. Is the recruitment strategy appropriate to the aims of the research?

We wanted to determine if the researcher explained the recruitment process. We focused on the discussion regarding the number of participants, their experience level, and why it was the most appropriate for the study to choose them.

QA5. Is the data analysis sufficiently rigorous?

The quality of data analysis was assessed. It concerns the description of the data analysis and the sufficiency and presentation of data. We focused on whether the researcher explained the potential bias during data analysis.

QA6. Is there a clear statement of findings?

This question assessed the explicitness of the findings, the existence of adequate discussion, and the credibility of the findings. It further concerns the discussion regarding the study's limitations.

QA7. Is the study of value for research or practice?

This question pertains to the discussed contribution of the study and its relevance. It concerns whether the researchers explained how their paper contributes to the existing knowledge. We assessed whether the researchers identified the new areas where research is necessary.

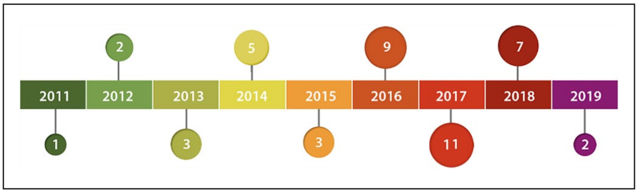

Figure 2. Publishing years of selected studies.

Data collection and analysis

We exploited the selected studies to collect the data that contribute to addressing the research questions. The data extracted from each study is the following:

1. Title;
2. Source;
3. Year of publishing;
4. Authors of the study;
5. Institutions and countries represented by authors;
6. Keywords;
7. Aim of the study;
8. Platforms used for human computation tasks;
9. Datasets used in the study;
10. Measurements used to assess segmentation results.

Extraction cards were used to collect information from the studies. The data are presented in tables to easily visualize and identify the answers to the RQs, trends, and limitations in the research of human computation for medical image analysis. The goal of the data analysis is to combine the evidence from the selected studies to answer the RQs. The data extracted in this review include both quantitative and qualitative data. The quantitative data include results of the performance metrics, and the qualitative data include the evaluation methods and datasets used.

Results and discussion

This section presents and discusses the findings of the review. First, an overview of the selected studies is presented. Next, the review findings to the RQs are reported and discussed.

Overview of the selected studies

Forty-three papers were systematically evaluated in this review. Among them, 22 (51.16%) papers were published in journals, and 21 (48.84%) appeared in conference proceedings. In Supplemental Table A.1 the data extraction cards of these studies are shown. The publication years of the papers are between 2011 and 2019, although the year filter was not set to any particular year. The increasing interest in the research of HbC is shown by the larger number of published studies in recent years. This increase coincides with the emergence of crowdsourcing platforms, more precisely AMT in 2005, and FE in 2008. These platforms enabled much research with small investments.

Figure 2 shows the publishing year of the selected studies. Figure 3 shows the number of institutions per origin country of the studies.

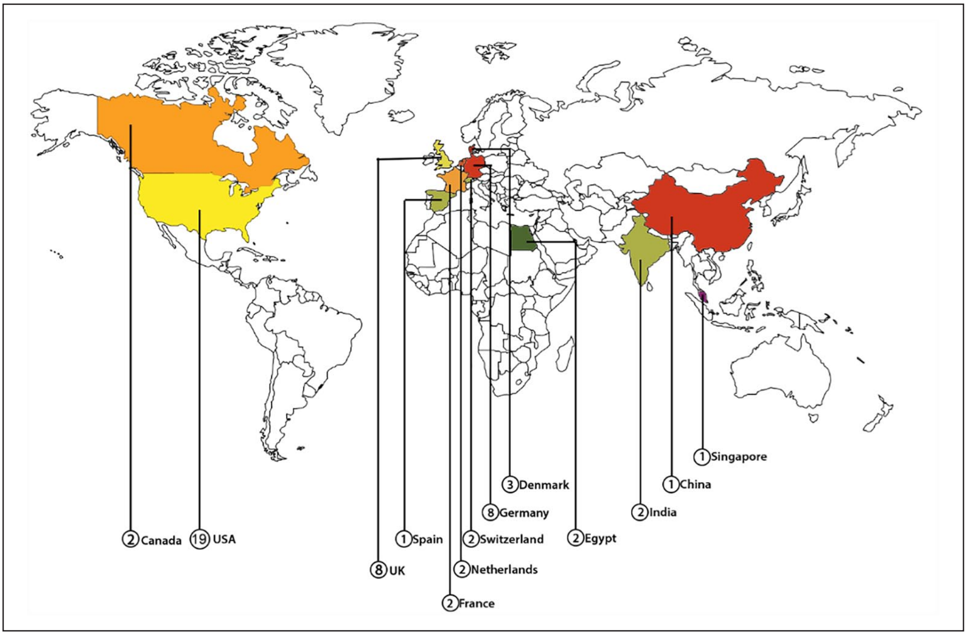

Figure 3. Countries of authors' institutions.

Regarding the types of the selected papers, all the papers pertain to experimental research except for one case study research (S2), and no survey research was found. We filled a form for each of the selected primary studies. To answer the RQs, we used visualization techniques such as graphs and charts.

Quality assessment of the studies

Using the previously shown QA questions, we obtained the results of each study's quality. In total, 18 studies (41.86%) were rated with the maximum number of points (7 points), 13 studies (30.23%) with 6.5 points, 11 studies with 6 points (25.58%), and 1 study with 5.5 points. Four QA questions were repeatedly not fully answered (QA3, QA4, QA5, QA6). In five studies, 0.5 was obtained for QA3 (S10, S11, S16, S17, S19). This is due to the non-existent explanation of the methods used. The discussion about why they were chosen and why they were appropriate for the study design was not provided. Nine studies failed in QA4 (S8, S10, S11, S24, S26, S32, S33, S39, S41). This is the case because a complete discussion about the recruitment process was not provided. Twelve studies earned half of the point in QA5 (S8, S12, S13, S17, S18, S19, S25, S33, S37, S38, S40, S41). In most cases, an in-depth description of the analysis process was not provided. Thirteen studies gained half a point for QA6 (S1, S2, S4, S13, S18, S23, S31, S35, S36, S38, S39, S40, S43). This is due to the unclear description of the findings. Studies with a complete failure to answer the QA questions did not occur (otherwise, 0 points would be given for them). Overall, we conclude that the selected studies are of high quality.

RQ1. What is the purpose of crowdsourcing HbC in medical image analysis?

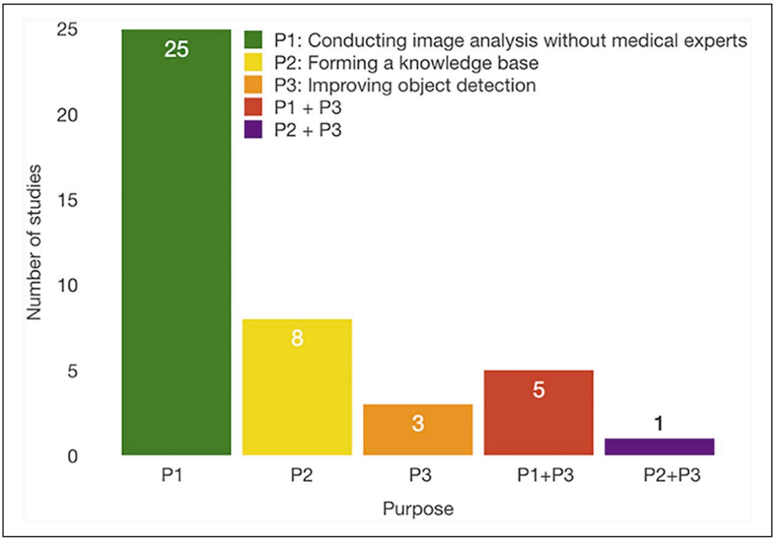

Figure 4. Crowdsourcing HbC purposes resulted by answering *RQ1*.

By answering the first RQ, we found three primary purposes of human computation for image analysis, as shown in Figure 4. It is noteworthy that some studies have multiple purposes. The primary purpose is to conduct image segmentation by non-experts. Non-experts are persons who do not have any medical training or experience. However, 74.42 percent of the studies supported this purpose (58.14% studies supported only this purpose, 11.63% combined it with forming knowledge base, and 1 study combined it with improving object detection), and this is due to the attempt to remove the burden of tasks from overworked medical experts through crowdsourcing HbC systems. Meanwhile, the cost of solving these tasks is decreased by hiring non-expert workers. The next purpose is forming a knowledge database, and 23.26 percent of studies supported this purpose (20.93% studies supported only this purpose and 1 study combined it with object detection). Forming a knowledge base is significant to train the automatic systems for image analysis tasks. This task is tedious and time consuming for a few workers. Therefore, distributing it to a vast number of people renders it easy and enjoyable to solve. Another purpose that is supported by 20.93 percent of the studies is improving the object detection of automatic systems by providing them with feedback (6.98% studies supported only this purpose and 11.63% supported both this and the first purpose, and 1 study supported it in combination with the second purpose). The feedback is of great significance because it enables the results of the algorithm to be improved, as well as improving the algorithm itself such that it can perform better in the future tasks.

RQ2. What kind of medical image analysis problem is solved?

Figure 5 presents the results of the *RQ2*. The findings show three primary tasks are typically solved by humans for image analysis problems: annotation, segmentation, and classification tasks. Classification tasks are defined as choosing a class of either a whole image or a segmented object. Segmentation tasks require the contour outlining of the object in the image. Annotation tasks add labels to the images or objects in the images. The difference between classification and annotation tasks is that classification tasks require only one class name to be associated with the image or

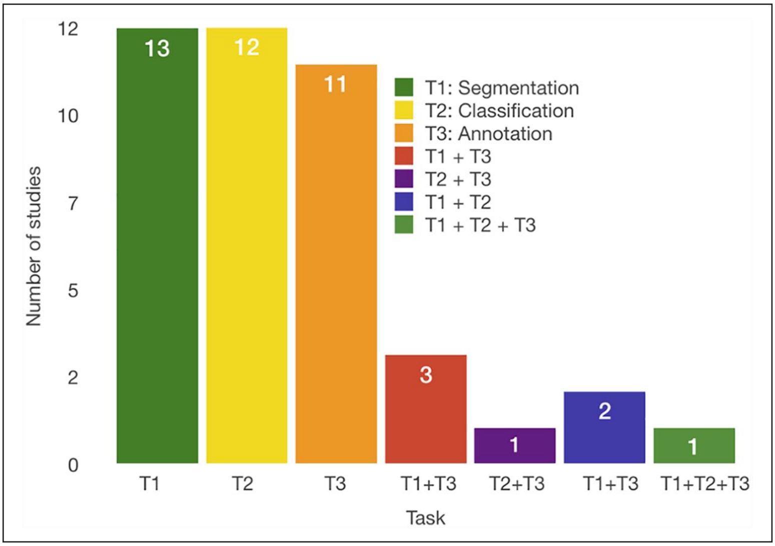

Figure 5. Crowdsourcing HbC tasks resulted by answering RQ2.

object in the image, while annotation tasks require more information from the user about that image or object. Annotation tasks can be cell or error detection (S9, S24). Other annotation tasks require answering the questions about image features (S3, S8, S27). Although in S21 the authors mentioned annotation tasks that require deciding whether polyps appeared in the image, we decided that this task is actually a classification task. Few studies explore more than one human computation task (S8, S9, S13). That is, 44.19 percent of the studies pertain to segmentation tasks (30.23% support only this task, 13.95% in combination with other tasks), 37.21 percent to classification (27.91% only classification tasks and 4.65% in combination with other tasks), and 37.21 percent to annotation tasks (25.58% only this task and 11.63% in combination with other tasks). These results are consistent with the RQ1 results because the crowdworkers typically perform the segmentation and classification tasks to replace the experts, and the classification and annotation tasks are primarily performed to form a knowledge database.

RQ3. What methods are used for the evaluation of the crowdsourcing worker's confidence?

The evaluation methods can be separated into three groups. The first group consists of pre-evaluation methods that are used before the completion of the tasks, the second requires quality control during the job completion, and the last one consists of post-evaluation methods that are used after the task completion.

The pre-evaluation methods concern the filtering of the crowdworkers. Typically, the worker must satisfy the requirements before conducting the tasks. AMT provides settings for a number of previously finished HITs and the overall approval rate, while FE categorizes the contributors based on their experience and provides the option to choose between three worker levels, where level 3 signifies the most trusted worker. Furthermore, both FE and AMT provide options for selecting contributors based on their location. Another means of quality control is the qualification test before the task, where the job designer can indicate the minimum accuracy for passing the test. FE also provides a setting for minimum time spent per task, so the contributors who spend less time working on the task are excluded. AMT has options for selecting workers based on different factors. Also, AMT provides custom qualifications where the job creator can indicate other qualifications needed for the task completion. 30.23 percent of the studies used overall approval and 34.88 percent used tests to exclude the underachieving contributors. To improve the results, in 79.07 percent of the studies, the researchers provided tutorials and examples that helped the workers to obtain more insights into the problem. Studies S4, S6, S7, S11, S28, and S34 showed that this is reasonable because the results demonstrate significant improvements in groups with previous training.

The quality control during the task execution concerns the identification and removal of spam workers. Studies S9, S18, S34, S35, and S38 used control objects during the task completion that need to be processed. If a user does not maintain the minimum accuracy on control tasks, he or she will be banned from further participation. S18 additionally prevents users from participating by measuring the frequency of answers such as "Not sure" and "Other." If 17 percent of a user's answers belong to that specific group, the user is blocked. S3 does not prevent users from participating but uses control images to give users feedback during job completion to improve their performance. S36 uses control images to calculate user's confidence and then aggregate the answers with other participants. S3 attempted to motivate users to perform better by giving them a bonus reward of US$5 for high-quality work.

The post-evaluation methods consist of either aggregating the results or further filtering the workers. Regarding the aggregation of results, we identified a few methods such as majority voting (51.16% of the studies), where the best results are chosen based on the number of times it occurs, and confidence maps (11.63% of the studies). An example of a confidence map for the segmentation task is an image with aggregated segmentation of the crowd presented with pixel intensities. This map provides a better analysis of the crowd results and the selection of the most reliable one. Further filtering is typically performed by setting a threshold for the specific measurement or defining the users' reliability (25.58% of the studies) by calculating Pearson's coefficient (S19), Spearman's coefficient (S27, S36, S41), Cohen's kappa coefficient (S13), or identifying the user's ranking using methods such as simultaneous truth and performance level estimation (STAPLE) (S29), and image-aware STAPLE (iaSTAPLE) (S32). Figure 6 shows the number of studies that implemented methods to evaluate users before task completion, Figure 7 shows the methods that belong to the second group and Figure 8 to the third group. From these results, we can conclude that it is advisable for the crowdsourcing HbC system to include at least one form of user evaluation. It is imperative to obtain the pre-evaluation of users to not waste time and money on cheaters and unreliable users. For a better performance, it is further advisable to use aggregation methods because the whole idea of crowdsourcing HbC systems is that one can incorporate many inputs for the same task and obtain the best solution for it.

RQ4. What are the types of information provided by the user?

We identified five types of information that a user can offer. Figure 9 shows these types. The most common types are the object's outline with support of 41.86 percent. Next is the class name to which the object belongs with 32.55 percent of support. The next type is filled form with 11.63 percent of studies. Coordinates of a point are the next type, typically the center of the segmented object, and it appeared in 9.30 percent of the papers. Furthermore, one study acquired the area of the object and one study acquired the bounding box of the object. These results agree with the different purposes and tasks of crowdsourcing HbC (RQ1 and RQ2). For the segmentation tasks, the essential information is the object's contour, and for classification it is the class name. The object's

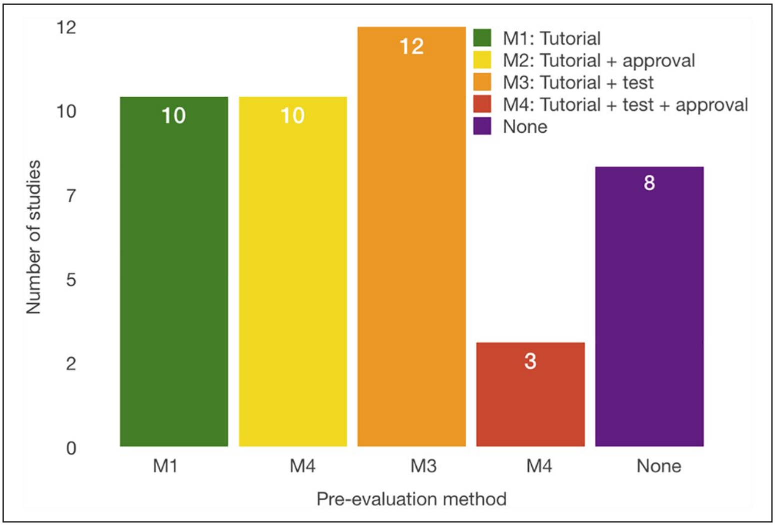

Figure 6. Evaluation methods that are used before task completion. "None" signifies that no method is used for the worker's pre-evaluation.

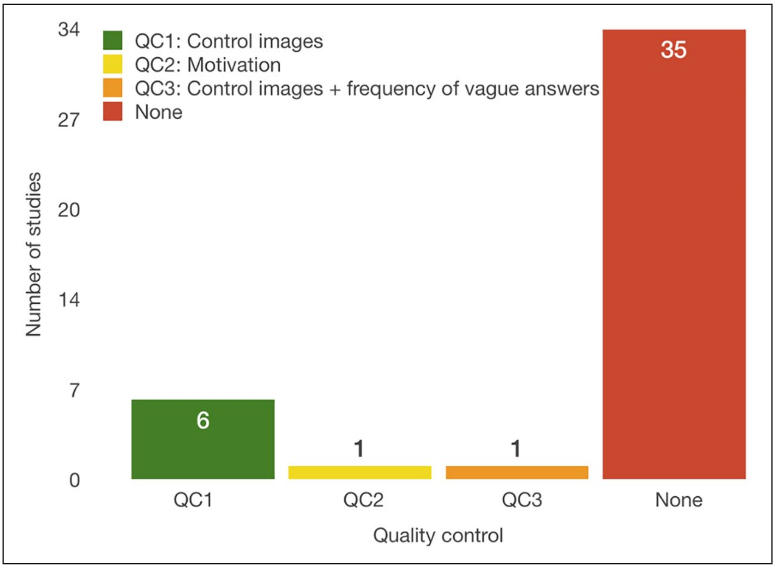

Figure 7. Methods for quality control during the task completion.

outline is used for the annotation tasks when the outline is a rectangle or a circle and does not represent the exact contour of the object. Point coordinates are required in both segmentation and annotation tasks. The object's area is used for segmentation tasks. A filled form is used in the annotation tasks for image labeling.

Given that automatic tools are not perfect in areas where the human vision system can perform outstandingly, it is not surprising that crowdsourcing HbC systems require various types of

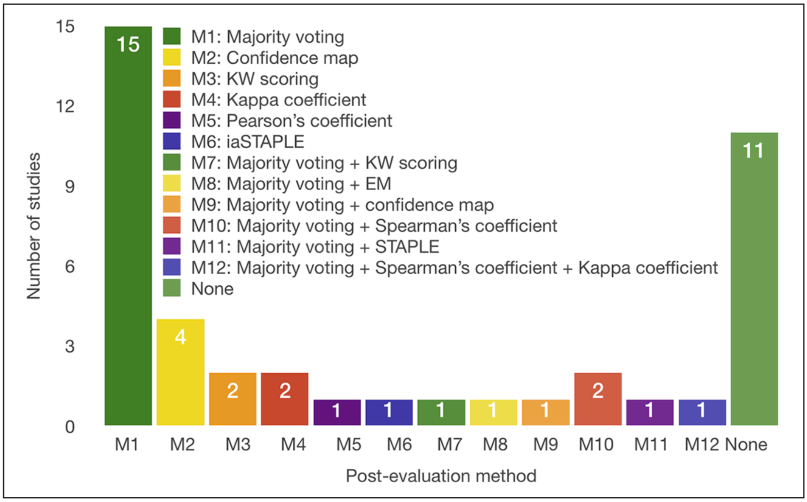

Figure 8. Evaluation methods used after task completion.

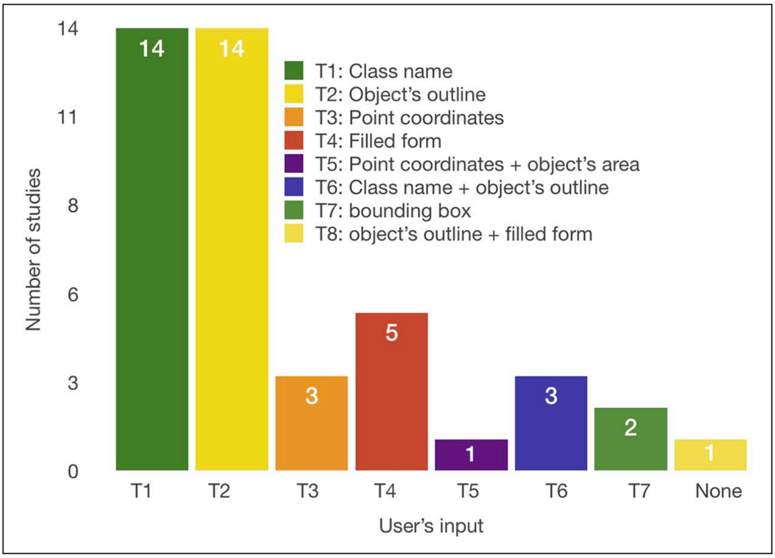

Figure 9. Types of information provided by users.

information that we found to be primarily used. The most significant challenge in gathering this information is misclassification in the classification tasks, and under or over-segmentations in the segmentation tasks. This problem can be solved in different approaches such as the pre-evaluation and post-evaluation of users, elimination of the cheater input, or choosing the right crowd for the specific assignment, as discussed previously in RQ3.

RQ5. What metrics are used for the evaluation of the obtained results?

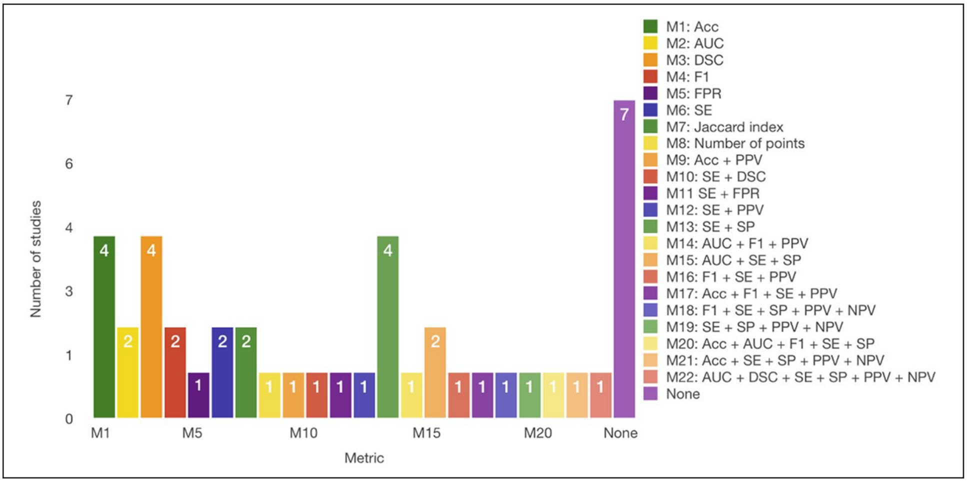

Figure 10. Used measures for evaluation of HbC system performance.

Figure 10 shows the measures reported in the selected papers. Sensitivity (SE) is primarily used, which appeared in 18 studies (41.86%), followed by accuracy (ACC) (37.21%), specificity (SP) (25.58%), precision (PPV) with 20.93 percent and both area under receiver-operating characteristic curve (AUC) and F1-measure (F1) with 16.28 percent. In addition, dice score (DSC) is used with a frequency of 13.95 percent and negative predictive value (NPV) with 9.30 percent. The least used measures were the false-positive rate (FPR) (S26, S39) and Jaccard index (S22, S23). The identification of the previously mentioned metrics as the most typically used is well expected because these are primarily used for the result evaluation in image segmentation problems. Furthermore, for classification and annotation tasks, several measures are typically chosen. They are the receiver-operating characteristic (ROC) curves, SE, SP, and ACC. In addition, F1 is typically used as a combined metric of precision and recall. It is especially useful when comparing several systems' performances (S9). It is recommended to use measures such as SE, SP, and F1 for the evaluation of classification tasks, whether they are conducted automatically or manually. This is because they are more reliable than accuracy itself, as they can provide more insight into the classification performance of individual classes. We noticed that although authors used the essential measures such as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), their values were not reported in the papers. These measures are the most important ones and they enable the derivation of other important measures, such as recall, SE, and F1. Furthermore, publishing their values enables the comparison with other studies' results that have not used the same derived metrics.

Meanwhile, for the segmentation tasks, the appropriate measures are different compared to the classification and annotation tasks. Apart from the SE, SP, and F1, more measures that are useful for the segmentation evaluation are available: overlap, pixel accuracy, and DSC. These metrics appear logical from the difference in the outline tasks, that is, to delineate some objects from the image. Hence, it is essential to measure the overlap between the outlined object and the ground truth, or even the pixel accuracy of that object. Even the number of points of the segmented contour can be relevant to exclude the under or over-segmented objects (S4).

As for the values of previously mentioned measures, Table 1 shows the comparison between different tasks and the HbC performance. From these results, we conclude that the workers perform

Table 1. The comparison between tasks and HbC performance.

Study	Measure	Task	Results (%)	Study	Measure	Task	Results (%)
S2	ACC	C	99	S2	SP	C	99
S6			72–77	S5			52–89
S11			64–71	S6			35–43
S16			92	S7			80–85
S28		A	95	S14			38
S40			81–95.1	S21			58–100
S2	SE	C	97.8	S34			100
S5			61–99	S9		S	97
S6			83–88	S20		A	66
S7			65–82	S28			71.3–83.3
S14			89	S31			86.5
S21			70–100	S2	PPV	C	96.6
S34			58–90	S5			49–100
S37			100	S9		S	86–89
S9		S	85–90	S24			23–72
S24			74–95	S20		A	88.7–92.4
S20		A	74	S20	NPV	A	63.6–72.7
S28			82.9–86.8	S9	F1	C	61
S31			90.4	S30			80–100
S35			87.88	S9		S	87–88
S38			80	S36			64–94
S6	AUC	C	65.6–93.8	S40			53.3–94
S15			86	S28		A	70
S34			89	S29	DSC	C	80.6
S1		S	85	S11		S	93
S3		A	50.1–97.6	S25			85
S28			95	S43			93.8

C: classification; S: segmentation; A: annotation; ACC: accuracy; SE: sensitivity; SP: specificity; PPV: precision; AUC: area under the receiver-operating characteristic curve; F1: F1-measure; DSC: dice score; NPV: negative predictive value.

equally well in every type of task. The difference in the results is not caused by the type of task but by other circumstances.

> *RQ6*. What are the results of the comparison between crowdsourcing HbC and other types of approaches?

We identified three types of other methods to solve the medical image analysis problem. These problems can be solved by medical experts, automatic systems, and fusion systems. Medical experts are trained professionals for a specific task. Automatic systems use different computing and machine-learning methods and algorithms for solving specific problems. Fusion systems are systems that include automatic algorithms and human feedback to improve these algorithms. The experts or crowdsourcing HbC systems can participate in the feedback process of the fusion system. The comparison was conducted in reference to the already available ground truth (GT), and the experts' results that were compared to others were not included in it. Figure 11 shows their distribution of papers. The results of the comparison are shown in Table 2.

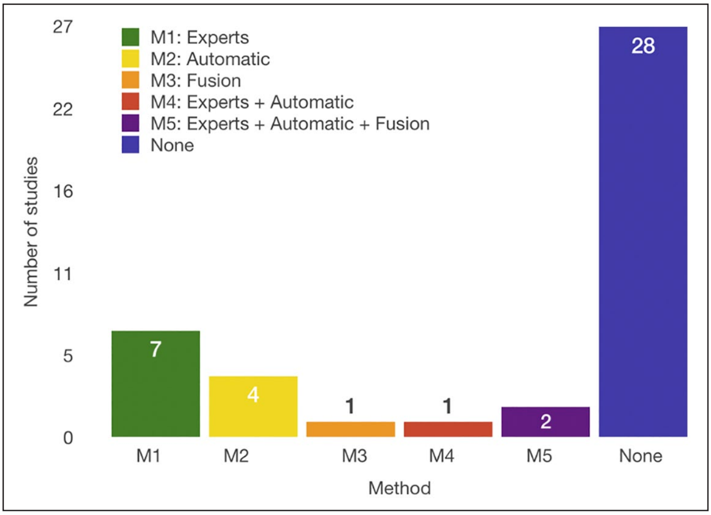

Figure 11. Number of papers that have compared the crowdsourcing HbC system performance with other available methods to solve the medical image segmentation problems: automatic systems, medical experts, and fusion systems. "None" signifies that a comparison was not conducted.

Table 2. Comparison of the systems' results.

Study (used metrics)	HbC (%)	Automatic (%)	Fusion (%)	Experts (%)
S1 (AUC)	85	86	91	–
S2 (ACC)	99	96	98	–
S8 (ACC)	86–89	89	–	94
S9 (ACC)	72–77	–	82–97	92
S12 (DSC)	82	36	–	85
S21 (SE)	65–82	–	–	86–92
(SP)	80–85			72–87
S31 (SE)	80	–	–	86.7
(SP)	86.5	–	–	87.2

AUC: area under the receiver-operating characteristic curve; ACC: accuracy; DSC: dice score; SE: sensitivity; SP: specificity; HbC: human-based computation.

Studies show that the experts are the best option for the highest performance, but contributors are not much worse than them. The reason is the simplification of the task; therefore, even untrained workers can effectively solve them. Furthermore, using the methods for aggregating the results mentioned in RQ3, the performance of untrained contributors can be further improved. In comparison with automatic systems, the crowdsourcing HbC systems proved to be equal or better in 75 percent of the cases. Fusion systems promise to be the best choice because they achieve better results than the crowdsourcing HbC systems and are cost-effective. These results show that untrained workers are highly reliable for annotation and segmentation tasks. Furthermore, we have to consider the methods that can improve the crowd results that did not appear in all of the previously compared papers. By including such methods, it could provide better results for the crowdsourcing HbC systems or even for the fusion systems that have

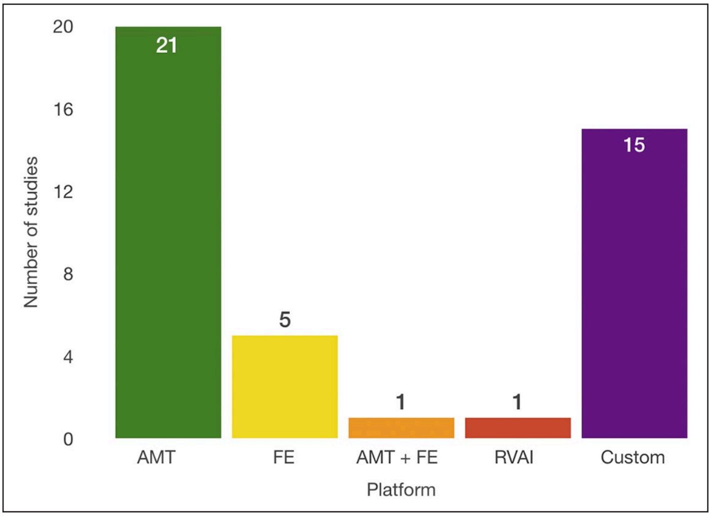

Figure 12. Crowdsourcing platforms used for solving medical image analysis problems. "Custom" signifies that the authors developed their own platform.

already been proven as the best choice. Studies S2 and S9 used the pre-evaluation of users, and studies S12 and S21 used the post-evaluation methods of the acquired data to achieve their results. Few papers reported experiments with various systems and processed the data with the correcting methods. Therefore, we do not possess the solid proof of them being effective in every case. Nevertheless, combining these results and the results of RQ3, we can assume that the performance of the systems described in studies S1 and S8 could be improved by including some corrective methods. In addition, not all studies used the same measures for assessing their results, nor were they trying to solve the same image segmentation problem. With this data, the comparison between systems was difficult and they were insufficient for a definite conclusion.

RQ7. What HbC/crowdsourcing platform is used?

RQ7 showed us two leading platforms that are used for crowdsourcing tasks. Figure 12 shows these platforms. Here, 51.16 percent of the studies used the AMT platform for the crowdsourcing tasks, and 13.95 percent used FE. One of them (S8) used both and one used Robovision AI platform. Furthermore, 34.88 percent of the studies used their own platform either in the form of serious game or an application for segmentation or annotation (S2, S4, S5, S13, S22, S25, S26, S28, S29, S36, S37, S39, S40, S41 and S42). Regarding AMT and FE, neither the workers' performance nor the final results of the finished tasks showed a significant difference.

For the shortest period of establishing the system and acquiring the results, some of the already established platforms that are proven to perform well for crowdsourcing HbC tasks are recommended. Meanwhile, if the development time is not an issue, the custom systems promise to be reliable, as well. They further have the advantage of more control during the system development. Using custom systems is especially advisable for gamesourcing systems, as studies S2, S25, and S26 showed that they have high performance.

RQ8. How many workers are required to obtain satisfying results, and at what cost and time frame?

Selected studies show that the number of workers can significantly vary while the results remain acceptable. Table 3 shows the results of nine papers that have compared the results obtained with different numbers of workers and different levels of experience evaluated with various pre-evaluation methods.

S2 shows that more workers increase the SE. Precisely, with more than 15 workers, the SE increased over 95%. Study S6 shows that with more workers, the results are not necessarily improved. Within the same level of experience, more workers that have higher levels of experience improve the results; however, this is not true with inexperienced workers. Here, fewer workers demonstrated much better results. Study S14 shows that a larger group of inexperienced workers exhibited better results than the smaller group; however, a difference is not observed with different numbers of experienced workers. Regardless of the number of people, S7 shows that experienced workers performed better than less experienced ones. Overall, all studies agree that not much difference is shown between experienced and inexperienced workers. In some cases, inexperienced workers can even over-perform experienced workers (S6). This is the case as shown by the SP values in S34, but it further shows that a smaller group of trained workers achieves a better SE than inexperienced workers. All of the studies show that a trade-off exists between the metrics, and it is closely related to the number of workers and their level of experience.

As for the post-evaluation methods, study S29 shows that multiple workers obtained better result than a single one, but with little improvement when the number of workers was more than 9. Study S24 shows that with the majority voting as the post-evaluation method, the SE increased, but the precision decreased with the increase in workers. Furthermore, study S8 shows that the best results obtained with the most efficient time and cost are with five workers. Studies S7 and S28 found that the optimal number of workers per image that achieved satisfying results is 10.

The total time of running the total number of tasks can vary. These times can be less than 24 h (S6, S5, S10, S11, S14, S16, and S34), approximately 2 weeks (S4 and S34) or a month (S8). From the total cost, we found that the cheapest experiment cost is US$1.21 (S19) for 22 workers. S27 shows that the cost of image labeling and nuclei labeling significantly differ because of the number and difficulty of the task. Therefore, they paid US$282 for image labeling and US$2820 for nuclei labeling. The first task required 4 h and the second 472 h to finish. S33 shows that training renders the workers more reliable, but consumes more time and money. Without training, 51 workers were paid US$84 for the total time of 40 min; with PowerPoint training, 19 workers were paid US$238 for 28 h; and with video training, 17 workers were paid US$213 for 18 days of work. Furthermore, S30 obtained acceptable results with a total cost of US$388.80 for 150 workers of 40 images. The cost of 1 HIT ranges between US$0.01 and US$0.50 in all studies, thereby rendering the crowdsourcing HbC systems highly cost-efficient.

Although not all the papers stated the criteria that are of our interest, we can conclude that by using the crowdsourcing HbC system, one can spend the minimum amount of money, and the minimum amount of time, yet still acquire enough data to achieve the satisfying solution to the initial problem.

Limitations of the review

This systematic review considered 43 studies to evaluate and assess the performance of various crowdsourcing HbC systems among themselves, and the automatic and fusion systems. The limitation of this review is that only 6 of the 43 studies compared the crowdsourcing HbC systems to one or more types of other approaches. Therefore, the comparison between different systems is not conclusive. While comparing the crowdsourcing HbC systems, each study uses different experimental settings that include datasets, pre-processing, post-processing and quality control methods, and performance metrics. The other limitation is that the study selection bias is a threat to the validity of this review. The selection of studies depends on the search strategy, literature sources, and selection criteria. We defined the search query using the RQs. Furthermore, we retrieved the relevant studies from eight electronic databases. However, some relevant studies may not use the terms related to the RQs. Therefore, we may have excluded these studies. To reduce this threat, we manually scanned the references and citation list of each relevant study to search for other relevant studies that were not included by the initial search. Although some studies have been missed, we believe that their number is small.

Guidelines for practitioners and scientists

The findings from conducting this systematic literature review inspired us to define some guidelines for practitioners and scientists. They can be further explored in crowdsourcing HbC systems for medical image analysis problems. From our discussion, we derived the following guidelines:

1. More studies for image segmentation should be conducted using the crowdsourcing HbC systems to obtain generalizable results. Few studies compared the crowdsourcing HbC systems with the experts, or the automatic and fusion systems. Therefore, it is necessary for future studies to compare the performance of these systems.
2. Few studies examined the effectiveness of crowdsourcing HbC systems that focus on improving object detection. Future studies may concentrate on this type of fusion systems.
3. It is advisable to use the general metrics and to compare them with the results of other approaches. The general metrics include TP, TN, FP, and FN. From these measures, other metrics can be derived. It is crucial to publish these results to provide a comparison with other studies.
4. Most of the studies did not use any pre-evaluation methods of the users. Future studies should consider some of these methods to improve their results. Furthermore, it is highly recommended to compare the results of pre-, post-evaluation and quality control methods to the experimental results without any evaluation methods. This would provide more insight into the importance of the evaluation methods and their development.
5. Every study must state the total number of workers, the number of workers per image, the time required for finishing the work, and the cost. This would provide a more precise comparison between the system's time and cost-efficiency.
6. A small number of studies used public datasets. We recommended publishing the dataset used, if possible. This would provide an accurate comparison between different approaches in solving the same problems.

The goal of these recommendations focuses on the fact that the field requires more robust research methods and more clarity in dissemination. Following these guidelines will have as a consequence more reproducible and comparable research.

Conclusion

This systematic review investigated crowdsourcing HbC systems for medical image analysis problems. We have conducted an extensive literature search of the relevant studies published until 2019 and identified 43 studies that are related to the eight RQs defined in this review. The primary findings are summarized as follows:

> *RQ1*. The three primary purposes of using crowdsourcing HbC systems for image analysis are as follows: conducting image analysis without medical experts, forming a knowledge base, and improving object detection. One that is widespread is the first one—conducting image analysis without medical experts. It is used to help medical professionals with tedious tasks that can be solved by untrained workers, thus enabling the medical professionals to explore and solve complex problems.
>
> *RQ2*. Three primary tasks are solved by human workers: segmentation, classification, and annotation. This is because humans are much better at solving these tasks than automated methods. Furthermore, it is cheaper for untrained workers to perform these kinds of tasks than to assign them to an expert, which could be a waste of time.
>
> *RQ3*. We identified three groups of methods that were used for the users' evaluation: pre-evaluation, quality control methods, and post-evaluation. The first group can eliminate unreliable users before the task completion. The second group requires quality control during the job completion. The third group focuses on the aggregation of the users' inputs and improving the results. All of these types are important to decrease the number of unreliable and cheating users and to improve the performance of the overall system.
>
> *RQ4*. We found several types of information that users provide. The types of information that are primarily used are the class name and the object's contour. This result is consistent with RQ2. The segmentation tasks require the object's contour, and the classification and annotations tasks require the class name of the object or the image.
>
> *RQ5*. We identified two groups of metrics for the tasks. For the classification and annotation tasks, the primarily used metrics are ACC, SE, SP, and F1. For the segmentation tasks, the additional metrics are overlap and pixel accuracy. All these measures are derived from the essential ones: TP, TN, FP, and FN. We found that the general metrics (TP, TN, FP, and FN) are primarily used but only published in a few studies.
>
> *RQ6*. Although the experts exhibited the best results, we found that the contributors demonstrated similar results for a much lower cost. It appears that the fusion systems have great potential as they provide better results with better time and cost-efficiency.
>
> *RQ7*. There are several platforms, but only three platforms were used in the works of HbC for medical image analysis included in this SLR: AMT, FE, and Robovision AI (only one study). No significant difference is found between them. The fourth option is a customized system that does not perform worse than the commercial platforms.
>
> *RQ8*. Our findings show that the untrained workers can perform as well as the experienced workers. They can even achieve better results depending on the task difficulty. A trade-off exists between SE and SP when comparing the trained and untrained workers. It is recommended to wisely choose the trade-off for each problem and to choose the correct level of the worker's experience. We found that the typical cost per HIT is from 3 to 5 cents. Experiments were even conducted for free. We concluded that using crowdsourcing HbC systems is highly efficient, considering the time and money spent, and the results achieved.

Table 3. Relation between the levels of workers' experience, number of workers, cost of tasks, and results; contributor levels 1, 2, and 3 are defined by the crowdsourcing platform, based on contributor's reliability.

Study	Metrics	Level of experience	Number of workers	Cost per HIT or total cost	Results (%)
S2	SE	Inexperienced	5	–	80
			10		90
			15–20		>95
S3	AUC	4 workers	88	US$0.01, US$5 bonus	83.6
		5 workers	88		88.9
S6	SE, SP	Inexperienced	69	US$0.03	99, 87
			152	US$0.03	98, 74
			127	US$0.03	99, 68
			72	US$0.05	96, 86
		Completed >5000 HITs with >99% approval	56	US$0.03	98, 89
			39	US$0.03	98, 85
			46	US$0.03	98, 52
			61	US$0.03	99, 64
S7	SE, SP	Completed >100 HITs with >97% approval	–	US$0.1	93.6, 67.8
		Improved training		US$0.1	100, 57
		Completed >500 HITs with >99% approval		US$0.15	100, 100
S8	ACC	Without specific level	1	US$0.05	72
			2	US$0.1	74
			3	US$0.15	77
			4	US$0.2	76
			5	US$0.25	77
S9	F1	Contributor level 1	–	–	60.87
		Contributor level 2			60.89
		Contributor level 3			66.41
S14	SE, SP	Inexperienced	78	US$0.05	88.80, 35.50
		Inexperienced	65		83.98, 43.97
		Completed >500 HITs with >90% approval	63		86.20, 39.79
			54		86.94, 26.10
S20	SE, SP, PPV, NPV	Nonmasters	25	–	87.5, 83.3, 92.9, 72.7
		Nonmasters with training	20	–	82.9, 77.8, 90.6, 63.6
		Masters	19	–	86.8, 71.3, 88.77, 68.3
S28	SE, SP, F1	Baseline tutorial	–	£7.5/h and volunteers	74, 66, 70
		Annotated images			80, 70, 76
		Interactive feedback			85, 60, 75
		Annotations + feedback			87, 59, 75
S34	SE, SP	Inexperienced	51	US$84	80.56, 66.67
		PowerPoint tutorial	19	US$238	91.67, 70.83

(Continued)

Table 3. (Continued)

Study	Metrics	Level of experience	Number of workers	Cost per HIT or total cost	Results (%)
		Video tutorial	17	US$213	97.22, 58.33
S35	SE	Unexperienced	143	US$0.30	90.4

HIT: human intelligence task; SE: sensitivity; AUC: area under the receiver-operating characteristic curve; SP: specificity; ACC: accuracy; PPV: precision; NPV: negative predictive value; F1: F1-measure.

From the findings of this systematic literature review, we derived guidelines for practitioners and scientists to help them improve their research on the topic. As a further work, we want to use the acquired knowledge in this SLR to design, implement, and validate crowdsourcing HbC systems for medical image analysis in order to diagnose support applications and to improve automatic medical image analysis.

Acknowledgements

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: We acknowledge the Ministerio de Economía, Industria y Competitividad (MINECO), the Agencia Estatal de Investigación (AEI), and the European Regional Development Funds (ERDF) for their support for the project TIN2016-81143-R (MINECO/AEI/ERDF, EU). We also acknowledge the Govern de les Illes Balears for its support for the projecte PROCOE/2/2017. N.P. benefited from the fellowship Euroweb + Program (ERASMUS-MUNDUS).

Author contributions

A.J. conceived the review. A.J. and N.P. proposed the research methodology. N.P. selected the manuscripts. G.M.-A. and J.V. reviewed the manuscripts. A.J. and N.P. applied the quality assessment. N.P. collected data. All authors analyzed the data and participated in the writing and the revision of the manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Antoni Jaume-i-Capó iD https://orcid.org/0000-0003-3312-5347

Supplemental material

Supplemental material for this article is available online.

References

1. Pham DL, Xu C and Prince JL. A survey of current methods in medical image segmentation. *Annu Rev Biomed Eng* 2000; 2: 315–337.
2. Xu Y, Zhu JY, Chang EIC, et al. Weakly supervised histopatholgy cancer image segmentation and classification. *Med Image Anal* 2014; 18: 591–604.

3. Balafar MA, Ramli AR, Saripan MI, et al. Review of brain MRI image segmentation methods. *Artif Intell Rev* 2010; 33: 261–274.
4. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal* 2017; 42: 60–88.
5. Sharma N and Aggarwal LM. Automated medical image segmentation techniques. *J Med Phys* 2010; 35(1): 3–14.
6. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imaging Graph* 2007; 31(4–5): 198–211.
7. Hesamian MH, Jia W, He X, et al. Deep learning techniques for medical image segmentation: achievements and challenges. *J Digit Imaging* 2019; 32(4): 582–596.
8. Shen D, Wu G and Suk HI. Deep learning in medical image analysis. *Annu Rev Biomed Eng* 2017; 19: 221–248.
9. McGuinness K and O'Connor NE. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognit* 2010; 43: 434–444.
10. McGuinness K and O'Connor NE. Toward automated evaluation of interactive segmentation. *Comput Vis Image Underst* 2011; 115: 868–884.
11. Olabarriaga SD and Smeulders AWM. Interaction in the segmentation of medical images. *Med Image Anal* 2001; 5(2): 127–142.
12. Fatima M and Pasha M. Survey of machine learning algorithms for disease diagnostic. *J Intell Learn Syst Appl* 2017; 9(1): 1–16.
13. Vezhnevets V and Konouchine V. GrowCut—interactive multi-label N-D image segmentation by cellular automata. In: *Graphicon*, 2005, pp. 150–156, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.8092&rep=rep1&type=pdf
14. Protiere A and Sapiro G. Interactive image segmentation via adaptive weighted distances. *IEEE Trans Image Process* 2007; 16(4): 1046–1057.
15. Snow R, Connor BO, Jurafsky D, et al. Cheap and fast—but is it good? Evaluation non-expert annotations for natural language tasks. In: *Proceedings of the 2008 conference on empirical methods in natural language processing*, 2008, pp. 254–263, https://www.aclweb.org/anthology/D08-1027.pdf
16. Von Ahn L. Human computation. In: *2008 IEEE 24th international conference on data engineering*, 2008, pp. 1–2, https://doi.org/10.1109/ICDE.2008.4497403
17. Quinn AJ and Bederson BB. Human computation: a survey and taxonomy of a growing field. In: *Proceedings of the international conference on human factors in computing systems (CHI 2011)*, 2011, pp. 1403–1412, https://doi.org/10.1145/1978942.1979148
18. Howe J. The rise of crowdsourcing. *Wired Magazine* 2006; 14(6): 1–4. https://www.wired.com/2006/06/crowds (accessed 17 July 2018).
19. Good BM and Su AI. Crowdsourcing for bioinformatics. *Bioinformatics* 2013; 29: 1925–1933.
20. Paolacci G, Chandler J and Ipeirotis P. Running experiments on Amazon Mechanical Turk. *Judgm Decis Mak* 2010; 5: 411–419.
21. Cooper S, Khatib F, Treuille A, et al. Predicting protein structures with a multiplayer online game. *Nature* 2010; 466(7307): 756–760.
22. von Ahn L, Maurer B, McMillen C, et al. reCAPTCHA: human-based character recognition via web security measures. *Science* 2008; 321(5895): 1465–1468.
23. Carlier A, Marques O and Charvillat V. Ask'nSeek: a new game for object detection and labeling. In: Fusiello A, Murino V and Cucchiara R (eds) *Computer vision—ECCV 2012. Workshops and demonstrations*. Berlin; Heidelberg: Springer, 2012, pp. 249–258.
24. Von Ahn L, Liu R and Blum M. Peekaboom. In: *Proceedings of the SIGCHI conference human factors in computing systems (CHI '06)*, 2006, https://doi.org/10.1145/1124772.1124782
25. Von Ahn L. Games with a purpose. *Computer* 2006; 39: 92–94.
26. Von Ahn L and Dabbish L. Labeling images with a computer game. In: *Proceedings of the SIGCHI conference on human factors in computing systems (CHI '04)*, 2004, pp. 319–326, https://doi.org/10.1145/985692.985733

27. Crequit P, Mansouri G, Benchoufi M, et al. Mapping of crowdsourcing in health: systematic review. *J Med Internet Res* 2018; 20(5): e187.
28. Kitchenham B and Charters S. Guidelines for performing systematic literature reviews in software engineering version 2.3. *Engineering* 2007; 45: 1051.
29. Critical Appraisal Skills Programme. CASP systematic review checklist, 2017, http://www.casp-uk.net/checklists (accessed 17 July 2018).
30. Dybå T and Dingsøyr T. Empirical studies of agile software development: a systematic review. *Inf Softw Technol* 2008; 50: 833–859.