Medical Monkeys: A crowdsourcing Approach to Medical Big Data

Lorenzo Servadei[1], Rainer Schmidt[1], Christina Eidelloth[1], Andreas Maier[2]
[1] University of Applied Sciences, Lothstraße 64,
80335 Munich, Germany
{lorenzo.servadei,c.eidelloth,rainer.schmidt}@hm.edu
[2] University of Erlangen-Nuremberg, Schloßplatz 4,
91054 Erlangen, Germany
andreas.maier@fau.edu

Abstract. Big data play a central role in eHealth and have been crucial for designing and implementing clinical decisions support systems. Those applications can avail on data analysis and response capabilities, often empowered by Machine Learning algorithms, which can help clinician in diagnostic as well as therapeutic decisions. On the other hand, in the context of eSociety, eCommunities can be essential actors for managing and structuring medical data. In fact, they can support in gathering, providing and labeling data. This last task is highly relevant for medical Big Data, as it is a key point for supervised Machine Learning algorithms, which need an extensive data annotation process. This improves prediction and analysis capabilities of the algorithms on large datasets. Our approach on the medical Big Data labeling problem is the design and prototyping of a crowdsourcing collaborative Web Application, used for the annotation of medical images, that we named Medical Monkeys. Under the principles of mutual advantage and collaboration researchers, online gamers, medical students and patients will be involved, within this platform, in a virtual and mutually beneficial cooperation for improving Machine Learning algorithms. Using our application on large scale data analysis, algorithms for image segmentation will become useful for clinical decisions support systems. Our application is the result of a collaboration of several universities and research institutes and has, as principal aim, the integration, in form of gaming tasks, of eCommunities for the implementation of a more accurate analysis and diagnostic on MRI or CT images.

Keywords: Free Innovation, Medical Big Data, Medical Innovation, eHealth, Gamification, Open Innovation, Medical Images Segmentation

1 Introduction

An exponential amount of sensors, monitoring, data storage systems, multimedia and devices has been generating what we identify as the Big Data phenomenon [1]. Big Data are large sets of data not merely outlined by their amount, but also by a higher degree of complexity as well as a larger value derived by the application of innovative analysis techniques [2]. Further than that, Big Data are characterized, in comparison to traditional Small Data, by a higher *velocity* and *variety* [3]. This means that they are often generated on real time (velocity) and composed by different sorts of data as images, audio, texts, etc. (variety) [3]. Due to their relevance for the business, such as the statistical trends forecasting and key indicators obtained, currently companies set more and more effort than ever in their gathering, storage and evaluation [4].

On the other hand, Machine Learning (ML) methods can extract meaning from Big Data as such [5]. Today this is already being applied in very different fields by all different kind of private companies [6]. But the establishment of own free data stocks for ML analysis is failing due to the high costs and the comparably low benefit for society. This complicates the role of research in universities, and take them out of a strategical role in the Big Data research [7].

With medical data such a data collection is not possible today, as the transfer of medical data is protected by law and usually complicated. Without explicit consent of the users, this would be highly unethical [8]. However, this is not seen in all countries, and there are also efforts all around the world to loosen these regulations [9].

In order to unify efforts of different actors and institutions for academic data gathering and annotation, we designed and prototyped a Web Application which avail of two pieces of collaborative software. A collaborative software is defined as an application whose intent is to realize a shared purpose, dividing the effort among users [10].This approach can generate solutions to common problems and thus create a solid base for innovation, generating a common innovation model [11]. The power of a collaborative software, if it is not directed towards economical compensation and comes from unpaid development, can be an essential part of a free innovation model - a project where the innovation designs are not being protected by the developers [12]. These constraints are motivated by the self-reward, which is based on benefits excluding compensated transactions, and is sometimes motivated by altruistic purposes [12]. Based on this path, we got inspired by two medical projects that have been created with the goal of a free sharing of innovations design and ideas, named *Patient-Innovation.com*[1] and *Nightscout*[2].

At this moment, to our knowledge, it does not exist any platform for extensive labeling of different medical images, which mantains their statistical analysis and results free and available for research scopes. Our paper presents an approach to develop such a platform and maintain labeled images within the academic community. Therefore, we propose requirements, research design and prototypes for a crowdsourcing Web Application, based upon the free innovation and collaborative paradigm, for labeling medical images. We explore the free cooperation among Web users, researchers and image donors as a gateway for enhancing the performance of ML algorithms on medical images. This will lead to better automatic segmentation and detection algorithms, improving clinical decisions support systems and reducing the human-based error on diagnostic and therapeutic evaluation [13].

In the following chapter of this paper, we will analyze first the literature sources which helped us to explore the main scientific areas of the project: medical Big Data and their analysis, free innovation patterns and collaborative innovation, crowdsourcing, crowdsourcing for medical images segmentation and its gamification.

In the third chapter, the core of our paper, we will proceed then enucleating our research design. We will first outline, through a use case scenario, the interaction among actors and the requirements for our medical images segmentation application.

After that we will enter the areas of data gathering and storage requirements of the system, where we will describe the functionalities of our distributed file system prototype. At last, we will briefly introduce the data analysis and algorithm evaluation step. In the fourth chapter we will then state the conclusions of the paper and the next steps for our project and research.

[1] https://patient-innovation.com/

[2] http://www.nightscout.info/

2 Literature

In Table 1 we summarized all the important pieces of literature for our research design.

Table 1. Selection of literature, categorized by main topic of interest.

Medical Big Data	Free and Collaborative Innovation	Crowdsourcing Innovation	Crowdsourcing for Medical Images Segmentation	Crowdsourcing, Gamification and Segmentation
[14] [15] [16] [17] [18]	[12] [19] [20] [21]	[22] [23] [24]	[25] [26] [27] [28]	[29] [30] [31] [32]

2.1 Medical Big Data

As a first step into our literature research, we deepened our knowledge on Medical Big Data. We first focused on their potential for medical analysis, gaining an overview over their importance for clinical diagnosis and research [14]. In particular, we looked towards the fundaments for the collection of medical imaging online, which is a sought-after topic in the scientific research [15]. On the side of data collection, the Internet of Things (IoT) and diffused sensors have been shown to be an important and pioneering step in the direction of data gathering for a multi-sources analysis [16]. But concerning the more related topic of Big Data in form of medical imaging, we evaluated some papers that introduce to their collection and analysis, which is the final purpose of the project [17] [18].

2.2 Free and Collaborative Innovation

On free innovation, the kind of innovation the project aims at, the research of Eric von Hippel provides a theoretical framework and pattern analysis [12]. In this work it is analyzed the motivation behind free innovators and, through a quantitative analysis on surveys over free innovators projects in Finland and Canada, it is provided an insight on motivation for free innovators [12] [19]. On the other hand, collaborative innovation relates to a cooperation in the prototyping and implementation of our project. Collaborative innovation is a very effective method to get a better contribution for different tasks from other individuals, enhancing the chances of excelling in a larger amount of assignments [21]. At the same time, a collaborative innovation tends to have a broader diffusion, reaching a larger share of potential users [20]. As a result, a collaborative work on the free innovation pattern can lead to an even better result, sharing the effort and the costs of the design and development and introducing new and effective ideas [12].

2.3 Crowdsourcing Innovation

For enhancing our perspective on performing the images segmentation, we deepened into the crowdsourcing innovation. For the development of free innovations, crowdsourcing gives the possibility of conveying additional expertise to the project: This could help to find better and more creative solutions to existing problems and features [22] [23]. The practical contribution of crowdsourcing in problem solving tasks comes from the vast experience and different background that a multitude of individuals takes along [22]. This mechanism explains the rising of crowdsourcing in solving specific tasks and long-term projects [24].

2.4 Crowdsourcing for Medical Images Segmentation

A general review of crowdsourcing for health-related tasks has been useful for a first orienteering within the technical aspects of this topic [26]. The first important fundament of our research is the effectiveness of crowdsourcing for medical image segmentation. The result of scientific works shows that a large crowd of non-expert can reach a high accuracy on image labelling for particular medical tasks, resulting comparable to the accuracy of expert medical doctors [27]. Crowdsourcing has shown its effectiveness for segmentation of medical images in tasks where the detection of particular entities has been required on large numbers of images. In specific tasks as cell mitosis in breast-cancer, Convolutional Neural Networks based on ground truth data generated by a crowd of non-experts has reached outstanding performance on diseased cells mitosis detection [25]. In order to collect a large number of data and collaborators, a Web Application based workflow has been previously proposed: Its online availability would increase the amount of people contributing to the tasks [28].

2.5 Crowdsourcing, Gamification and Segmentation

For attracting a crowd of users to accomplish medical tasks, gamification has been proven to be a very effective strategy. This has been shown for medical students [29] as well as for crowds of non-experts [31]. This technique is a gateway to gain better motivation for users even for non-trivial tasks [30]. Regarding objects segmentation, a statistical analysis over the crowd contribution and the filtering out of inconsistent segmentations dramatically improves the result of a non-expert crowd. This leads to the status where the performance of non-experts moves very close to the performance of professionals [27]. Through semi-supervised objects segmentation, it is possible to keep track of the algorithm performances and adapt the algorithms to the proposed task. A precise quantification of time involved and difficulty level of the task helps furthermore to elaborate a better gaming interface and attract a larger crowd to the proposed online project [32].

3 Research Design

The research design has been structured in four steps, corresponding to the main points of the research. In order to conduct a structured research, we will borrow concepts and guidelines from the design science [33]. In particular, we will refer to the design science in the Information Systems Research. The artifacts that we are going to develop present in fact an innovative solution to an existing problem, and its utility is going to be evaluated on our specific domain, from a technical and business related point of view [33].

3.1 Step I: Design and Planning

After an accurate review on literature and the achievement of a structured theoretical background, the first step of the research is the design and planning of an evolutionary prototype for a Web Application [34], which is used to accomplish the goal expressed for the free innovation development: Create better automatic segmentation algorithms for medical organs images. As the evolutionary prototype pattern implies, the prototype will be robust and will constitute a reliable basis for a final version of the application [34].

In order to explain the functionalities of the software proposed, we will borrow the concept of use cases. Use cases are defined as a succession of events generated by actors, which point out dependencies and functional structure of the software [35]. The three main actors identified in the Medical Monkeys Application are the *Image Donor*, the *Solver – Gamer* and the *Researcher*.

The Image Donor provides own medical images for the Web Application. He donates his images for medical purposes and manages them, editing or even excluding them from the game.

The *Solver – Gamer* participates to the segmentation game on the medical images. His performance will be tested by the system and, in case of enough accuracy, his result will be submitted and the data collected and stored, for further statistical evaluation.

The *Researcher* has the role of analyzing and interacting with the data submitted by Image Donors and Solvers. In the first case, the researcher will be in charge of an evaluation over the uploaded images (so that they can be relevant for the research). In the second case, the researcher will be collecting and analyzing data results from the segmentation game.

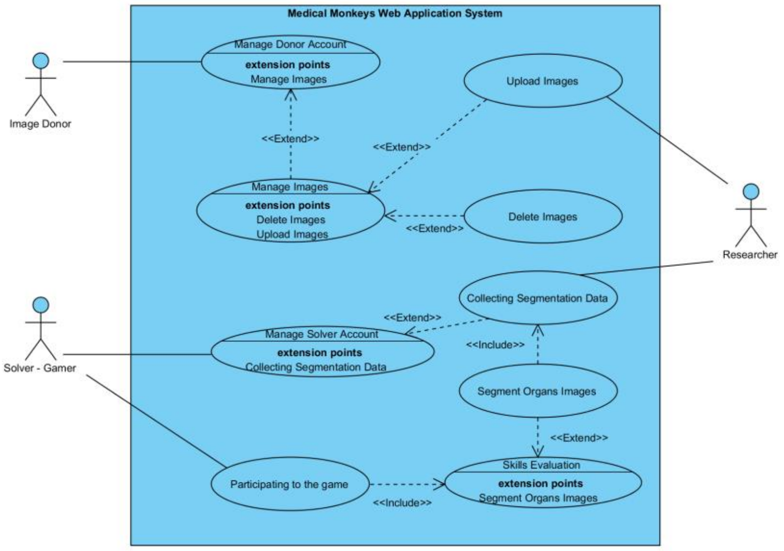

Fig 1. Use Case Scenario Diagram.

Given these actors, we can build up a scenario for the use cases. A scenario is defined as an ordered amount of interactions among partners, who are mostly represented by external actors and a given system [36]. In the use case scenario, the succession of events (e.g. the work steps and interactions) is pointed out. The diagram in Fig.1 is representing graphically our use case scenario.

The medical images received from the image donors are collected and categorized following the organ representing. These images will be inserted, as a 3D model

representation, in the Web Application, which will offer a game interface, where a crowd of users can participate to segment the donated organs images.
The gaming interface will be organized in levels and will motivate users with scoring and competition mechanisms. Through that, users will be triggered in keeping focused on the task and improve their performance. The levels will be structured by increasing difficulty, in order to stimulate solvers to get more accurate. The results of the game labelling task will be taken and analyzed by the researchers. A continuous improvement of the algorithms will be ensured by an increasing amount of gamers and donors. The larger the amount of images and the more the segmentation gamers, the more accurate will be the algorithms for organs images segmentation [27].

3.2 Step II: Web Application Implementation

As a second step, the tool will be implemented on a Web Platform. From this point, the Web Application will be available, and data of patients and game-solver will be collected. During the initial part of the data collection, the gamer will play against automatic segmentation algorithms, in order to obtain an own score. The Web Application will be gradually implemented and deployed for mobile devices as well, enhancing the chance of increasing participation by game-solvers. For each annotated layer the user gets points, which he can then post on social media. A public high-score list should then also allow the formation of groups, which then allows a competition between institutions. Currently, the Web prototype is to be found online, together the informative webpage. Image Donors and Gamers can register and manage their own profile, images and account[3]. The segmentation game is still in development. We have as well a public repository for our Web Application[4].

3.3 Step III: Data Collection and Storage

In the current step of our project, we request patients for data donations after medical investigations. Image data are particularly suitable for that, since these are often available as a DVD for the patients. Images of patients will be sent to us together with an agreement certificate. Once provided, we will store the images in a private cloud service. Before dispatching, the donor taps a TAN on the envelope, with which he then accesses his data online and can revoke the user authorization. Furthermore, the donor will get access to all research results obtained with his data. All the scientific work based on these data will be therefore published as open access.

For a first implementation of the storage, we realized a fault-tolerant and scalable distributed file system [37] based on Hadoop and managed by a local application. This is used mainly for uploading administering the medical images. Apache Hadoop enables Big Data to be stored, accessed and processed in a distributed way across clusters of commodity servers [38]. In order to provide an appropriate Hadoop based architecture for the storage and uploading of the medical images, we outlined first the requirements for our architecture, as shown in Fig. 2.

[3] http://medicalmonkeys.ddns.de

[4] https://github.com/Lorenzo1985/Monkey_BackBone.git

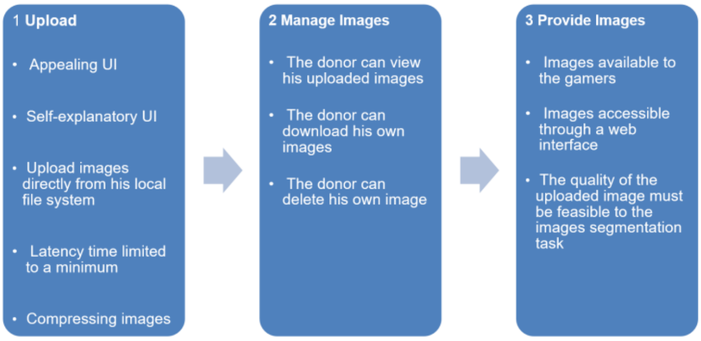
1 Upload
Appealing UI
Self-explanatory UI
Upload images directly from his local file system
Latency time limited to a minimum
Compressing images
2 Manage Images
The donor can view his uploaded images
The donor can download his own images
The donor can delete his own image
3 Provide Images
Images available to the gamers
Images accessible through a web interface
The quality of the uploaded image must be feasible to the images segmentation task

Fig 2. Distributed File System Requirements

The local application we prototyped for fulfilling these requirements has a lightweight frontend-solution which directly interacts with the distributed file system. Therefore, the system consists of three major components which implement all necessary interactions with the data storage mechanism. The diagram in Fig. 3 shows the components together with their utilized interfaces.

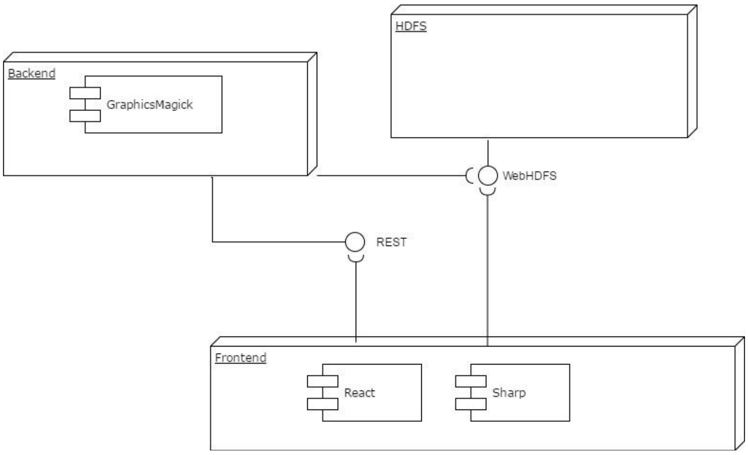

Fig 3. Local Application connected to the Distributed File System

Both, frontend and backend of the application interact with HDFS using WebHDFS [37]. To initialize the backend processes, the frontend uses a REST interface. The architecture is implemented with JavaScript, and we show with the following flowcharts, the necessary steps for storing and administering images in accordance to the given requirements (see Fig. 4). The algorithm implemented splits the images in tiles, which will be separately compressed in order to enhance the speed for reading the images and show them online for gaming and administration purposes. The introduction of Hadoop consent data as well as algorithm parallelism in our artifact. This means, faster writing as well as reading operations. For a deeper look into the implementation, it is possible to refer to this repository[5].

[5] https://github.com/chrissike/saveimages.git

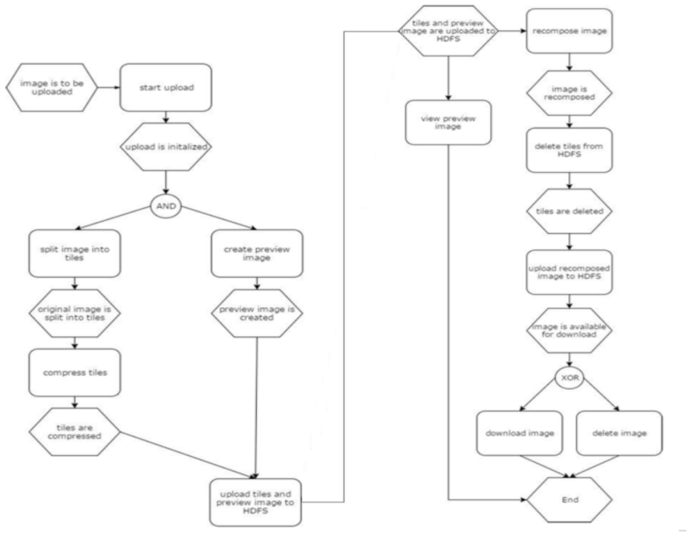

Fig 4. Local Application Flow Chart – Administering Images

3.4 Step IV: Data Analysis and Algorithm Evaluation

The last step will focus on the evaluation of the data obtained and the derived machine learning algorithms. Through statistical analysis, data coming from anomalous or not focused player will be ignored in the processing. This will lead to an improvement in the data collection and analysis [27]. In this step, a ground truth for the segmentation algorithm will be formed from the the valid values obtained by gamers. Given the ground truth created, ML algorithms will be measured and their accuracy will be evaluated against the state-of-the-art performances.

4 Conclusions and Future Work

ML algorithms for medical images segmentation are sensitive to the lack of large labelled training sets [39]. The reasons for that are mainly to be found in the privacy policies [39] and in the missing labelling which, differently from public internet images, cannot easily be performed by a non-expert crowd [40]. For solving this problem, we propose through Medical Monkeys a collaborative free innovation pattern that, using crowdsourcing to increase diffusion, enhances the amount of medical images and the segmentation accuracy of the algorithms, through a copious labeling. This paper presented the research design, requirements and prototype of our crowdsourcing Web Application, developed over a free innovation pattern, for medical images segmentation. We explored the possibility of free cooperation from Web users, researchers and images donors for improving the application, as a gateway for enhancing the performance of machine learning algorithms on medical images. This will lead to better automatic segmentation and detection algorithms, improving clinical decisions support systems and reducing the human-based error on anomalies evaluation [13]. The next steps of our research will lead to exploring effective way for designing our segmentation game and increasing the participation of the crowd. At the same time, we will continue with the data collection from hospitals and cooperation with institutions, in order to create a larger dataset of images.

References

1. Cai, L., Zhu, Y.: The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Sci. J. 14, 2 (2015).
2. Ward, J.S., Barker, A.: Undefined By Data: A Survey of Big Data Definitions. ArXiv13095821 Cs. (2013).
3. De Mauro, A., Greco, M., Grimaldi, M.: A formal definition of Big Data based on its essential features. Libr. Rev. 65, 122–135 (2016).
4. LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S., Kruschwitz, N.: Big data, analytics and the path from insights to value. MIT Sloan Manag. Rev. 52, 21–32 (2011).
5. Qiu, J., Wu, Q., Ding, G., Xu, Y., Feng, S.: A survey of machine learning for big data processing. EURASIP J. Adv. Signal Process. 2016, (2016).
6. Einav, L., Levin, J.: The Data Revolution and Economic Analysis. Innov. Policy Econ. 14, 1–24 (2014).
7. Schutt, C.O., Rachel: Doing Data Science.
8. Cios, K.J., William Moore, G.: Uniqueness of medical data mining. Artif. Intell. Med. 26, 1–24 (2002).
9. Aicardi, C., Del Savio, L., Dove, E.S., Lucivero, F., Tempini, N., Prainsack, B.: Emerging ethical issues regarding digital health data. On the World Medical Association Draft Declaration on Ethical Considerations Regarding Health Databases and Biobanks. Croat. Med. J. 57, 207–213 (2016).
10. Johnson-Lenz, P., Johnson-Lenz, T.: Post-mechanistic groupware primitives: rhythms, boundaries and containers. Int. J. Man-Mach. Stud. 34, 395–417 (1991).
11. West, J., Gallagher, S.: Challenges of open innovation: the paradox of firm investment in open-source software. R Manag. 36, 319–331 (2006).
12. Hippel, E. von: Free innovation. (2017).
13. Zhou, S.K., Greenspan, H., Shen, D.: Deep Learning for Medical Image Analysis. Academic Press (2017).
14. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. 2, (2014).
15. Steinbrook, R.: Personally controlled online health data--the next big thing in medical care? N. Engl. J. Med. 358, 1653–1656 (2008).
16. Dimitrov, D.V.: Medical Internet of Things and Big Data in Healthcare. Healthc. Inform. Res. 22, 156 (2016).
17. Aji, A., Wang, F., Saltz, J.H.: Towards building a high performance spatial query system for large scale medical imaging data. Presented at the (2012).
18. Van Horn, J.D., Toga, A.W.: Human neuroimaging as a “Big Data” science. Brain Imaging Behav. 8, 323–331 (2014).
19. de Jong, J.P.J., von Hippel, E., Gault, F., Kuusisto, J., Raasch, C.: Market failure in the diffusion of consumer-developed innovations: Patterns in Finland. Res. Policy. 44, 1856–1865 (2015).
20. Ogawa, S., Pongtanalert, K.: Exploring Characteristics and Motives of Consumer Innovators: Community Innovators vs. Independent Innovators. Res.-Technol. Manag. 56, 41–48 (2013).
21. Akgün, A.E., Keskin, H., Byrne, J.C.: Procedural Justice Climate in New Product Development Teams: Antecedents and Consequences. J. Prod. Innov. Manag. 27, 1096–1111 (2010).

22. Jeppesen, L.B., Lakhani, K.R.: Marginality and Problem-Solving Effectiveness in Broadcast Search. Organ. Sci. 21, 1016–1033 (2010).
23. Afuah, A., Tucci, C.L.: Crowdsourcing As a Solution to Distant Search. Acad. Manage. Rev. 37, 355–375 (2012).
24. The Rise of Crowdsourcing | WIRED, https://www.wired.com/2006/06/crowds/.
25. Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., Navab, N.: AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images. IEEE Trans. Med. Imaging. 35, 1313–1321 (2016).
26. Ranard, B.L., Ha, Y.P., Meisel, Z.F., Asch, D.A., Hill, S.S., Becker, L.B., Seymour, A.K., Merchant, R.M.: Crowdsourcing—Harnessing the Masses to Advance Health and Medicine, a Systematic Review. J. Gen. Intern. Med. 29, 187–203 (2014).
27. Maier-Hein, L., Mersmann, S., Kondermann, D., Bodenstedt, S., Sanchez, A., Stock, C., Kenngott, H.G., Eisenmann, M., Speidel, S.: Can Masses of Non-Experts Train Highly Accurate Image Classifiers? In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014. pp. 438–445. Springer International Publishing (2014).
28. Chávez-Aragón, A., Lee, W.-S., Vyas, A.: A crowdsourcing web platform-hip joint segmentation by non-expert contributors. In: Medical Measurements and Applications Proceedings (MeMeA), 2013 IEEE International Symposium on. pp. 350–354. IEEE (2013).
29. LEBA, M., IONICĂ, A., APOSTU, D.: Educational Software based on Gamification Techniques for Medical Students. Wseas Us. 225–230 (2013).
30. Spampinato, C., Palazzo, S., Giordano, D.: Gamifying Video Object Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. PP, 1–1 (2016).
31. Carlier, A., Salvador, A., Cabezas, F., Giro-i-Nieto, X., Charvillat, V., Marques, O.: Assessment of crowdsourcing and gamification loss in user-assisted object segmentation. Multimed. Tools Appl. 75, 15901–15928 (2016).
32. Salvador, A., Carlier, A., Giro-i-Nieto, X., Marques, O., Charvillat, V.: Crowdsourced object segmentation with a game. In: Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia. pp. 15–20. ACM (2013).
33. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Q. 28, 75–105 (2004).
34. Overmyer, S.: Revolutionary Vs. Evolutionary Rapid Prototyping: Balancing Software Productivity and Hci Design Concerns. Proc. Fourth Int. Conf. ….
35. Jacobson, I.: Object Oriented Software Engineering: A Use Case Driven Approach, http://www.citeulike.org/group/8357/article/348273.
36. Seybold, C., Meier, S., Glinz, M.: Scenario-driven modeling and validation of requirements models. Presented at the (2006).
37. An introduction to Apache Hadoop | Opensource.com, https://opensource.com/life/14/8/intro-apache-hadoop-big-data.
38. Ishwarappa, Anuradha, J.: A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. Procedia Comput. Sci. 48, 319–324 (2015).
39. Cho, J., Lee, K., Shin, E., Choy, G., Do, S.: How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? ArXiv151106348 Cs. (2015).
40. Startups, R. for: Deep Learning in Healthcare: Challenges and Opportunities, https://medium.com/the-mission/deep-learning-in-healthcare-challenges-and-opportunities-d2eee7e2545, (2016).