SMI Swiss Medical Informatics
SGMI SSIM SSMI
Schweizerische Gesellschaft für Medizinische Informatik
Société Suisse d'Informatique Médicale
Società Svizzera d'Informatica Medica
Swiss Society for Medical Informatics

Crowdsourcing for medical image classification

Alba García Seco de Herrera, Antonio Foncubierta-Rodríguez, Dimitrios Markonis, Roger Schaer, Henning Müller

University of Applied Sciences Western Switzerland, Sierre

Abstract

To help manage the large amount of biomedical images produced, image information retrieval tools have been developed to help access the right information at the right moment. To provide a test bed for image retrieval evaluation, the ImageCLEFmed benchmark proposes a biomedical classification task that automatically focuses on determining the image modality of figures from biomedical journal articles.

In the training data for this machine learning task, some classes have many more images than others and thus a few classes are not well represented, which is a challenge for automatic image classification. To address this problem, an automatic training set expansion was first proposed. To improve the accuracy of the automatic training set expansion, a manual verification of the training set is done using the crowdsourcing platform Crowdflower. This platform allows the use of external persons to pay for the crowdsourcing or to use personal contacts free of charge. Crowdsourcing requires strict quality control or using trusted persons but it can quickly give access to a large number of judges and thus improve many machine learning tasks. Results show that the manual annotation of a large amount of biomedical images carried out in this project can help with image classification.

Introduction

Images are produced in hospitals in ever-increasing numbers [1] and provide crucial information for diagnosis, treatment planning and other tasks. Besides in clinical settings, images are also made available via biomedical publications. The biomedical open access literature of PubMed Central alone contained almost 2 million images in 2014. This creates a need for searching in the immense collection of images in institutions and on the World Wide Web, making the data accessible for reuse. Many tools have been developed for these tasks over the past 20 years [2].

Retrieval and classification of medical images have been explored to get additional information for reading and interpretation of medical cases [1] when open questions remain and thus help clinicians in their daily work. Although text queries are commonly used, the visual information of the images can enrich the search. Thanks to benchmarks such as ImageCLEF [4] or Visceral [5, 6] the retrieval and classification algorithms have been further studied and compared with sometimes more than 20 research groups participating. In particular, ImageCLEFmed [7] has been proposing several retrieval and classification tasks since 2005. The image and case-based retrieval tasks and the modality classification task are using articles from the biomedical open access literature. The goal of the retrieval tasks is to retrieve images/cases that are similar to a given image/case description. The image modality classification (for modalities such as X-ray, ultrasound, computed tomography, etc.) is used as one of the most important filters to limit the search results in existing systems. Such filtering can improve the precision of the search [8] and reduce the search space [9]. ImageCLEFmed also proposes a medical image classification task based on a proposed hierarchy including medical modalities and other image types appearing in the biomedical literature.

In ImageCLEF 2013 a training set consisting of approximately 2,900 images was distributed to participants and classification methods were evaluated with a test set of approximately 2,600 images. Both sets were obtained from a subset of PubMed Central [10] of more than 300,000 images. In the training set some of the image categories were represented by only a few images. Therefore, a training set expansion strategy was applied to our multimodal (visual and textual based) classification approach to improve the accuracy precision (from 69.63% to 71.87%) [11].

Crowdsourcing has recently emerged as a tool in bioinformatics for solving large volumes of simple human tasks [12]. In this article we propose using crowdsourcing for two tasks: to verify the automatically detected modality of approximately 17,000 images and to reclassify the images identified as wrongly classified. Each of these tasks can be solved in a short amount of time (a few seconds) by users familiar with medical images. A short tutorial is also given in the crowdsourcing platform to explain the task and allow quality control. Crowdsourcing was recently used for image annotation in medical imaging, e.g. for evaluation of medical pictograms [13] or for retinal fundus photography classification [14]. Results show that the manual annotation can improve automatic classification tasks.

Methods

In this section the ImageCLEFmed tasks used in this project are presented. The details of the crowdsourcing performance are also explained.

Medical ImageCLEF tasks

The ImageCLEFmed benchmark proposes a standard test bed for medical image retrieval that allows researchers to compare their approaches on large and varied data sets including manually generated ground truth [2]. The image-based retrieval task aims to retrieve images for a precise information need, expressed through text and example images. On the other hand, the case-based task aims to retrieve cases that are similar to the query case and are useful in differential diagnosis.

Using the modality information of the images can help in the retrieval process to focus on one modality or to remove nonclinical images entirely, thus improving the retrieval performance [14]. The goal of the ImageCLEFmed modality classification task [7] is to classify the images into medical modalities and other image types, such as computed tomography, X-ray or general graphs using the modality hierarchy shown in Figure 1. The work presented in this paper aims to improve the modality classification accuracy and to integrate it into the medical retrieval system to enhance and filter its results.

Training set expansion

Previous work [2] describes the baseline used for the automatic image modality classification. It consists of a multimodal approach based on multiple visual descriptors and a Lucene [16] baseline using text information. This approach achieved an accuracy of almost 69%.

To improve the classification accuracy, a training set expansion strategy was applied to better represent all image categories [3]. For this purpose, the dataset of ImageCLEF 2013 medical image retrieval task was used. Each image from the original training set was queried against this dataset and the top results were assigned the class labels of the query. Results that were retrieved by multiple queries belonging to different classes were discarded. This automatic labelling resulted into a larger but "noisy" training set.

Crowdsourcing

Continuing the work in [17], the Crowdflower [18] platform was chosen since it provides an internal interface to be used by a known set of experts. For our experiments, eight experts in the medical imaging domain participated in the crowdsourcing job.

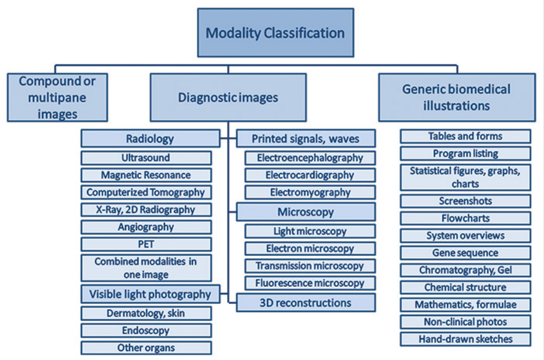


Figure 1

The image class hierarchy used for image classification.

Given the training set from the modality classification task with approximately 2,900 images, an expanded set of images was automatically classified as described in the previous section. The internal crowdsourcing interface was used to verify the automatically assigned class for each of the 17,002 images of the new set.

A first crowdsourcing task was set up to verify the given tag but, since a large amount of images are compound or multipane images (about 50% of the figures in the biomedical open access literature [4]), an option to correctly define this class of images was added. Therefore, each image was presented with a key question formulated as follows (fig. 2):

Does the figure correspond to the category?:

- *Yes, perfect classification*
- *No, compound image*
- *No, wrong category*
- *Not sure*

Using an iterative approach as in [17], images that were wrongly classified by the automated training set expansion were manually reclassified in a second crowdsourcing iteration. The same procedure was applied to the "not sure" category.

In this case the user has to decide to which class each of the images belongs to, across the hierarchy presented in Figure 1. Therefore, the task was presented in a hierarchical structure where a broad class is first asked and then the subclass (fig. 3).

Since this was a more difficult task, each of the images was classified by two users. A third user judged the images in case of disagreement between the first two answers.

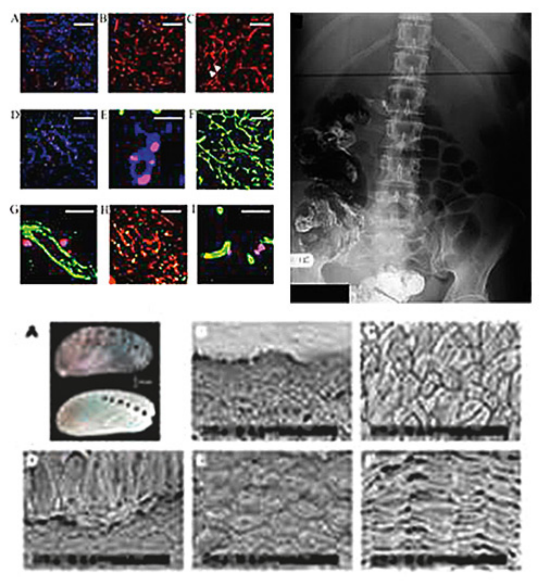


Figure 2

Images automatically classify as "Compound", "X-ray" and "Electron microscopy" respectively. Crowdsourcing was used to verify this image modality classes.

Results

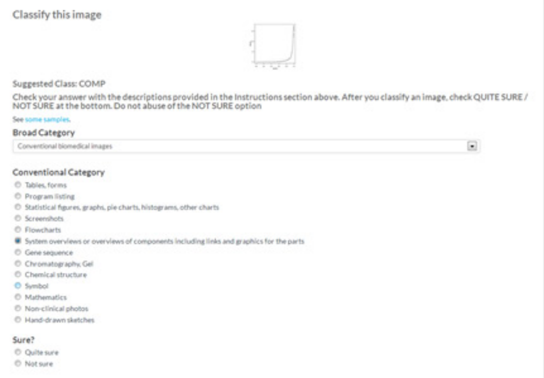
The crowdsourcing in this experiment was done with an internal team to limit errors in classification. A total of eight experts in the medical imaging domain participated in the task. In the past, external experts were used but strong quality control is necessary in this case, but on the other hand, tasks can be finished extremely quickly.

In the first step, 50% of the images were verified by crowdsourcing to augment the training set and automatically classify the remaining images. The results of the crowdsourcing task show that the automatic classification achieved an accuracy of 60% for this additional data set. Thanks to the first question in the platform, 21% of the images were reclassified as compound figures during the same crowdsourcing job. Almost 20 % of the images then had to be reclassified manually (fig. 3).

In the second part of the experiment, the correctly classified images and the images classified as compound were added to the initial training set. The new training set contained more than 9,000 images and was used to automatically reclassify the non-labeled images (images tagged as “wrong category” or “not sure”). A total of approximately 1,600 images were automatically reclassified and then verified via crowdsourcing. 16% of the previously wrongly classified images were now correctly labelled after the automatic reclassification. Figure 5 shows some images that were correctly reclassified.

In general, the second crowdsourcing task was more difficult for the experts. More knowledge about the modality classes was necessary and indeed the classes were not always easy to identify. Figure 6 shows some examples of images incorrectly labelled which experts found also difficult to classify. Often full-size versions of the images were necessary to clearly determine the modality.

Experiments showed that the automatic expansion of the training set improves the accuracy of the ImageCLEFmed modality classification task from ca. 69% to ca. 72%. Even better accuracy is expected using the new manually verified and larger training sets.



Classify this image

Suggested Class: COMP

Check your answer with the descriptions provided in the Instructions section above. After you classify an image, check QUITE SURE / NOT SURE at the bottom. Do not abuse of the NOT SURE option

See some samples.

Broad Category

Conventional biomedical images

Conventional Category

○ Tables, forms
○ Program listing
○ Statistical figures, graphs, pie charts, histograms, other charts
○ Screenshots
○ Flowcharts
● System overviews or overviews of components including links and graphics for the parts
○ Gene sequence
○ Chromatography, Gel
○ Chemical structure
○ Symbol
○ Mathematics
○ Non-clinical photos
○ Hand-drawn sketches

Sure?

○ Quite sure
○ Not sure

Figure 3

Screenshots of the crowdsourcing interface for image modality classification.

Conclusions

The goal of this project was to use crowdsourcing to improve the quality of an automatic modality classification task that uses the visual information of the images and the text of the figure captions. Increasing the size of the training set demonstrated how to improve the quality of the automatic classification. Manual correction of such a noisy training set can also significantly improve performance, and a crowdsourcing platform can help to simplify the process with a simple environment that can be used free of charge with known persons but can also be used with stricter quality control using a larger number of people participating in these platforms for a very limited cost.

The iterative nature of the shown task will continue to progressively generate a large and discriminative training set so all images of PubMed Central can be automatically classified and then made accessible for retrieval tasks.

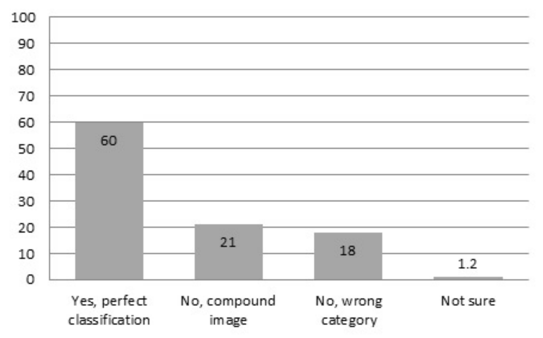


Figure 4

Each bar represents the distribution of each of the answers in the verification crowdsourcing task.

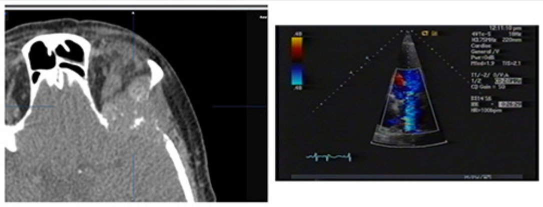


Figure 5

Images correctly reclassified after the training set expansion verification.

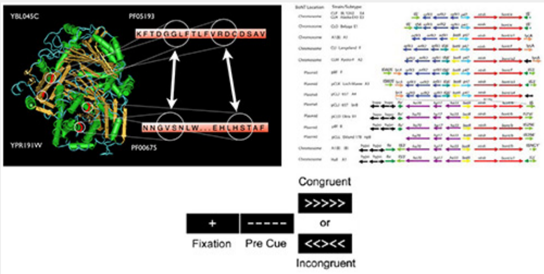


Figure 6

Images incorrectly classified automatically but that were also difficult to classify manually.

Correspondence:
Alba García Seco de Herrera
Techno-Pôle 3
CH-3960 Sierre
alba.garcia[at]hevs.ch

References

1 Akgül C, Rubin D, Napel S, Beaulieu C, Greenspan H, Acar B. Content-based image retrieval in radiology: current status and future. Digital Imaging. 2011;208–22.

2 Kalpathy-Cramer J, García Seco de Herrera A, Demner-Fushman D, Antani S, et al. Evaluating performance of biomedical image retrieval system – an overview of the medical image retrieval task at ImageCLEF 2004–2014. Computerized Medical Imaging and Graphics, 2014.

3 Uwimana E, Ruiz ME. Integrating an automatic classification method into the medical image retrieval process. In: AMIA Anual Symposium Proceedings, 2008.

4 http://imageclef.org/

5 http://www.visceral.eu/

6 Langs G, Müller H, Menze BH, Hanbury A. VISCERAL: Towards large data in medical imaging – challenges and directions. In: MCBR-CDS MICCAI workshop, 2013.

7 García Seco de Herrera A, Kalpathy-Cramer J, Demner Fushman D, Antani S, Müller H. Overview of the ImageCLEF 2013 medical tasks. In: CLEF 2013 (Cross Language Evaluation Forum), Valencia, Spain, 2013.

8 Rahman MM, You D, Simpson MS, Antani SK, Demner-Fushman D, Thoma GR. Multimodal biomedical image retrieval using hierarchical classification and modality fusion. Int J Multimedia Information Retrieval. 2013;23:15973.

9 Kalpathy-Cramer J, Hersh W. Multimodal medical image retrieval: image categorization to improve search precision. In: International Conference on Multimedia Information Retrieval, New York, USA, 2010.

10 http://www.ncbi.nlm.nih.gov/pmc/

11 Markonis D, García Seco de Herrera A, Müller H. Semi-supervised learning for medical image classification. In: Conference and Labs of the Evaluation Forum (CLEF), Sheffield, UK, Submitted.

12 Good BM, Su AI. Crowdsourcing for Bioinformatics. Bioinformatics. 2013; 29(16):1925–33.

13 Yu B, Willis M, Sun P. Wang J. Crowdsourcing parcipatory evaluation of medical pictograms using Amazon Mechanical Turk. J Med Internet Research. 2013;15(6).

14 Mitry D, Peto T, Hayat S, Morgan JE, Khaw K-T, Foster PJ. Crowdsourcing as a novel technique for retinal fundus photography classification: Analysis of images in the EPIC Norfolk Cohort on behalf of the UKBiobank Eye and Vision Consortium. PLoS One. 2013;8(8).

15 García Seco de Herrera A, Müller H. Fusion techniques in biomedical information retrieval. In: Information Fusion in Computer Vision for Concept Recognition, 2014, pp. 20–28.

15 García Seco de Herrera A, Markonis D, Schaer R, Eggel I, Müller H. The medGIFT group in Image CLEF med 2013. In: CLEF 2013 (Cross Language Evaluation Forum), Valencia, Spain, 2013.

16 http://lucene.apache.org/

17 Foncobierta-Rodríguez A, Müller H. Ground truth generation in medical imaging: a crowdsourcing based interative approach. In: Workshop on Crowdsourcing for Multimedia, ACM Multimedia, Nara, Japan, 2012.

18 http://www.crowdflower.com/

19 Chhatkuli A, Markonis D, Foncubierta-Rodríguez A, Meriaudeau F, Müller H. Separating compound figures in journal articles to allow for subfigure classification. In: SPIE Medical Imaging, Orlando, FL, USA, 2013.

Figures (large format)

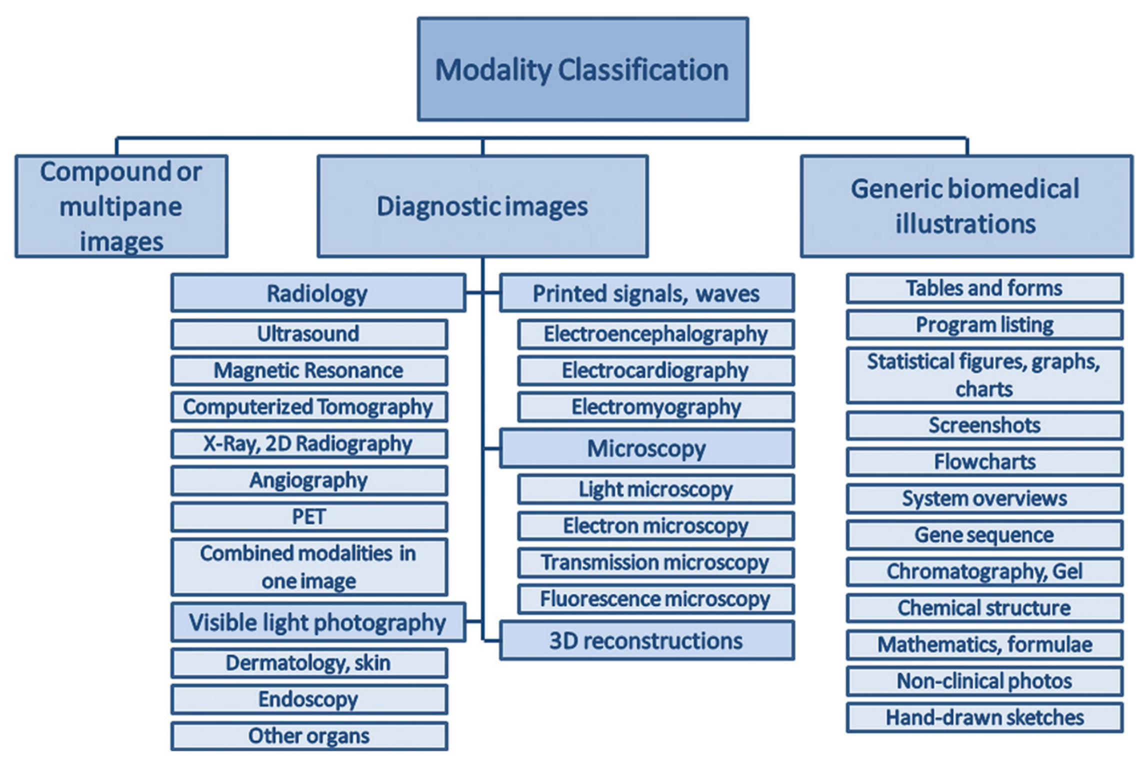


Figure 1

The image class hierarchy used for image classification.

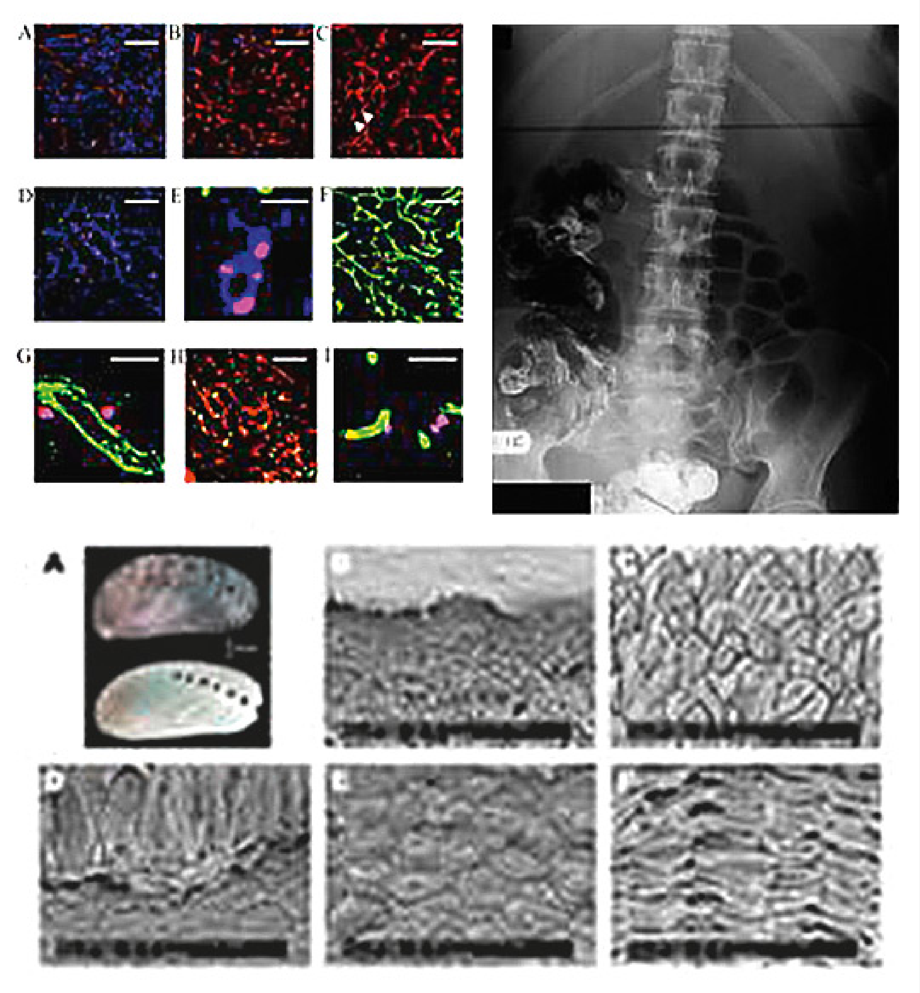
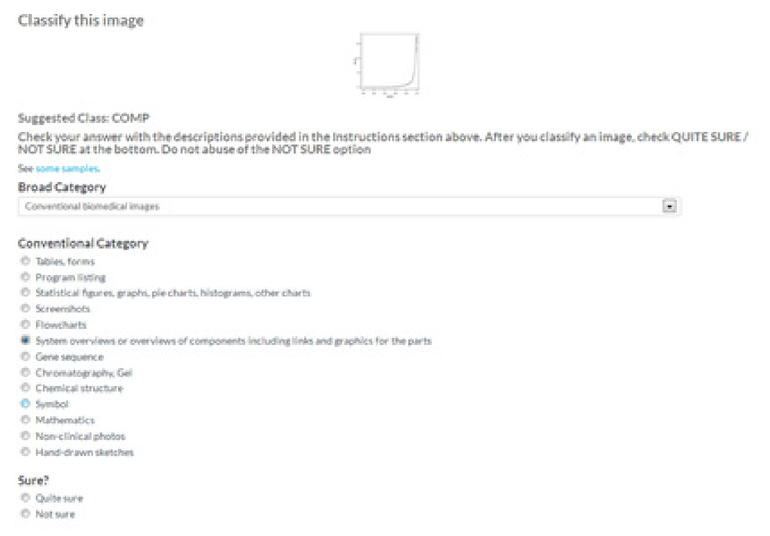


Figure 2

Images automatically classify as “Compound”, “X-ray” and “Electron microscopy” respectively. Crowdsourcing was used to verify this image modality classes.



Classify this image

Suggested Class: COMP

Check your answer with the descriptions provided in the Instructions section above. After you classify an image, check QUITE SURE / NOT SURE at the bottom. Do not abuse of the NOT SURE option

See some samples.

Broad Category

Conventional biomedical images

Conventional Category

- ○ Tables, forms
- ○ Program listing
- ○ Statistical figures, graphs, pie charts, histograms, other charts
- ○ Screenshots
- ○ Flowcharts
- ● System overviews or overviews of components including links and graphics for the parts
- ○ Gene sequence
- ○ Chromatography, Gel
- ○ Chemical structure
- ○ Symbol
- ○ Mathematics
- ○ Non-clinical photos
- ○ Hand-drawn sketches

Sure?

- ○ Quite sure
- ○ Not sure

Figure 3

Screenshots of the crowdsourcing interface for image modality classification.

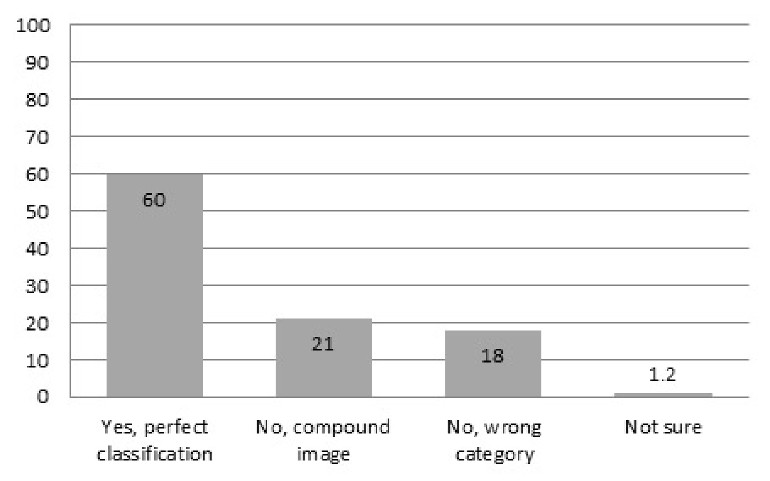


Figure 4

Each bar represents the distribution of each of the answer in the verification crowdsourcing task.

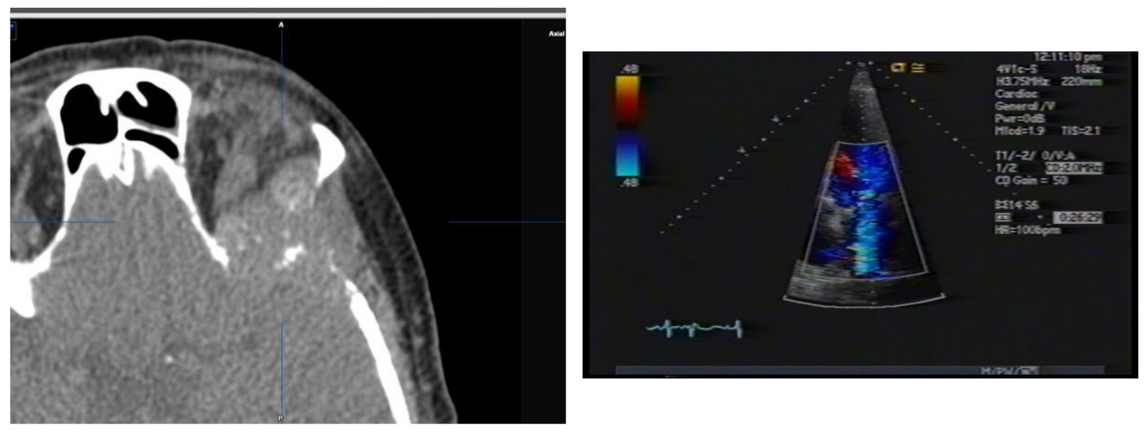


Figure 5

Images correctly reclassified after the training set expansion verification.

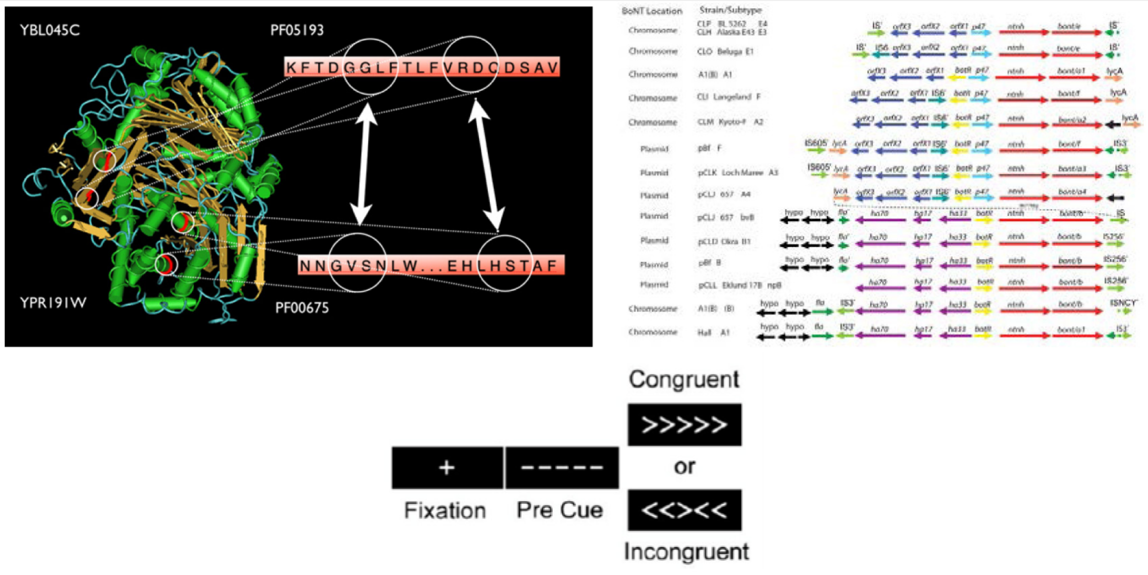


Figure 6

Images incorrectly classified automatically but that were also difficult to classify manually.