

# Project Proposal



Sara EL-ATEIF

---

## 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 project's aim is to detect if a child is affected by pneumonia or not by feeding a chest x-ray image to a ML model that will classify the image into healthy or containing pneumonia.

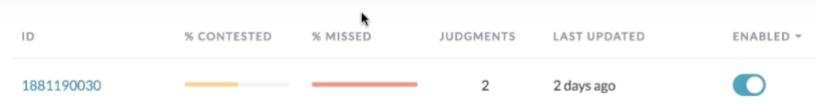

ML will help doctors quickly identify cases of pneumonia in children, which will reduce the time that a doctor would spend on reading x-ray results of chests that are healthy and spend more time on treating children that actually suffers from pneumonia.

### 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 choose to treat it as a classification problem where the ML will help flag a chest x-ray image as healthy or not as it seems easy to realize and safe because in some of the images I have gone through it's extremely challenging to detect where are the areas that are affected. I also choose to add a scaling factor on how sure a labeler is of their answer so that the images with low confidence could be reviewed by a professional to help lift the doubt and avoid any misshape that could endanger someone's life.

## Test Questions & Quality Assurance

<p><b>Number of Test Questions</b></p> <p>Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?</p>	<p>I choose to go with 16 questions 44% of them flagged as No (doesn't suffer from pneumonia or no signs) and another 44% as flagged as Yes (presence of pneumonia signs) while another 13% flagged as Unknown to cover uncertainty.</p> <p>I would have preferred to cover more questions, but I found very few `No` answers and many `Yes` answers, to avoid bias issues I took the above decision.</p>
<p><b>Improving a Test Question</b></p> <p>Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?</p>	 <p>I would first review the question, try to understand from where the problem comes and then consult with the annotators to redesign and adjust the test questions and the whole annotation job.</p>
<p><b>Contributor Satisfaction</b></p> <p>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.)</p>	 <p>From the above results, it seems that the most urgent to improve are the Test Questions, but I would still go over the examples with a few annotators and ask for feedback to improve and clarify the instructions, give better examples and design better test questions.</p>

## Limitations & Improvements

<b>Data Source</b>  Consider the size and source of your data; what biases are built into the data and how might the data be improved?	From my exploration of the data, I found out that there are many data points with pneumonia more than healthy ones. To improve this dataset I would search for other similar data points that covers different conditions of the image quality and capture positions and try to keep the healthy and pneumonia examples even.
<b>Designing for Longevity</b>  How might you improve your data labeling job, test questions, or product in the long-term?	After deploying the product after an initial run, I would collect more data points from hospitals, get user feedback and work in collaboration with them to implement their suggestions and update the model to be more accurate and generalized.