Crowdsourcing-Based Segmentation of Medical Images

Danni Beaulieu
Computer Science and Engineering
Washington University in St Louis
b.danni@wustl.edu

Kaushik Dutta
Imaging Science
Washington University in St Louis
kaushik.dutta@wustl.edu

Abstract

Accurate segmentation of medical images is extremely crucial in diagnosis, prognosis, and subsequent treatment planning in modern clinical practices. In recent times, algorithms based in machine-learning have shown great potential in performing automatic segmentation. One major limitation of these methods is that it requires large amounts of accurately annotated data for training which is not only time and labor intensive but also costly and requires domain expertise to obtain. The guiding hypothesis of this paper is to analyze whether a crowd-powered algorithm can effectively annotate large scale medical image databases. To that end, we formulated a crowd-source segmentation framework to address three different hypotheses: I) whether crowdsourcing is effective in producing segmentation labels with quality similar to radiologists, II) whether the performance of the crowd correlates to self-evaluated rating of difficulty level, and III) whether interactions between workers actually improve performance. We found that aggregation via majority vote holds promise, and that while the expectation-maximization algorithm STAPLE experiences convergence issues with small samples, it warrants further study with bigger sample sizes. Difficulty rating and performance may appear to have some correlation. Worker collaboration yields improved results with some subjects but not others, although it still reduces overall variability on par with the STAPLE algorithm.

1 Introduction

Crowdsourcing has proven to be an effective method for leveraging collective intelligence of workers to solve labor-intensive computer vision tasks like image labelling and image annotation (1; 2; 3; 4; 5). The emergence of large datasets has helped to fuel the development and training of state-of-the-art image classification and segmentation techniques for computer vision problems (6; 7). However, within the realm of medical imaging, the advances are limited by availability of large-scale annotated datasets (8; 9). Generating high-quality data annotations for medical imaging is time-consuming and expensive due to scarcity of domain experts. In this project, we explore the use of crowdsourcing as an alternative method for segmenting glioblastoma tumors from brain MRI images. To that end, we developed a framework for gathering annotations from crowdworkers, applying label-aggregation algorithms to optimize the performance, and finally, comparing the performance to ground truth.

1.1 Motivation

In current practice, the task of labelling medical images relies heavily on domain experts and even moreso in cancer detection and diagnosis. Unfortunately, access to domain experts for completion of these tasks is intrinsically either limited and/or expensive. Advances in deep-learning based algorithm have shown great efficacy in automatic tumor segmentation in clinical setting, but these

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

techniques require large-scale accurately annotated datasets to train (10; 11; 12). Through this project, we investigate the potential of leveraging non-expert workers via crowdsourcing for accurate segmentation of medical imaging modalities. Ideally, this will facilitate a more rapid diagnosis and treatment process.

1.2 Related Works

Inspired by the success of crowdsourcing in high-quality image annotation for everyday objects (13; 2), it is increasingly being adopted for medical imaging tasks. The most important aspect of crowdsourcing in medical image annotation is ensuring task design guides the interaction between the annotator and the underlying objective. It is mainly applied to medical imaging for classification, rating and localization, or segmentation. Classification tasks include assigning labels to particular regions-of-interest (ROI) or to the whole image based on abnormality detection, visual attributes, and image quality/distortions for different modalities (14; 15; 16). Rating tasks were designed in the form of identifying presence/absence of specific features (17), counting cells (18), assessing intensity of staining (19), and discriminating disease types through scoring mechanisms (20). In segmentation-based tasks, crowdworkers usually draw bounding boxes or outlines for creating new annotation maps in applications ranging from CT and MR organ segmentation to nuclei segmentation for histopathology images (21; 22; 23; 24; 25). Another approach was to utilize the crowd to update or evaluate existing annotations as a means to indicate the location of errors in segmentations produced by automated algorithms (26; 27).

One of the main benefits of crowdsourcing is the fast and cost-effective collection of large-scale data which can then be aggregated to generate noise-free labels. Majority voting is the most common aggregation algorithm owing to its computational and conceptual simplicity (28; 29; 30). While Keshavan et al. (16) utilized an advanced version of majority voting called weighted voting by utilizing XGBoost algorithm to estimate annotators weights, Brady et al. (31) estimated annotators weight by a probabilistic approach based on the difficult level of the task. Expectation maximization (EM) algorithms provide a more probabilistic approach to label aggregation assuming heterogeneity among worker skill and are applied in context of medical image segmentation (32; 33; 34). Simultaneous Truth and Performance Level Estimation (STAPLE), based on the EM algorithm principle, considers a collection of segmentations and computes a probabilistic estimate of the true segmentation weighted against a measure of the performance level represented by each segmentation (35; 36; 37). Understanding behavioral patterns of human annotators is important to optimize and evaluate the performance of crowdsourcing frameworks and is explored in different studies (38; 39; 40), although very few studies correlated the user behaviour patterns with their quality of annotations. Heim et al. performed a regression analysis to estimate the quality of annotations with respect to features extracted from click-stream patterns (41). Similarly, Gurari et al. analyzed the correlation between annotation performance and annotation time (42).

1.3 Aims and Objectives

In this study we aimed to develop a framework for utilizing crowdsourcing with three different objectives:

I) Achieve effective segmentation of glioblastoma tumors from brain MRI. To that end, we generated multiple manual segmentation maps using the crowdworkers and label-aggregation algorithms (Majority Voting and STAPLE) were implemented to aggregate the different labels. Further, the performance of the framework was evaluated according to the radiologist (expert) segmentation maps.

II) Discover any correlation between individual performance and self-evaluation of the task. Each participant was asked to provide a difficulty rating on a scale of 1 (easy to detect and segment the tumor) to 5 (difficult to detect and segment the tumor).

III) Understand whether interactions between crowdworkers increase performance in niche tasks like medical image segmentation. A human-collaborative component was analyzed in which two workers were paired up for a team segmentation session to inform each other's perspectives.

2 Methods

We intended to design or find a user-friendly segmentation tool for effective segmentation of regions-of-interest from medical images. We implemented the the BRATS glioblastoma (brain tumor) dataset (43) hosted by the Medical Imaging Segmentation Decathlon. We planned to deploy the tool to a small group of inexperienced workers (our classmates) for annotating the region-of-interest following a brief description and tutorial of the task. The segmentation task could be either semantic (all instances should be labeled as a one group) or instance-wise (each instance should have its own label). For simplicity, we chose to perform a binary semantic segmentation. An additional classification task of interest was separating the datasets into "easy" and "hard" images or some similar spectrum of difficulty. This could act as a rather casual categorization for the purposes of detecting the value of future work along these lines. Further, we also included a human-collaborative component to our study where two workers would be allowed to engage in discussion while annotating a case.

2.1 Experiments

We utilized the available MedSeg AI tool for obtaining the segmentation in our data gathering phase. Our utilized dataset for the individual segmentations consists of a single subject with brain glioblastoma. Both the input image as well as the output segmented mask were processed using the nifty file format widely utilized in large scale medical image processing. We chose a single subject with variable number of slices because we felt that would be a good representative as an introductory task to be delegated to the crowd for segmentation rather than just giving complex images from various subjects; we chose eleven slices. For the pair segmentation component, we used two more glioblastoma subjects in their entirety (all slices).

One of the challenges of this BRATS dataset was that it had multiple image contrasts in different dimensions. The multi-contrast acquisition is helpful for the radiologist in diagnosing different biological phenomena in parallel to tumor segmentation, but due to this multi-contrast acquisition of the dataset, the tumor was not evident in various dimensions. As tumor segmentation is our primary task utilizing the crowd, we implemented a preprocessing algorithm to select and optimize the dimensions with the best tumor contrast to get the best results from the crowdsourcing workers.

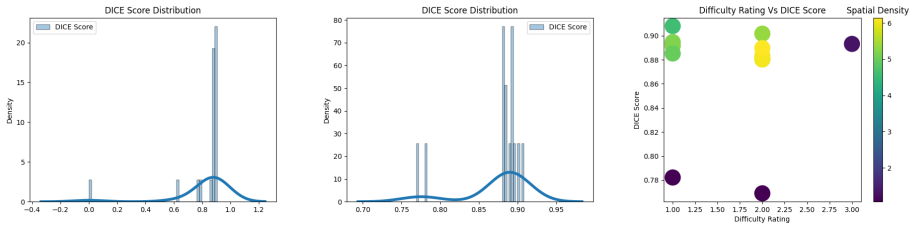
2.2 Task Design

In presenting the task, we provided a video and live demonstration for worker convenience. This included instructions on how to navigate to and use the MedSeg AI tool, how to download the dataset, and how to upload the segmented slices. In our classroom experiment, we allotted approximately 5 minutes for the segmentations. We also requested workers complete a Google form alongside their task, and asked them to rate the difficulty of the task given to them on a scale of 1-5, with 1 being the easiest and 5 being the most complex. For the collaborative component two additional subjects were segmented by two crowd-source workers individually. Then, after a week they again segmneted the same subjects in un-timed interactive discussion setting. The segmentations were treated with one of two algorithms for aggregation of lables to decrease the inter-labeller noise: (1) majority vote and (2) STAPLE. Two metrics were utilized to determine the success of the algorithms: (1) DICE score and (2) Hausdorff distance.

2.3 Results

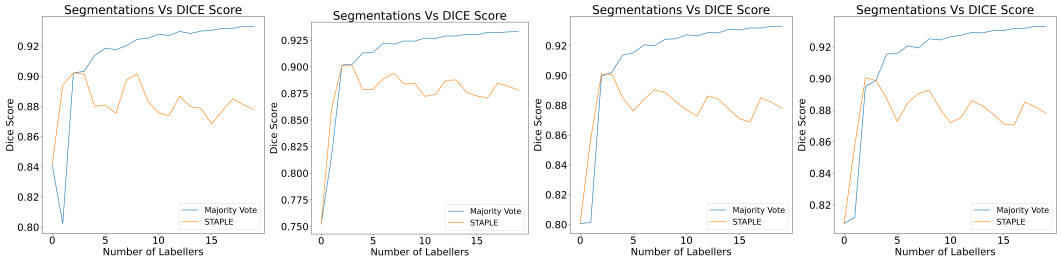
2.3.1 Individual Segmentation

We received a total of 19 submissions from the classmates and 2 segmentations were generated by us (the authors). Out of these 21 total collected labels, two labels are considered not to adhere to the segmentation standards. While one worker made no attempt to take the task seriously and segmented absurdly with line drawings, it appears most other workers were not "bad actors," and provided usable data for analysis. Another worker left all slices blank, although this could have been by accident. Aside from the empty submission, all other labels were utilized in the experiments for a final total of 20 segmentations. The DICE scores were mostly distributed between about 0.88 and 0.91 (figures 1a and 1b). We sampled these segmentations to discover their relationship to the two success metrics. Majority vote performed well, smoothly rising to produce a DICE score of almost 0.94 when all segmentations were included (figures 2, 4). STAPLE varied between about 0.88 and 0.9, decreasing



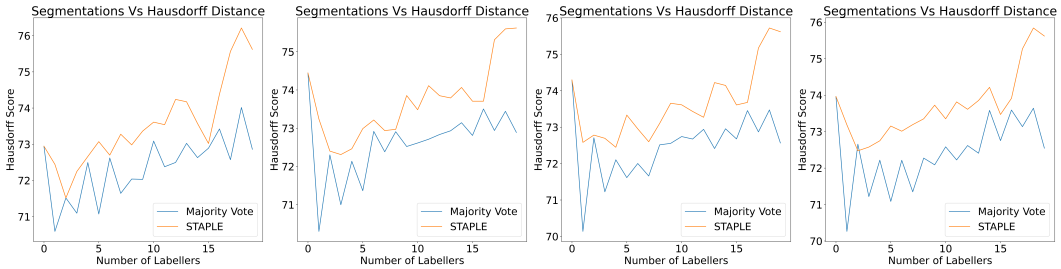
(a) All Participants (b) Survey Participants (c) Rating Vs DICE Score

Figure 1: DICE Score Distribution



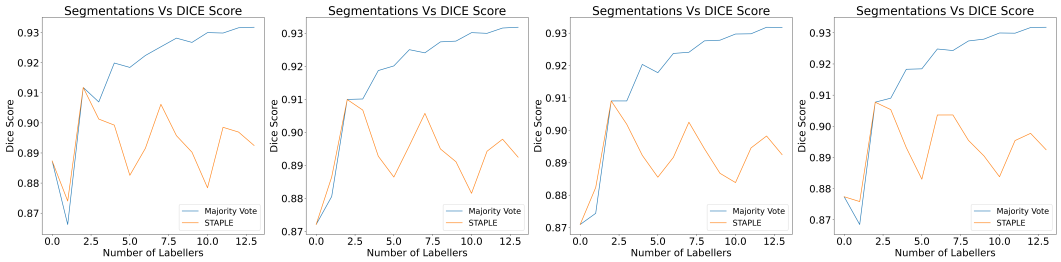
(a) 25 Iterations (b) 50 Iterations (c) 75 Iterations (d) 100 Iterations

Figure 2: DICE Score All Participants



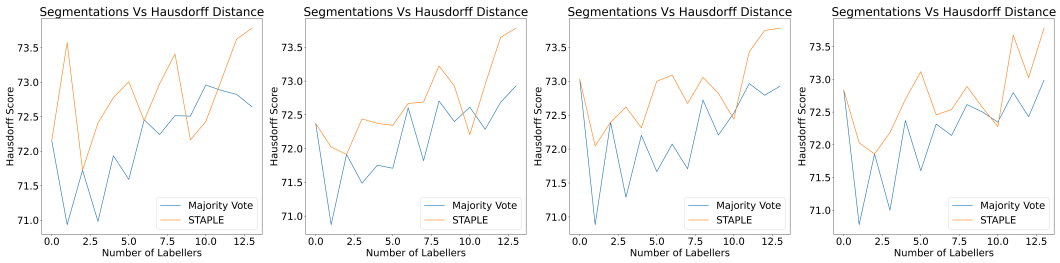
(a) 25 Iterations (b) 50 Iterations (c) 75 Iterations (d) 100 Iterations

Figure 3: Hausdorff Score All Participants



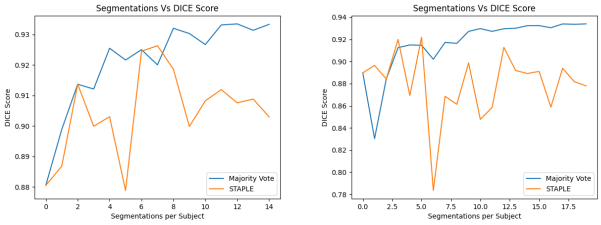
(a) 25 Iterations (b) 50 Iterations (c) 75 Iterations (d) 100 Iterations

Figure 4: DICE Score Survey Participants



(a) 25 Iterations (b) 50 Iterations (c) 75 Iterations (d) 100 Iterations

Figure 5: Hausdorff Score Survey Participants



(a) Highest 75% Performers (b) All Performers

Figure 6: Remove Lowest 25% of Performers

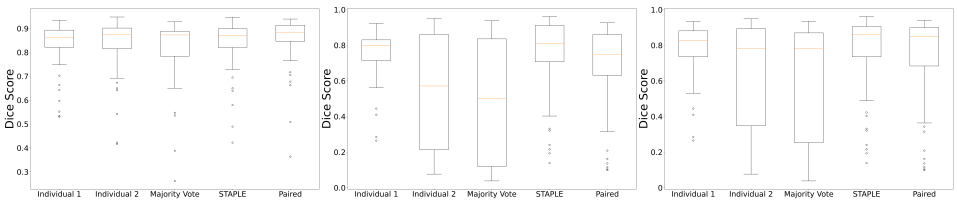
slightly overall as more segmentations were sampled in (figures 2, 4). The Hausdorff distance for majority vote varied between 72 and 73 once all segmentations were sampled in, while for STAPLE it ended near 74 for survey participants and near 76 when all participants were included (figures 3, 5). We also manually removed the lowest quartile performers and re-ran the algorithms; this yielded similar results (figure 6).

2.3.2 Rating Difficulty

We received 18 responses to the form, with 13 out of 18 rating the task a low difficulty of 1 or 2. Only 1 worker rated the task a 5, and the remaining 4 rated the task a 3. This result aligns with the goal of making the task accessible to workers with no prior experience in medical segmentation. Unfortunately, some workers filled out a survey but did not successfully upload segmentations, while others successfully uploaded segmentations but did not fill out the survey. For this reason, we show results both for all participants and for only survey participants. There were 14 workers who both filled out the survey and uploaded segmentations. In comparing the ratings and the DICE scores of survey participants, it appears there may be a correlation between higher ratings and lower DICE scores, although more data is needed (figure 1c).

2.3.3 Human Collaborative Effect

During the interactive component, the individual segmentations were accumulated using the two algorithms and this was contrasted with the pair segmentation using DICE score. Overall, the STAPLE



(a) Subject 1 (b) Subject 2 (c) Both Subjects

Figure 7: Collaborative Study

algorithm outperformed majority vote, and the pair segmentation was comparable to STAPLE (figure 7). In Subject 1, there was relatively low variability across all segmentations and accumulations, and the paired segmentation slightly outperformed the others (figure 7a). In Subject 2, there was a high degree of variability introduced by the worker "Individual 2"; this comes through in the majority vote result but is reduced in both the STAPLE and the pair result (figure 7b).

3 Discussion

Surprisingly, in the main phase of our experiments with our classmates as crowd workers, majority vote outperformed STAPLE by both the DICE and Hausdorff metrics. The sub-optimal performance of the STAPLE algorithm, which usually outperforms majority vote is likely due to our small sample sizes. This is discussed both theoretically and empirically in the paper, "A Cautionary Analysis of STAPLE Using Direct Inference of Segmentation Truth," which explores convergence issues in cases like ours (44). Scores for only survey participants were slightly higher than those for all participants. This indicates the survey may have acted as a kind of quality filter which helped remove workers who labelled segmentations but who did not really wish to actively participate. Similarly, marginal benefit for STAPLE was observed when removing the lowest quartile of performers. A possible correlation between higher difficulty ratings and lower DICE scores among survey participants was anticipated. The data appears that it may support this hypothesis, however this needs to be validated with further experimentation and higher sample sizes. For the pairwise experiments, the disparity between the two subjects reinforces the notion that subject complexity is critical and requires more in-depth inquiry among a spectrum of images. Specifically, pairwise segmentation outperformed STAPLE and majority vote only for Subject 1. This indicates network effects as studied here may be more effective for some types of images over others. The conduciveness of an image to pairwise or group justification warrants further study. Another interesting result is that STAPLE performed relatively well in the pairwise experiment in comparison to the classroom experiment. This may be due to lack of convergence issues with only two people.

3.1 Conclusions

This work shows aggregating crowdsourcing labels for medical image segmentation holds promise, particularly using majority vote. Human collaborative component though very subject dependent might provide some benefits in terms of decreasing the variability in delineation. Ensuring quality of workers and appropriate sample sizes is key to further examination in this area.

3.2 Future Work

There are several directions we can extend our work in the future. We chose binary segmentation for this project; in the future, we want to test our framework on multi-class segmentation for tumor grading. In the human collaborative component, we can introduce a justification framework, either live or text-based such that each labeller leaves a note or justification with the segmentation map which can be introduced to other labellers in case of disagreements. For this experiment, we chose a relatively easy case of segmentation and selected the slices which contain the tumor; in real cases, the radiologists select the slices. In the future, perhaps the crowdworkers could be additionally tasked with this selection process themselves. Since we saw variability in our collaborative study across subjects, we want to explore other forms of the cancer segmentation task across a spectrum of complexity and image modalities (PET imaging, optical etc). Further, we want to track the amount of time and number of clicks taken to complete the task. Finally, to boost our performance, especially when we deploy it to a large crowd, we would like to add a filtering mechanism on top of the aggregation algorithm. This could function either to filter out low quality workers or as a rectification algorithm to counterbalance the weaknesses of STAPLE or both.

[1]

[1]Project location: https://github.com/danni-beaulieu/human-in-the-loop-project

References

[1] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala, "Material recognition in the wild with the materials in context database," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3479–3487.

[2] Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, Kristen Grauman, et al., "Crowdsourcing in computer vision," *Foundations and Trends® in computer graphics and Vision*, vol. 10, no. 3, pp. 177–243, 2016.

[3] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[4] Christopher H Lin, Mausam Mausam, and Daniel S Weld, "Crowdsourcing control: Moving beyond multiple choice," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[5] Antonio Torralba, Bryan C Russell, and Jenny Yuen, "Labelme: Online image annotation and applications," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1467–1484, 2010.

[6] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao, "Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.

[7] Hossein Mobahi, Shankar R Rao, Allen Y Yang, Shankar S Sastry, and Yi Ma, "Segmentation of natural images by texture and boundary compression," *International journal of computer vision*, vol. 95, no. 1, pp. 86–98, 2011.

[8] Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[9] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[10] Wu Deng, Qinke Shi, Miye Wang, Bing Zheng, and Ning Ning, "Deep learning-based hcnn and crf-rrnn model for brain tumor segmentation," *IEEE Access*, vol. 8, pp. 26665–26675, 2020.

[11] Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li, "Deep learning-based image segmentation on multimodal medical imaging," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 162–169, 2019.

[12] Xiaomei Zhao, Yihong Wu, Guidong Song, Zhenye Li, Yazhuo Zhang, and Yong Fan, "A deep learning model integrating fcnns and crfs for brain tumor segmentation," *Medical image analysis*, vol. 43, pp. 98–111, 2018.

[13] Hao Su, Jia Deng, and Li Fei-Fei, "Crowdsourcing annotations for visual object detection," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

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

[15] Alba García Seco de Herrera, Antonio Foncubierta-Rodríguez, Dimitrios Markonis, Roger Schaer, and Henning Müller, "Crowdsourcing for medical image classification," *Swiss Medical Informatics*, 2014.

[16] Anisha Keshavan, Jason D Yeatman, and Ariel Rokem, "Combining citizen science and deep learning to amplify expertise in neuroimaging," *Frontiers in neuroinformatics*, vol. 13, pp. 29, 2019.

[17] Adam M Sonabend, Brad E Zacharia, Michael B Cloney, Aarón Sonabend, Christopher Showers, Victoria Ebiana, Matthew Nazarian, Kristin R Swanson, Anne Baldock, Henry Brem, et al., "Defining glioblastoma resectability through the wisdom of the crowd: a proof-of-principle study," *Neurosurgery*, vol. 80, no. 4, pp. 590–601, 2017.

[18] Peter Smittenaar, Alexandra K Walker, Shaun McGill, Christiana Kartsonaki, Rupesh J Robinson-Vyas, Janette P McQuillan, Sarah Christie, Leslie Harris, Jonathan Lawson, Elizabeth Henderson, et al., "Harnessing citizen science through mobile phone technology to screen for immunohistochemical biomarkers in bladder cancer," *British journal of cancer*, vol. 119, no. 2, pp. 220–229, 2018.

[19] Francisco J Candido Dos Reis, Stuart Lynn, H Raza Ali, Diana Eccles, Andrew Hanby, Elena Provenzano, Carlos Caldas, William J Howat, Leigh-Anne McDuffus, Bin Liu, et al., "Crowdsourcing the general public for large scale molecular pathology studies in cancer," *EBioMedicine*, vol. 2, no. 7, pp. 681–689, 2015.

[20] Sam Mavandadi, Stoyan Dimitrov, Steve Feng, Frank Yu, Uzair Sikora, Oguzhan Yaglidere, Swati Padmanabhan, Karin Nielsen, and Aydogan Ozcan, "Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study," *PloS one*, vol. 7, no. 5, pp. e37245, 2012.

[21] Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab, "Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1313–1321, 2016.

[22] Eric Heim, *Large-scale medical image annotation with quality-controlled crowdsourcing*, Ph.D. thesis, 2018.

[23] Alison Q O'Neil, John T Murchison, Edwin JR van Beek, and Keith A Goatman, "Crowdsourcing labels for pathological patterns in ct lung scans: can non-experts contribute expert-quality ground truth?," in *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pp. 96–105. Springer, 2017.

[24] Alberto Chávez-Aragón, Won-Sook Lee, and Aseem Vyas, "A crowdsourcing web platform-hip joint segmentation by non-expert contributors," in *2013 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2013, pp. 350–354.

[25] D Mitry, T Peto, S Hayat, JE Morgan, KT Khaw, et al., "Crowdsourcing as a novel technique for retinal fundus photography classification: Analysis," 2013.

[26] Sebastian Bittel, Veith Roethlingshoefer, Hannes Kenngott, Martin Wagner, Sebastian Bodenstedt, Tobias Ross, Stefanie Speidel, and Lena Meier-Hein, "How to create the largest in-vivo endoscopic dataset," *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, 2017.

[27] Melanie Ganz, Daniel Kondermann, Jonas Andrulis, Gitte Moos Knudsen, and Lena Maier-Hein, "Crowdsourcing for error detection in cortical surface delineations," *International journal of computer assisted radiology and surgery*, vol. 12, no. 1, pp. 161–166, 2017.

[28] Yu-chi Hu, Gikas Mageras, and Michael Grossberg, "Multi-class medical image segmentation using one-vs-rest graph cuts and majority voting," *Journal of Medical Imaging*, vol. 8, no. 3, pp. 034003, 2021.

[29] Antonis Daskalakis, Dimitris Glotsos, Spiros Kostopoulos, Dionisis Cavouras, and George Nikiforidis, "A comparative study of individual and ensemble majority vote cdna microarray image segmentation schemes, originating from a spot-adjustable based restoration framework," *computer methods and programs in biomedicine*, vol. 95, no. 1, pp. 72–88, 2009.

[30] Henrietta Toman, Laszlo Kovacs, Agnes Jonas, Lajos Hajdu, and Andras Hajdu, "A generalization of majority voting scheme for medical image detectors," in *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 2011, pp. 189–196.

[31] Christopher John Brady, Lucy Iluka Mudie, Xueyang Wang, Eliseo Guallar, and David Steven Friedman, "Improving consensus scoring of crowdsourced data using the rasch model: development and refinement of a diagnostic instrument," *Journal of medical Internet research*, vol. 19, no. 6, pp. e7984, 2017.

[32] Ramiro Donoso, Alejandro Veloz, and Héctor Allende, "Modified expectation maximization algorithm for mri segmentation," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2010, pp. 63–70.

[33] Mohamed Ali Mahjoub et al., "Image segmentation by adaptive distance based on em algorithm," *arXiv preprint arXiv:1204.1629*, 2012.

[34] Sergiu Nedevschi, Anca Ciurte, and George Mile, "Kidney ct image segmentation using multi-feature em algorithm, based on gabor filters," in *2008 4th International Conference on Intelligent Computer Communication and Processing*. IEEE, 2008, pp. 283–286.

[35] Simon K Warfield, Kelly H Zou, and William M Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *IEEE transactions on medical imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[36] Olivier Commowick and Simon K Warfield, "Estimation of inferential uncertainty in assessing expert segmentation performance from staple," *IEEE transactions on medical imaging*, vol. 29, no. 3, pp. 771–780, 2010.

[37] Olivier Commowick and Simon K Warfield, "Incorporating priors on expert performance parameters for segmentation validation and label fusion: a maximum a posteriori staple," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2010, pp. 25–32.

[38] Wedyan Alswiti, Ja'far Alqatawna, Bashar Al-Shboul, Hossam Faris, and Heba Hakh, "Users profiling using clickstream data analysis and classification," in *2016 Cybersecurity and Cyberforensics Conference (CCC)*. IEEE, 2016, pp. 96–99.

[39] Gang Wang, Tristan Konolige, Christo Wilson, Xiao Wang, Haitao Zheng, and Ben Y Zhao, "You are how you click: Clickstream analysis for sybil detection," in *22nd USENIX Security Symposium (USENIX Security 13)*, 2013, pp. 241–256.

[40] Clint Feher, Yuval Elovici, Robert Moskovitch, Lior Rokach, and Alon Schclar, "User identity verification via mouse dynamics," *Information Sciences*, vol. 201, pp. 19–36, 2012.

[41] Eric Heim, Alexander Seitel, Jonas Andrulis, Fabian Isensee, Christian Stock, Tobias Ross, and Lena Maier-Hein, "Clickstream analysis for crowd-based object segmentation with confidence," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2814–2826, 2017.

[42] Danna Gurari, Mehrnoosh Sameki, and Margrit Betke, "Investigating the influence of data familiarity to improve the design of a crowdsourcing image annotation system," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2016, vol. 4, pp. 59–68.

[43] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al., "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[44] Koan Van Leemput and Mert R. Sabuncu, "A cautionary analysis of staple using direct inference of segmentation truth," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, 2014, pp. 398–406.