CrowdDetective: Wisdom of the Crowds for Detecting Abnormalities in Medical Scans

Veronika Cheplygina¹

¹ Medical Image Analysis, Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, 5612 AZ, the Netherlands

Correspondence: Veronika Cheplygina, Medical Image Analysis, Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, 5612 AZ, the Netherlands

Email: v.cheplygina@tue.nl

Grant Application to: Netherlands Organisation for Scientific Research Veni scheme 2018

Abstract

Machine learning (ML) has great potential for early diagnosis of disease from medical scans, and at times, has even been shown to outperform experts. However, ML algorithms need large amounts of annotated data – scans with outlined abnormalities – for good performance. The time-consuming annotation process limits the progress of ML in this field.

To address the annotation problem, multiple instance learning (MIL) algorithms were proposed, which learn from scans that have been diagnosed, but not annotated in detail. Unfortunately, these algorithms are not good enough at predicting where the abnormalities are located, which is important for diagnosis and prognosis of disease. This limits the application of these algorithms in research and in clinical practice.

I propose to use the “wisdom of the crowds” – internet users without specific expertise – to improve the predictions of the algorithms. While the crowd does not have experience with medical imaging, recent studies and pilot data I collected show they can still provide useful information about the images, for example by saying whether images are visually similar or not. Such information has not been leveraged before in medical imaging applications. I will validate these methods on three challenging detection tasks in chest computed tomography, histopathology images, and endoscopy video.

Understanding how the crowd can contribute to applications that typically require expert knowledge will allow harnessing the potential of large unannotated sets of data, training more reliable algorithms, and ultimately paving the way towards using ML algorithms in clinical practice.

KEYWORDS

machine learning, artificial intelligence, medical imaging, crowdsourcing, computer-aided diagnosis

Research proposal

General information

Institution of employment at time of application

Eindhoven University of Technology

Prospective host institution

Eindhoven University of Technology

Main field

Artificial Intelligence, Expert Systems

Other fields

Medicine, other

Number of words

Description of the proposed research: 1999 (max. 2,000 words)

Knowledge utilisation: 719 (max. 750 words)

Description of the proposed research

Overall aim and key objectives

Machine learning (ML) has seen tremendous successes in recent years, for example in classifying everyday objects such as cats in images. This progress

FIGURE 1 Supervised learning and MIL in medical imaging

is driven by the availability of powerful computational resources, sophisticated algorithms, but most of all, large amounts of labelled data. Data is annotated at different levels: outlining the cat in the image (a strong annotation) or tagging the whole image (a weak annotation). Strong labels are the most effective, but weak labels can also be used by weakly supervised or multiple instance learning (MIL) algorithms (Fig.1).

In medicine, ML offers invaluable opportunities for diagnosis of disease, reaching expert performance [1] or even outperforming the experts [2, 3]. However, while the amount of medical data is growing rapidly¹, the data is often only weakly annotated (a scan with a diagnosis, but without outlined abnormalities). Although MIL is gaining popularity in medical imaging [4–9], there is an overlooked, fundamental problem. MIL algorithms are optimized to predict weak annotations [10], but the classifier best at predicting weak annotations, is often not the best at predicting strong annotations [6, 8, 11, 12]. In practice this means that without strong annotations, MIL algorithms are poor at localizing abnormalities [13].

I propose a novel way of using crowdsourcing [14] to improve the abnormality localization. Crowdsourcing is the process of gathering input by enlisting the services of a large number of people, either paid or unpaid, typically via the internet, allowing large-scale data annotation [15]. A widely used example is tagging people on social media, used to improve face recognition. Crowdsourcing has therefore been successful in computer vision [15, 16] with non-expert tasks, such as recognizing everyday objects. However, internet users are not medical experts – how could they annotate abnormalities? A key insight is that the crowd does not need to mimic the experts to improve MIL for detecting abnormalities. For example, for detecting abnormalities in chest CT images, I propose to instead leverage the human visual system by asking:

- ▷ To outline airways, which are easily recognizable structures (Fig.2)
- ▷ Whether patches are similar to each other

These annotations can be intuitively provided by the crowd, as demonstrated by my pilot results [17, 18]. Although such annotations cannot be used to directly train algorithms, they are still informative.

The visual information in these annotations can help the MIL algorithm to

find better representations for the data via *multi-task learning* [19] with labels for related tasks, such as outlining airways, and with *similarity-based learning* [20, 21] with patch similarities.

The goal of this project is to improve the prediction of strong annotations by MIL, which is important for medical imaging, but also applications in other fields where annotations are scarce. Furthermore, the project will provide insight into the value of crowdsourcing in expert-level tasks. This is essential to leverage

FIGURE 2 Airway in a chest CT patch

the value and scientific potential of big datasets, which are growing at an exponential rate.

Originality and innovative aspects. The problem of predicting strong labels with MIL is underexplored [10, 12]. Using crowdsourcing to improve MIL is novel in general - in a recent survey of crowdsourcing in computer vision [16] none of the 195 papers address MIL, while a recent book on MIL [22] does not mention crowdsourcing. While techniques for multi-task and similarity-based learning exist, these often have assumptions not compatible with medical images annotated by crowds. I will develop methods which better address such data.

Crowdsourcing is new to medical imaging, with most studies published in the last five years [17, 23–27]. The crowd typically mimics the experts in labelling images with a diagnosis or outlining structures in image patches, and the results are compared to expert labels. Several studies show promising results, while in other cases the crowd does not achieve the level of the experts [25, 27]. Note that in most studies the crowd only labels the images – the labelled images are then not used in ML approaches. Only [27] uses the labels to train algorithms, but collects labels and does not address the MIL scenario. There is no prior work on collecting labels for related tasks and similarities, except in my recent pilot studies [17, 18].

Next to my expertise in these topics I started building a community around the topic of crowdsourcing for medical imaging (MICCAI LABELS workshop, recently funded eScience-Lorentz workshop), have many international collaborators and interest from industry. The project also benefits from my outreach and open science efforts (Section on Knowledge Utilisation).

Approach. I will investigate three strategies in which expert weak labels and crowd annotations can be combined to train more robust MIL algorithms:

- ▷ Experts + crowd labels for the expert task
- ▷ Experts + crowd labels for related tasks
- ▷ Experts + crowd similarities

My hypothesis is the best strategy will depend on the application, as ex-

1. http://ec.europa.eu/eurostat/statistics-explained/index.php/Healthcare_resource_statistics_-_technical_resources_and_medical_technology

FIGURE 3 Chest CT patches with healthy (left and middle) tissue and emphysema (right). Labelling these is difficult for non-experts, but they can still characterize the visual similarity of the patches.

plained below.

Labels from the crowd can be combined with expert labels to create a more robust classifier by *decision fusion (DF)* [28, 29]. In traditional DF approaches each classifier is trained on a (random) subset of the data [30–32] – classifiers of similar quality, and their decisions are averaged. An alternative is to weight classifiers by their estimated accuracy, which requires additional data or assumptions, such as that per image, at least half of the classifiers are correct. This assumption may not always hold in medical images. In more difficult tasks where expert knowledge is needed, I expect that this strategy will not outperform the baseline, where the MIL algorithm is trained only on the expert weak labels.

Labels for related tasks, such as outlining airways, when the target is to detect abnormalities, can be leveraged via *multi-task learning (MTL)* [19]. The intuition is that there are underlying features helpful for both tasks, which helps to find a more robust classifier. This is achieved by optimizing a joint loss function for the classifier. Such approaches are used in computer vision to simultaneously detect different everyday objects in images [33]. In medical imaging MTL was used to simultaneously predict an Alzheimer's diagnosis and a cognitive test score [34], and to simultaneously predict different abnormalities related to lung cancer [35].

However, the computer vision approaches have assumptions that might be incompatible with medical imaging, such as well-defined boundaries for each object, and the medical imaging methods do not address MIL and expect expert labels for both tasks. I will investigate the applicability of these and similar approaches, while addressing the specifics of crowd-annotated medical images, for example, by weighting the loss function such that the expert labels are given more emphasis. I expect that the MTL strategy will be most successful when the related task is simple enough for the crowd to do, but has some relevance to the target task.

Similarities, such as “image patch A is more similar to B than to C” (Fig. 3) can also add information during training. The idea is that such judgements are more intuitive for crowds to provide, even if they are not able to label the images. The similarities provide relative information about the expected classifier outputs, favouring classifiers for which these relationships hold. Examples include metric learning [20] and similarity embedding [21, 36], which I group under the name *similarity-based learning (SBL)*.

Again, the assumptions of current methods that focus on MIL may not be applicable. For example [37] assumes that an image label can apply to at most one patch in that image, which is not true for many abnormalities. Based on my

experience with different (medical and non-medical) MIL problems [38, 39], I will investigate how to create approaches with assumptions more suitable for medical images. I expect SBL to be successful when the crowd can focus on a single aspect of similarity, such as image texture. If there are multiple aspects to focus on, the similarities are likely to be inconsistent across annotators, and might hurt performance of MIL instead.

I will develop MIL algorithms which learn both from expert weak labels and crowd annotations, via DF, MTL or SBL. The strategies I will develop are general, therefore I will use them both with traditional [40, 41] and more recent [39, 42] MIL classifiers. Similarly, the strategies will apply both to traditional features and features extracted by deep learning.

To investigate which strategy is better in which scenario, I will apply these algorithms to three applications, with different image characteristics and task difficulty:

- ▷ Localizing patches with emphysema, in chest CT images [43, 44]. This is a task of medium difficulty, where I expect best results from MTL because the underlying disease affects the appearance of airways, or SBL based on recent pilot results [18].
- ▷ Localizing different surgical instruments in endoscopy videos [45]. This is a relatively easy task where DF is already likely to be successful.
- ▷ Localizing cancerous regions in histopathology images [46]. This is a difficult task, where I would expect best results for SBL, or MTL if a simple but powerful related task can be defined in collaboration with experts.

The datasets have expert weak labels, but also some expert strong labels available, which I will use for validation purposes. I will collect annotations via two platforms, with different trade-off between collection time and quality [47]:

- ▷ [Amazon Mechanical Turk](#) with paid crowds, which reduces the time to 1-Rajpurkar2017 days, but could reduce quality.
- ▷ [Zooniverse](#) with unpaid volunteers interested in science, which increases time, but also quality.

The next step is to decide which patches to annotate and how many annotators to assign to each patch. Combining annotators increases the quality but, given the same budget, decreases the total amount of annotated data. Because ML (and thus also MIL) algorithms benefit from seeing a larger set of data and my goal is not to produce expert-quality annotations, I expect it will be more valuable to maximize the amount of annotated patches. To keep the number of annotations within budget, I will investigate active selection of patches, for example based on their uncertainty or diversity according to the MIL algorithm.

Lastly I will compare the strategies to baselines, such as training MIL only

FIGURE 4 Learning curves, method 1 is better with few annotations and vice versa.

	Y1			Y2			Y3		
WP1									
WP2									
WP3									
WP4									
Knowledge utilization (next section)	UM	Visit DKFZ		UM	Visit ETS	Workshop	UM		Lectures

TABLE 1 Time Plan. UM = user meeting

with expert weak labels, and to each other. In this process I will vary the number of expert labels and crowd annotations used, in order to create more scenarios than the three applications investigated.

By analysing these results from three different applications, I will aim to extract general rules on when each strategy is best to use.

Research plan

The project has four work packages:

- WP1: Decision fusion with labels
- WP2: Multi-task learning with related labels
- WP3: Representation learning with similarities
- WP4: Generalization

In WP1-WP3, I will crowdsource the corresponding type of annotations and develop MIL algorithms for combining these annotations with expert weak labels. The deliverables are extended versions of my publicly available MIL toolbox [48] and papers published in high impact conferences and journals. In WP4 I will compare the strategies and draw up recommendations for their use. The deliverable is a paper with a related blog post.

I will conduct the research at the [Medical Image Analysis group](#), TU/e. The main collaborators (I will visit both for 3 months are

- ▷ [Prof. Lena Maier-Hein](#) (DKFZ, Germany). I will collaborate with her on crowdsourcing aspects, and apply my algorithms to endoscopy video data from her group.
- ▷ [Prof. Eric Granger](#) (ETS, Canada) and [Dr. Luke McCaffrey](#) (McGill, Canada). I will collaborate with Prof. Granger on MIL and apply my algorithms to histopathology data provided by Dr. McCaffrey.

I also have contacts in medical imaging and crowdsourcing communities whom I could approach for advice, including Dr. Marleen de Bruijne (Erasmus MC), Dr. Javed Khan (TU/e), Dr. Alessandro Bozzon (TU Delft).

Knowledge utilisation

Potential

Research. In medical imaging the results of my research will be beneficial to many research groups, as the annotation problem is present in many different applications. In machine learning (ML) the research will demonstrate the potential of improving algorithms with crowd annotations which are not

typically leveraged. This is an underexplored area, but is likely to spark interest in the community. I expect this to assist in development of (general purpose) algorithms focusing on such annotations.

Furthermore I expect interest from other applications where annotated data is scarce, such as remote sensing [49] and ecology [50]. Lastly, the behaviour of the crowd (who contributed what and why) could be of interest to research groups in human-computer interaction [47] and social sciences.

Industry. There is already great interest from the industry in ML, and medical imaging is catching up, for example in 2017 the leading conference [MICCAI](#) had IBM, Nvidia, Siemens and many medical imaging startups as sponsors.

The interest in my project is demonstrated by the companies in [Table 2](#), including IBM Research. These companies could integrate the research outcomes in their products, which are designed to be user-friendly and secure. This will enable translation of the proposed research to products and (in case of medical imaging) the clinic.

Society. The research area is relevant to several questions on the National Science Agenda, including health (questions [105](#), [89](#), [94](#)) and society and technology ([112](#), [108](#)). In health, there are long term benefits of improved medical imaging algorithms, such as better prognosis and diagnosis of disease, and better use of the experts' time.

For society, there are benefits in involving the public as annotators, which can raise awareness about health. In the long term, contributing to such projects could create jobs or volunteering opportunities, accessible to people who are be unable to work due to health or care responsibilities [51], increasing their well-being.

Implementation

Research. I will publish in relevant venues (MICCAI, Pattern Recognition) and make the papers/data/code available online. I will continue organizing events I'm already involved in (MICCAI LABELS, eScience-Lorentz workshop, NVPHBV) and giving talks about my work. In years 1 and 2 my research visits to Germany and Canada will help in reaching other groups I am not yet in contact with. I expect other groups will use my methods for their data from the first year of the project.

In year 2 I will organize a workshop on the intersection of ML and crowdsourcing in label-scarce applications beyond medical imaging. I will invite

researchers via my network and online communities I'm a member of ([ml-news](#), [crowd-hcomp](#)). The impact is more difficult to predict, but I expect that at least some groups will be interested in either the data and/or code generated by my project, and translating these to their own applications.

Industry. I have established a user group (Table 2) of industry representatives interested in the research outcomes. I will organize meetings with this group **every year**. Translation to industry could start within 2-Cheplygina2013 years of starting the project. This collaboration could lead to other joint projects, ensuring impact even after completion of the current research.

Group	Contact	Role
IBM Research	(anonymized)	IBM Research is working on a crowdsourcing solution for “internal crowds” (e.g. colleagues) and is interested in a medical imaging application.
Thirona, Clinical-Graphics	(anonymized)	Thirona and ClinicalGraphics develop software for medical imaging. They train employees for annotating images, and want to scale up the annotations without an equal increase in cost.
Cosmonio	(anonymized)	Cosmonio develops an app that allows interaction between the medical expert and the algorithm. My research will help to optimize the type of interactions needed.

TABLE 2 User Group

Society. I will reach a broader group of people (people with an interest in science and/or annotators from AmazonMT and ZooUniverse platforms) via outreach through my blog (3K visits per month) and Twitter (1.8K followers). I will blog about my project **every quarter**, explaining my project to an audience interested in science, but without a technical background.

In **year 3** I will record lectures for an online course on [DataCamp](#)², where I am already setting up an image analysis course. The course will combine machine learning, medical imaging and crowdsourcing, but will not require technical prerequisites. This is a more long-term strategy than the others.

Cost estimates

Total budget requested

€250.000

Intended starting date

September 1, 2018

Application for additional grants

No

	Personnel	Communication ³	Teaching	Equipment/ Material
Year 1	73.000	4.000		5.000
Year 2	75.400	5.000		3.000
Year 3	77.700	2.000	2.000	1.900
Total	226.100	11.000		9.000

TABLE 3 Cost Estimates

Data management plan

Will data be collected or generated that are suitable for reuse?

Yes. I will collect annotations (labels and similarities) for patches extracted from already available medical images. The annotations will be stored in the widely reusable JSON format.

I will also generate features – numerical data which describes the images, but from which the images cannot be reconstructed. The features will be stored as CSV files to facilitate reuse across different platforms.

The annotations and features (*hereafter referred to as relevant data*) can be reused in development of other machine learning algorithms.

Where will the data be stored during the research?

All data is stored electronically. The annotations will be collected via the cloud (Amazon Mechanical Turk or Zooniverse) and thus a secure back-up will always be available. The annotations will then be copied to secure network drives available at the host institute. This is also where the generated features will be stored. Daily back-ups are made on this storage facility.

Upon publication of a paper, I will also upload the relevant data to the [Figshare](#) repository (see below).

After the project has been completed, how will the data be stored for the long-term and made available for the use by third parties? To whom will the data be accessible?

I will store the relevant data on [Figshare](#) which will ensure its long-term preservation. Figshare has agreements with publishers such as Nature, PLOS, and others to ensure the data persists for a minimum of 10 years. I will share the relevant data under the Creative Commons Attribution-NonCommercial-ShareAlike (CC-BY-NC-SA) license.

Which facilities (ICT, (secure) archive, refrigerators or legal expertise) do you expect will be needed for the storage of data during the research and after the research? Are these available?*

**ICT facilities for data storage are considered to be resources such as data storage capacity, bandwidth for data transport and calculating power for data processing.*

I will use in-house computing facilities, which are already available at the host institute.

2. The author now distances herself from this company. She would not use their software after an investigation has shown an incident of sexual misconduct from the former CEO and the ensuing mismanagement of this incident. For transparency, the editors are maintaining the company name.

3. Conferences, outreach, publication, and travel

Ethics

Use of extension clause

No

Ethical aspects

Approval from a recognised (medical) ethics review committee	Not applicable
Approval from an animal experiments committee	Not applicable
Permission for research with the population screening Act	Not applicable

TABLE 4 Ethical aspects

Declarations. By submitting this form I endorse the code of conduct for laboratory animals and the code of conduct for biosecurity/possibility for dual use of the expected results and will act accordingly if applicable.

- ☒ I have completed this form truthfully.
- ☒ By submitting this document I declare that I satisfy the nationally and internationally accepted standards for scientific conduct as stated in the *Netherlands Code of Conduct for Scientific Practice 2014* (Association of Universities in the Netherlands)
- ☐ I have submitted non-referees.

Name: Veronika Cheplygina

Place: Eindhoven

Date: 8 January 2018

Society

Public summary

Crowds as medical detectives (ENG). *Dr.ir. V. (Veronika) Cheplygina (v), TU/e – Computer Science*

Detecting abnormalities in medical images is essential for diagnosis and treatment of illness. Computer algorithms can learn to do this using manually annotated scans, but the annotation process is costly for experts. This project studies how annotations made by untrained internet users can improve the detection accuracy of computer algorithms.

Internetgebruikers als medische experts? (NL). *Dr.ir. V. (Veronika) Cheplygina (v), TU/e – Informatica*

Het detecteren van afwijkingen in medische scans is belangrijk voor diagnose en behandeling van de ziekte. Computeralgoritmes kunnen dit leren van handmatig geannoteerde beelden, maar dit kost experts veel tijd. Dit project onderzoekt hoe annotaties van gewone internetgebruikers de automatische detectie van afwijkingen kunnen verbeteren.

Reviews

Grant

Vernieuwingsimpuls Veni ENW 2018

Title

CrowdDetective: wisdom of the crowds for detecting abnormalities in medical scans

Applicant

Dr. ir. V.V. Cheplygina

File number

016.Veni.192.066

Your rebuttal on the referee reports is in progress

Referee report of referee 1

Assessment of the quality of the researcher.

Explanation. Criteria - Quality of the researcher:

- ▷ in terms of profile fit in the target group;
- ▷ from an international perspective belongs to the 10 to 20% of his/her peer group;
- ▷ academic excellence as demonstrated by the PhD thesis, publications and/or other relevant achievements in the field
- ▷ inspiring enthusiasm for research and/or technology;
- ▷ persuasiveness;
- ▷ clear indications of an outstanding talent for academic research.

The Veni scheme aims at outstanding researchers only: the top 10-20% of his/her international peer group.

Question a: *What is your opinion on the past performance of the researcher (as demonstrated by his/her doctoral thesis, publications, and other relevant scientific achievements)?*

Comments: The applicant shows a clear focus in her research on multiple instance learning (MIL), one of the key components of the present application. She has several years of postdoctoral experience dealing with the application of MIL related algorithms to tasks in medical image analysis. I see some publications in journals, such as Pattern recognition and Pattern recognition letters, one Miccai paper. She has co-organized a workshop at the same conference, dealing with data annotation and crowd sourcing, the other key component of the present application. She also organized a workshop at ICML, and lists several international research visits. After only two years of national postdoc experience, she became assistant professor.

Overall, the applicant presents herself with a profile dedicated to an academic career, a driving research interest, all aligned well with the present research proposal. As a minor grain of salt I would have hoped for (more) high impact publications, i.e., either highly cited papers (and 8 years after the MSc there may be one or two highly cited papers, whatever the journal or conference), or contributions to high visibility conferences, such as ICML, NIPS, CVPR, or a second or third MICCAI paper.

Question b: *Does the applicant belong to the top 10-20% of his/her international peer group? Which scientific achievements or talents of the applicant show he/she belongs to this top?*

Comments: I see a number of activities, such as international collaborations and research visits, co-organization of events and workshops that actively shape the discussion of the research community. She actively disseminates research

ideas via new media with followers being interested in her views and opinions. I think this does set her apart from many of her peers. There are some noticeable publications, although with a 5 year PhD and 3 years of postdoc (and assistant professor) level research, a number of other researchers in the field might have a stronger publication record. Depending on how to weigh both aspects she may be among the top 20% of her peers.

Assessment of the quality, innovative character and academic impact of the proposal

Explanation. Criteria - Quality, innovative character and academic impact of the proposed research

- ▷ challenging in terms of content;
- ▷ originality of the research topic;
- ▷ innovative scientific elements;
- ▷ potential to make an important contribution to science;
- ▷ effective in terms of proposed methodology.

Question a: *Please comment on the relevance of the problem and on the originality and challenging content of the proposal.*

Comments: The problem of how to include crowd-sourced expert and non expert annotations is a relevant problem in machine learning and, hence, in medical image processing research. The problem is not solved yet and any solution would have the potential to impact significantly on the design and dissemination of machine learning in diagnostic clinical image analysis. Still, the proposed project could be stronger if it would not only focus on the comparison of different (more or less) existing techniques to a few selected (and more or less well defined) problems, but would promise to contribute to the advancement of related machine learning algorithms itself. I would see that generating these algorithms may be a natural second step, but - as it is - the application only promises to "investigate three strategies" that are likely to be data set dependent (as the applicant suggests). Overall, I feel the application promises to deliver solid and systematic research that, however, is far from offering new innovative concepts and contributions to the field.

Question b: *What are the innovative aspects of the proposal? Will the research break new ground by generating new concepts, a deeper understanding, new methods, etc.?*

Comments: The main contribution will be a systematic comparison of different analytic strategies on different data sets. As such, it promises some 'best practice' guidance in a field that would, indeed, would benefit from such systematic research.

Question c: *What is your opinion on its potential to make a major contribution to the advancement of scholarship, science or technology (academic impact)?*

Comments: see 2.b)

Question d: *To what extent is the proposed method effective? Please comment.*

Comments: The research strategy is well described, and the aims the applicant is presenting are likely to be reached.

Assessment of the knowledge utilisation

Explanation. Criteria - Knowledge utilisation (= KU)

NWO uses a broad definition of KU: not only innovative end of pipe product

creation is considered, but also purposeful interaction with (potential) knowledge users, such as industry or public organisations and knowledge transfer to other scientific disciplines. NWO asks applicants to fill out the paragraph on KU. An applicant may however explain that KU cannot be expected given the nature of the research project. In that case, we still kindly ask you to assess whether the applicant has provided reasonable argument or evidence sufficiently.

Potential:

- ▷ contribution to society and/or other academic areas;
- ▷ disciplines and organisations that might benefit from the results.

Implementation:

- ▷ action plan to allow the outcomes of the research project to benefit the potential knowledge users;
- ▷ if and how the potential knowledge users will be involved;
- ▷ (concrete) outcomes for society;
- ▷ the period over which KU is expected to occur.

Question a: *What is your opinion on the described relevance of the results of the research?*

Comments: In the best case the project will help paving the road for simplifying research in medical image computing and the translation of medical image computing technology into clinical practice. The outcome of the project, i.e., a description of the optimal strategy for structuring diagnostic information linked to a given medical image set, can be used both in the design, or evaluation, or continuous quality control of these technologies. As such, I would consider the overall research direction to be interesting and relevant also from a wider perspective.

Question b: *Please comment on the effectiveness and feasibility of the proposed approach for knowledge utilisation.*

Comments: Data sets and best practice recommendations, together with related algorithms, will be the promised output: "By analysing these results from three different applications, I will aim to extract general rules on when each strategy is best to use". Whether these general rules exist, will only be known upon completion of the projection. I am somewhat missing a 'basic methodological research' component in the research objectives, for example, exploring one particular machine learning algorithm for MIL on top of the promised systematic comparison (whether this is the ubiquitous deep learning or any other). Similarly, a 'driving clinical problem' that would be solved at the end of the project, would have been nice as well (e.g., new solutions to one interesting problem that would be relevant - whether these rules generalize to other tasks, or not).

Question c: *Only answer this question in case the applicant argued that knowledge utilisation is not to be expected given the nature of the research proposal: Does the applicant convincingly explain why knowledge utilisation is not applicable for his/her research project (see also the information under criterion 3 listed above)?*

Comments: Knowledge Utilization is expected.

Final assessment

Question a: *How do you assess the entire application? Please give your final scoring (A+/A/B/UF/U).*

Comments: 03, B

Question b: *Could you please summarize (point by point) the strengths and weaknesses of the grant application focussing on the candidate, proposal and knowledge utilisation?*

Comments: The applicant presents herself with a dedicated and good career path in medical image computing. There is a focus on MIL, that is relevant for the present project, although I feel that publications are cited for systematic comparisons, benchmarks, and implementations than for innovative methodological contributions. The proposed project is very timely and, in case "general rules" can be found, has the potential to have significant impact. It promises these contributions from a systematic study, rather than from new innovative concepts and ideas.

Feedback datamanagement

Question. Feedback datamanagement

Comments: no

Referee report of referee 2

Assessment of the quality of the researcher.

Explanation. Criteria - Quality of the researcher:

- ▷ in terms of profile fit in the target group;
- ▷ from an international perspective belongs to the 10 to 20% of his/her peer group;
- ▷ academic excellence as demonstrated by the PhD thesis, publications and/or other relevant achievements in the field
- ▷ inspiring enthusiasm for research and/or technology;
- ▷ persuasiveness;
- ▷ clear indications of an outstanding talent for academic research.

Question a: *What is your opinion on the past performance of the researcher (as demonstrated by his/her doctoral thesis, publications, and other relevant scientific achievements)?*

Comments: The researcher has a very good profile, with organising of workshops and interesting publications in good venues. It is a junior researcher so major impact in terms of citations can not be expected. The research direction is in line with her previous work and she has experience with research visits as well.

Question b: *Does the applicant belong to the top 10-20% of his/her international peer group? Which scientific achievements or talents of the applicant show he/she belongs to this top?*

Comments: She has received recognition by being part of the workshop organisers at LABELs and has had several interesting publications.

Assessment of the quality, innovative character and academic impact of the proposal

Explanation. Criteria - Quality, innovative character and academic impact of the proposed research

- ▷ challenging in terms of content;
- ▷ originality of the research topic;
- ▷ innovative scientific elements;
- ▷ potential to make an important contribution to science;

▷ effective in terms of proposed methodology.

Question a: *Please comment on the relevance of the problem and on the originality and challenging content of the proposal.*

Comments: The problem of limited annotation is an important problem in medical image analysis and a major limiting factor of machine learning in image analysis. Another important factor is the amount of data available, which increasingly gets easier with large repositories such as TCIA and TCGA that were not even mentioned in the text.

Question b: *What are the innovative aspects of the proposal? Will the research break new ground by generating new concepts, a deeper understanding, new methods, etc.?*

Comments: Using Crowdsourcing has been done many times in medical image analysis and a few examples are mentioned. The link of weak annotations is somewhat novel but it seems that there are other approaches that would need to be combined. Just getting more data is often not enough and getting the right images annotated, so those that are on the decision boundaries would seem most important. I did not see any strategy of quality control of the crowdsourced annotations and this seems like the major factor that is important.

Question c: *What is your opinion on its potential to make a major contribution to the advancement of scholarship, science or technology (academic impact)?*

Comments: There is an opportunity to advance the area medical image annotation but to a limited degree with the approaches give if no quality control is done and if only weak labels are given. MIL is important and finding a link between the annotation and the best approaches could be very interesting.

Pretty much all medical images have reports associated to them, so ignoring the available weak labels would be a pity.

These can be radiology and pathology reports and may be more effective than getting labels of limited quality.

Question d: *To what extent is the proposed method effective? Please comment.*

Comments: It is very hard to judge if the method will work. Some approaches have been using crowdsourcing in the past and they show that with strong quality control this works well. It is not clear how this will be leveraged by the proposed approach.

Assessment of the knowledge utilisation

Explanation. Criteria - Knowledge utilisation (= KU)

NWO uses a broad definition of KU: not only innovative end of pipe product creation is considered, but also purposeful interaction with (potential) knowledge users, such as industry or public organisations and knowledge transfer to other scientific disciplines. NWO asks applicants to fill out the paragraph on KU. An applicant may however explain that KU cannot be expected given the nature of the research project. In that case, we still kindly ask you to assess whether the applicant has provided reasonable argument or evidence sufficiently.

Potential:

- ▷ contribution to society and/or other academic areas;
- ▷ disciplines and organisations that might benefit from the results.

Implementation:

- ▷ action plan to allow the outcomes of the research project to benefit the potential knowledge users;
- ▷ if and how the potential knowledge users will be involved;
- ▷ (concrete) outcomes for society;
- ▷ the period over which KU is expected to occur.

Question a: *What is your opinion on the described relevance of the results of the research?*

Comments: The results have a potential to increase clinical decision making if it is working well. Still, all relies on the techniques to work and the annotation to be of good quality and there are currently no methods for quality control, so this is somewhat limited.

Question b: *Please comment on the effectiveness and feasibility of the proposed approach for knowledge utilisation.*

Comments: The research proposes contact with industry and has an industrial panel. It is not clear how exactly the interaction will be done and how intellectual property rights can be shared. AI in medicine is a hot topic in industry as well, so there definitely is potential if things work well.

Question c: *Only answer this question in case the applicant argued that knowledge utilisation is not to be expected given the nature of the research proposal: Does the applicant convincingly explain why knowledge utilisation is not applicable for his/her research project (see also the information under criterion 3 listed above)?*

Comments: knowledge utilisation is expected.

Final assessment

Question a: *How do you assess the entire application? Please give your final scoring (A+/A/B/UF/U).*

Comments referee. 03, B

Question b: *Could you please summarize (point by point) the strengths and weaknesses of the grant application focussing on the candidate, proposal and knowledge utilisation?*

Comments:

Strong points:

- ▷ important domain of medical data annotation to train machine learning classifiers
- ▷ crowdsourcing has shown strong potential with good quality control
- ▷ good links with the MICCAI community via the LABELS workshop
- ▷ good to adapt machine learning to specific types of annotations

Weak points:

- ▷ quality control is not mentioned and this seems essential
- ▷ active learning, so selecting the best images to annotate with maximised information gain is not mentioned
- ▷ there are many more papers using crowdsourcing for medical imaging than mentioned and the background should really be checked
- ▷ existing large data repositories are not mentioned such as TCGA and TCIA
- ▷ why are expert labels not used to control crowd labels? Why would combinations be useful? If expert labels exist than crowd labels do not seem necessary
- ▷ the similarity between images or patches extremely subjective and texture

is not a concept where people will have consistent answers; there is much literature on subjective perception and non of it is mentioned

- ▷ it is not clear which images are annotated and how many images are needed; who will provide images? who will test the system?

Feedback datamanagement

Question. Feedback datamanagement

Comments: The data management only concentrates on the annotations to be collected and not on the raw data. Where will the images originate from? How many are available? Who will test the algorithm and who will generate the ground truth? Is ethics approval for the CT and histopathology images available? By whom? How will it be insured that these data are treated properly? These are human data!

Sharing only annotations without the raw data would have a very limited usefulness.

Rebuttal

File number

016.Veni.192.066

Name of candidate

Veronika Cheplygina

Title

CrowdDetective: wisdom of the crowds for detecting abnormalities in medical scans

I would like to start by thanking the committee and the reviewers for taking the time to provide feedback on my application. Below I first address the overall opinion of the reviewers and then discuss a few specific points. Direct quotes from the reviewers are in *red*, and direct quotes from my proposal are in *blue*. The page numbers refer to pages in my submitted PDF, where my research proposal is on pages 4 through 9.

The reviewers are positive about my profile as researcher, mentioning important publications and my leadership role in the community around my research topic. R1 comments that my publication record could have been stronger. I would disagree with this, since several of my publications have been cited at 5 or 10 times the impact factor of the venue. Furthermore, since submitting the Veni, my citations increased from 230 to 254, and my h-index from 9 to 10. Two papers I published in 2017 already have 6 and 4 citations, therefore I would expect the h-index to further increase 12 in 2018. This is exceptional for somebody at my career stage, after 6.5 years in total (not 8, as the reviewer calculated) spent on research.

The reviewers agree that the problem is important, the method has potential and describe the project as well-defined. R1 is concerned with a more innovative contribution of the project, and suggests it would have been better to focus on one specific MIL algorithm, AND on one clinical problem. First, a key innovation of the proposal is to focus on different types of annotations that have been collected from the crowd, which has not been addressed before in medical imaging. Furthermore, I have specifically chosen to focus on a range of methods and applications, providing general guidelines for the field. I would

argue that this is more innovative than developing specific methods for specific applications, which is what is regularly being done at most conferences on the topic.

In contrast to R1, R2 seems to find the proposal too innovative, and suggests it would be better to follow the existing approach – collecting labels from the crowd, and comparing them to expert labels. As I discuss in the proposal, this is likely not to be an optimal strategy. My proposed methods, which focus on alternative (not yet investigated) types of annotations, are more promising in this regard. Since they rely on more intuitive characteristics of the images, the quality control is also less of an issue than suggested by the reviewer. Of course, I will still perform validation, as described on page 7 of the proposal – *“The datasets have expert weak labels, but also some expert strong labels available, which I will use for validation purposes”*, R2 suggests a number of other improvements, most of which are either addressed in my proposal, or could not be addressed within the scope of the project. I briefly respond to these below:

Active learning. This is indeed an important point to investigate. I mention this in my proposal on page 7 *“will investigate active selection of patches, for example based on their uncertainty or diversity according to the MIL algorithm”*. Although I did not mention this explicitly, both of my collaborators have recent work on active learning [Carbonneau2017] and the related concept of uncertainty estimation [Moccia2018]. These methodologies can be incorporated in the algorithms I develop.

Existing large repositories such as TCIA and TCGA. TCGA is a repository of genomic data, which is not relevant to my proposal. TCIA could be an interesting resource, but does not provide local annotations, which is precisely what is necessary for validation/quality control. As I describe in the proposal, I choose to focus on three applications for which local annotations are available for validation.

Patient reports. Patient reports provide weak labels for images, and are indeed often the basis of the expert weak labels I have available for my datasets. It is incorrect that I ignore these labels – these are in fact the expert labels my methods will use, in combination with the crowd labels. Processing the patient reports with natural language processing is outside the scope of my research.

Redundancy of expert and crowd labels. The reviewer writes *“if expert labels exist than crowd labels do not seem necessary”*. This is incorrect. The use of expert weak labels alone leads to unstable MIL algorithms, as I have detailed on page 4, *“In practice this means that without strong annotations, MIL algorithms are poor at localizing abnormalities [13].”*

Questions on data management. These questions, together with the other comments, suggest that the reviewer has overlooked an entire page of my proposal (page 7), where I discuss the public datasets I will use, and which already have expert labels available for validation. Overall, given the many positive comments of the reviewers, and the fact that several weak points are not justified, I hope that the committee will consider my proposal for the interview stage.

References

Carbonneau, M. A., Granger, E., & Gagnon, G. (2017). Bag-Level Aggregation for Multiple Instance Active Learning in Instance Classification Problems.

arXiv preprint arXiv:1710.02584.

Moccia, S., Wirkert, S. J., Kenngott, H., Vemuri, A. S., Apitz, M., Mayer, B., ... & Maier-Hein, L. (2018). Uncertainty-aware organ classification for surgical data science applications in laparoscopy. *IEEE Transactions on Biomedical Engineering*.

Decision

Project number

016.Veni.192.066

Applicant

V. V. Cheplygina

Title

CrowdDetective: wisdom of the crowds for detecting abnormalities in medical scans

Scores

Quality of the applicant: 4.2

Quality of the research proposal: 5.8

Knowledge utilization: 3.9

Final score: 4.8

Qualification: Your research proposal received the qualification “good”, based on the application, the reviewer reports and the rebuttal.

Quality of the Candidate

The committee and reviewers agree that the candidate has a clear research focus and an average to good publication record, although high-impact publications are still missing. One reviewer is therefore hesitant to place the candidate in the top 20% of her international cohort, which is agreed upon by the committee. The candidate was however found by the committee to be an ambitious researcher who has spent a significant amount of time on academic services (workshop and conference organisation, reviewing duties, board member) and outreach.

Quality of the Proposal

The reviewers agree that the proposal tackles a very timely and relevant research topic within medical image processing. They also notice that the methodology is logical, though not overly compelling. One reviewer questions the novelty of the envisioned contributions while another reviewer raises the issue of crowdsourcing quality control that should have been included in the proposal. The committee shares the doubt of the reviewers on these aspects.

Knowledge Utilization

The committee and the reviewers find the knowledge utilization plan convincing. The plan aligns well with the candidate’s prior experiences and targets different audiences with diverse activities. The inclusion of an industry panel is valued by the committee. One reviewer misses more details on issues of intellectual property rights as well as further details on the setup of the industry panel, to which the committee agrees that more details should be provided.

References

- [1] Kooi, T., Litjen, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., & Karssemeijer, N. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Medical image analysis*, 35, 303–312. <https://doi.org/10.1016/j.media.2016.07.007>
- [2] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., & Shpanskaya, K. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning (arXiv preprint arXiv: 1711.05225).
- [3] Bejnordi, B. E., Veta, M., van Diest, P. J., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A., Hermesen, M., Manson, Q. F., & Balkenhol, M. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22), 2199–2210. <https://doi.org/10.1001/jama.2017.14585>
- [4] Manivannan, S., Cobb, C., Burgess, S., & Trucco, E. (2016). Sub-category Classifiers for Multiple-instance Learning and Its Application to Retinal Nerve Fiber Layer Visibility Classification. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, & W. Wells (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (pp. 308–316). Springer International Publishing. https://doi.org/10.1007/978-3-319-46723-8_36
- [5] Cheplygina, V., Sorensen, L., Tax, D. M. J., Pedersen, J. H., Loog, M., & de Bruijne, M. (2014). Classification of COPD with Multiple Instance Learning. *2014 22nd International Conference on Pattern Recognition*, 1508–1513. <https://doi.org/10.1109/icpr.2014.268>
- [6] Kandemir, M., & Hamprecht, F. A. (2015). Computer-aided diagnosis from weak supervision: A benchmarking study. *Computerized Medical Imaging and Graphics*, 42, 44–50. <https://doi.org/10.1016/j.compmedimag.2014.11.010>
- [7] Melendez, J., van Ginneken, B., Maduskar, P., Philipsen, R. H. H. M., Reither, K., Breuninger, M., Adetifa, I. M. O., Maane, R., Ayles, H., & Sanchez, C. I. (2014). A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays. *IEEE Transactions on Medical Imaging*, 34(1), 179–192. <https://doi.org/10.1109/tmi.2014.2350539>
- [8] Cheplygina, V., Sørensen, L., Tax, D. M. J., de Bruijne, M., & Loog, M. (2015). Label Stability in Multiple Instance Learning. In N. Navab, J. Hornegger, W. M. Wells, & A. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 539–546). Springer International Publishing. https://doi.org/10.1007/978-3-319-24553-9_66
- [9] Quellec, G., Lamard, M., Abràmoff, M. D., Decencièrre, E., Lay, B., Erginay, A., Cochener, B., & Cazuguel, G. (2012). A multiple-instance learning framework for diabetic retinopathy screening. *Medical Image Analysis*, 16(6), 1228–1240. <https://doi.org/10.1016/j.media.2012.06.003>
- [10] Quellec, G., Cazuguel, G., Cochener, B., & Lamard, M. (2017). Multiple-instance learning for medical image and video analysis. *IEEE reviews in biomedical engineering*. <https://doi.org/10.1109/rbme.2017.2651164>
- [11] Vanwinckelen, G., Fierens, D., & Blockeel, H. (2016). Instance-level accuracy versus bag-level accuracy in multi-instance learning. *Data Mining and Knowledge Discovery*, 30(2), 313–341. <https://doi.org/10.1007/s10618-015-0416-z>
- [12] Carbonneau, M.-A., Granger, E., Raymond, A. J., & Gagnon, G. (2016). Robust multiple-instance learning ensembles using random subspace instance selection. *Pattern Recognition*, 58, 83–99. <https://doi.org/10.1016/j.patcog.2016.03.035>
- [13] Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.-J., & Fei-Fei, L. (2018). Thoracic Disease Identification and Localization with Limited Supervision. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8290–8299. <https://doi.org/10.1109/CVPR.2018.00865>
- [14] Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, 14(6), 1–4.
- [15] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *European conference on computer vision (ECCV)* (pp. 740–755). Springer.
- [16] Kovashka, A., Russakovsky, O., Fei-Fei, L., & Grauman, K. (2016). Crowdsourcing in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 10(3), 177–243. <https://doi.org/10.1561/06000000071>
- [17] Cheplygina, V., Perez-Rovira, A., Kuo, W., Tiddens, H. A. W. M., & de Bruijne, M. (2016). Early Experiences with Crowdsourcing Airway Annotations in Chest CT. In G. Carneiro, D. Mateus, L. Peter, A. Bradley, J. M. R. S. Tavares, V. Belagiannis, J. P. Papa, J. C. Nascimento, M. Loog, Z. Lu, J. S. Cardoso, & J. Cornebise (Eds.), *Deep Learning and Data Labeling for Medical Applications* (pp. 209–218). Springer International Publishing. https://doi.org/10.1007/978-3-319-46976-8_22
- [18] Ørting, S. N., Cheplygina, V., Petersen, J., Thomsen, L. H., Wille, M. M. W., & de Bruijne, M. (2017). Crowdsourced Emphysema Assessment. In M. J. Cardoso, T. Arbel, S.-L. Lee, V. Cheplygina, S. Balocco, D. Mateus, G. Zahnd, L. Maier-Hein, S. Demirci, E. Granger, L. Duong, M.-A. Carbonneau, S. Albarqouni, & G. Carneiro (Eds.), *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis* (pp. 126–135). Springer International Publishing. https://doi.org/10.1007/978-3-319-67534-3_14
- [19] Caruana, R. (1998). Multitask Learning. In S. Thrun & L. Pratt (Eds.), *Learning to Learn* (pp. 95–133). Springer US. https://doi.org/10.1007/978-1-4615-5529-2_5
- [20] Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10, 207–244.
- [21] van der Maaten, L., & Weinberger, K. (2012). Stochastic triplet embedding. *2012 IEEE International Workshop on Machine Learning for Signal Processing*, 1–6. <https://doi.org/10.1109/MLSP.2012.6349720>

- [22] Herrera, F., Ventura, S., Bello, R., Cornelis, C., Zafra, A., Sánchez-Tarragó, D., & Vluymans, S. (2016). *Multiple Instance Learning*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-47759-6>
- [23] Maier-Hein, L., Mersmann, S., Kondermann, D., Stock, C., Kenngott, H. G., Sanchez, A., Wagner, M., Preukschas, A., Wekerle, A.-L., Helfert, S., Bodenstedt, S., & Speidel, S. (2014). Crowdsourcing for Reference Correspondence Generation in Endoscopic Images. In P. Golland, N. Hata, C. Barillot, J. Hornegger, & R. Howe (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014* (pp. 349–356). Springer International Publishing. https://doi.org/10.1007/978-3-319-10470-6_44
- [24] Maier-Hein, L., Mersmann, S., Kondermann, D., Bodenstedt, S., Sanchez, A., Stock, C., Kenngott, H. G., Eisenmann, M., & Speidel, S. (2014). Can Masses of Non-Experts Train Highly Accurate Image Classifiers?: A Crowdsourcing Approach to Instrument Segmentation in Laparoscopic Images. In P. Golland, N. Hata, C. Barillot, J. Hornegger, & R. Howe (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014* (pp. 438–445). Springer International Publishing. https://doi.org/10.1007/978-3-319-10470-6_55
- [25] Mitry, D., Peto, T., Hayat, S., Blows, P., Morgan, J., Khaw, K.-T., & Foster, P. J. (2015). Crowdsourcing as a Screening Tool to Detect Clinical Features of Glaucomatous Optic Neuropathy from Digital Photography (W. H. Merigan, Ed.). *PLOS ONE*, 10(2), e0117401. <https://doi.org/10.1371/journal.pone.0117401>
- [26] Nguyen, T. B., Wang, S., Anugu, V., Rose, N., McKenna, M., Petrick, N., Burns, J. E., & Summers, R. M. (2012). Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. *Radiology*, 262(3), 824–833. <https://doi.org/10.1148/radiol.11110938>
- [27] Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., & Navab, N. (2016). Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging*, 35(5), 1313–1321. <https://doi.org/10.1109/tmi.2016.2528120>
- [28] Kittler, J. (1998). Combining classifiers: A theoretical framework. *Pattern Analysis & Applications*, 1(1), 18–27. <https://doi.org/10.1007/bf01238023>
- [29] Kuncheva, L. I. (2004). *Combining pattern classifiers: Methods and algorithms*. John Wiley & Sons.
- [30] Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24(2), 123–140. <https://doi.org/10.1023/A:1018054314350>
- [31] Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8), 832–844. <https://doi.org/10.1109/34.709601>
- [32] Dietterich, T. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2), 139–157. <https://doi.org/10.1023/a:1007607513941>
- [33] Vezhnevets, A., & Buhmann, J. M. (2010). Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3249–3256. <https://doi.org/10.1109/CVPR.2010.5540060>
- [34] Cheng, B., Liu, M., Suk, H.-I., & Shen, D. (2015). Multimodal manifold-regularized transfer learning for MCI conversion prediction. *Brain imaging and behavior*, 9(4), 1–14. <https://doi.org/10.1007/s11682-015-9356-x>
- [35] Bi, J., Xiong, T., Yu, S., Dundar, M., & Rao, R. B. (2008). An Improved Multi-task Learning Approach with Applications in Medical Diagnosis. In W. Daelemans, B. Goethals, & K. Morik (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 117–132). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-87479-9_26
- [36] Hoffer, E., & Ailon, N. (2015). Deep Metric Learning Using Triplet Network. In A. Feragen, M. Pelillo, & M. Loog (Eds.), *Similarity-Based Pattern Recognition* (pp. 84–92). Springer International Publishing. https://doi.org/10.1007/978-3-319-24261-3_7
- [37] Law, M. T., Yu, Y., Urtasun, R., Zemel, R. S., & Xing, E. P. (2017, July). Efficient multiple instance metric learning using weakly supervised data. *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE.
- [38] Cheplygina, V., Tax, D. M. J., & Loog, M. (2015). Multiple instance learning with bag dissimilarities. *Pattern Recognition*, 48(1), 264–275. <https://doi.org/10.1016/j.patcog.2014.07.022>
- [39] Cheplygina, V., Tax, D. M. J., & Loog, M. (2016). Dissimilarity-based ensembles for multiple instance learning". *IEEE Transactions on Neural Networks and Learning Systems*, 27(6), 1379–1391. <https://doi.org/10.1109/tnnls.2015.2424254>
- [40] Chen, Y., Bi, J., & Wang, J. (2006). MILES: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12), 1931–1947. <https://doi.org/10.1109/tpami.2006.248>
- [41] Andrews, S., Tsochantaridis, I., & Hofmann, T. (2002). Support vector machines for multiple-instance learning. *Advances in neural information processing systems (NIPS)* (pp. 561–568).
- [42] Carbonneau, M.-A., Cheplygina, V., Granger, E., & Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77, 329–353. <https://doi.org/10.1016/j.patcog.2017.10.009>
- [43] Depeursinge, A., Vargas, A., Platon, A., Geissbuhler, A., Poletti, P.-A., & Müller, H. (2012). Building a reference multimedia database for interstitial lung diseases. *Computerized medical imaging and graphics*, 36(3), 227–238. <https://doi.org/10.1016/j.compmedimag.2011.07.003>
- [44] Pedersen, J. H., Ashraf, H., Dirksen, A., Bach, K., Hansen, H., Toennesen, P., Thorsen, H., Brodersen, J., Skov, B. G., & Døssing, M. (2009). The Danish randomized lung cancer CT screening trial-overall design and results of the prevalence round. *Journal of Thoracic Oncology*, 4(5), 608–614. <https://doi.org/10.1097/jto.0b013e3181a0d98f>
- [45] Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., Kenngott, H., Speidel, S., Kopp-Schneider, A., Maier-Hein, K., & Maier-Hein, L.

- (2018). Exploiting the potential of unlabeled endoscopic video data with self-supervised learning (arXiv preprint arXiv: 1711.09726). *International Journal of Computer Assisted Radiology and Surgery*, 13(6), 925–933. <https://doi.org/10.1007/s11548-018-1772-0>
- [46] Veta, M., Diest, P. J. V., Willems, S. M., Wang, H., Madabhushi, A., Cruz-Roa, A., Gonzalez, F., Larsen, A. B., Vestergaard, J. S., & Dahl, A. B. (2015). Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical image analysis*, 20(1), 237–248. <https://doi.org/10.1016/j.media.2014.11.010>
- [47] Mao, A., Kamar, E., Chen, Y., Horvitz, E., Schwamb, M. E., Lintott, C. J., & Smith, A. M. (2013). Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. *First AAAI conference on human computation and crowdsourcing*.
- [48] Tax, D. M. J., & Cheplygina, V. (2016). *MIL, a Matlab toolbox for multiple instance learning*. prlab.tudelft.nl/
- [49] Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F., & Obersteiner, M. (2009). Geo-wiki. org: The use of crowdsourcing to improve global land cover. *Remote Sensing*, 1(3), 345–354. <https://doi.org/10.3390/rs1030345>
- [50] Fink, D., Damoulas, T., Bruns, N. E., Sorte, F. A. L., Hochachka, W. M., Gomes, C. P., & Kelling, S. (2014). Crowdsourcing meets ecology: Hemisphere-wide spatiotemporal species distribution models. *AI magazine*, 35(2), 19–30. <https://doi.org/10.1609/aimag.v35i2.2533>
- [51] Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018). A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–14. <https://doi.org/10.1145/3173574.3174023>

License

Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020