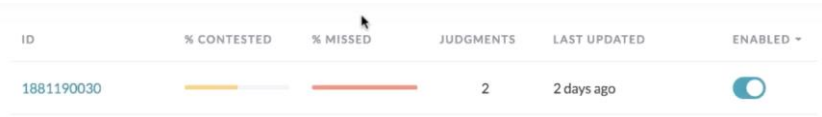
Project Proposal

Ashkan Yousefi

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	The radiologist needs to check many images of the patients in a daily basis and in many cases the image is normal. I am planning to build a tool which can identify the suspicious cases of pneumonia and help the doctor to focus more on cases which are not normal. This approach could save time of the doctor and subsequently save money for the whole healthcare chain including hospitals, medical clinics as well as patients.
Choice of Data Labels What labels did you decide to add to your data? And why did you decide on these labels vs any other option?	I have three labels 1-normal 2-pneumonia3-other. In addition, I have added the confidence level as well for the labeled images. I decided to use only three labels to make it easy for the annotators and it is enough for the product design to work efficiently. The normal and the pneumonia label will help to distinguish the healthy cases from the not healthy cases and the other label will help the annotators to not become confuse when unable to select the right choice for selecting the normal case from the pneumonia. In addition, the confidence level will help to filter the images which marked with low confidence level by the annotators.

Test Questions & Quality Assurance

Number of Test Questions Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?	I have developed eight test questions covering both normal x-ray image as well as pneumonia x-ray images. The annotators need to pass 75% to be able to move into the next step which is annotating the x-ray images.
Improving a Test Question Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?	 I will try to break down the task into few steps such as is the lung areas clearly shown? Then if the answer is positive, the second question will appear by asking "Is the diaphragm area and the shadow shown clearly?" and the third question could be "Is there any scattered cloudy area in the image? ". Then the possible suggestion will be shown and the annotator can select the suggestted option which in the described case would be normal x-ray image. Alternatively, the serial questions could be toward marking the x-ray image as Pneumonia as an example "Are both lungs clearly shown in the image without any cloudy area?". If the answer is false, then the next question could be "Is any of the lungs covered by the cloudy area?". The questions could be based on a decision tree which lead to selection of normal or pneumonia case and simply the decision-making process by breaking down the task into smaller bits.

Contributor Satisfaction

Say you've run a test launch and gotten back results from your annotators; the instructions and test questions are rated below 3.5, what areas of your Instruction document would you try to improve (Examples, Test Questions, etc.)

I would look at the breakdown of the statistics and try to improve the weakest points which in this case includes instructions and test questions. I am going to review the test questions and try to understand if an individual can answer the test questions with no background about the subject with only focusing on the instructions. In addition, I will revise the task given to see if there is a possibility to break it down to the smaller tasks which is easier to handle for the annotators by deciding a decision tree and creating a series of questions which could lead to specific answer and decision after few questions. Furthermore, I will try to add more interactive examples such as creating more graphical instructions using gif files to show the important points which annotator need to focus on and follow to be able to reach to the correct decision.

Limitations & Improvements

Data Source Consider the size and source of your data; what biases are built into the data and how might the data be improved?	The source of data could be biased as a result of selection from the specific source or the size of the data set could be small and does not reflect the real world. To solve this issue, it is necessary to use combination of the available data sets. In addition, feeding the data into the model need to be balanced between the labels as an example 100 normal image and 100 images with pneumonia.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	After the first product release the frequent updates to the model is required by using the new source of data and a feedback loop from the product can help to adjust the products and push it to the next level. As an example, analyzing the accuracy indexes such as the F1 score and the accuracy and recall values can help to provide a feedback on how the model is functioning. The new data feed to the model could be an effective approach for improving the indexes and enhance the functionality of the model. Particular attention is required to feed the new data sources which the model shows weak and not satisfactory performance.