

Project Proposal

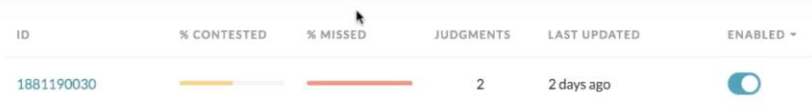



<Rahul Sarkar>

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 goal of this project is to build a classifier which is capable of detecting from the lungs image, whether a person is suffering from pneumonia. This will help to automate the existing process and reduce the patient-doctor cycle time in areas with limited doctors by paying immediate assistant to those patients first who are found positive in this process.
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 chose binary classification in this process (i.e. 0 for healthy or 1 for pneumonia) as a patient can have only 2 outcomes in ideal scenario i.e. either he/she will be suffering from pneumonia or healthy. Advantage of this label is that it helps to directly identify the real outcome. But one disadvantage is that it doesn't help in identifying the stage of the disease i.e. whether the patient is in initial stage or advanced stage of the disease.

Test Questions & Quality Assurance

<p>Number of Test Questions</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>An ideal size of the test set can be around 5% of the total observations in the dataset. This will help to check the number of false positive and false negative cases in the result and try to understand the factors that could have led to wrong output label.</p>
<p>Improving a Test Question</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>	 <ul style="list-style-type: none"> • First, we should provide clear description on what can be the possible reason for the wrong labeling to the annotator and how it can be improved. • We can also think on how the data source can be improved to avoid such cases. • We should also try to cover all types of possible cases in our <i>example set</i>, so that annotators can get a clear idea. • We should ask the confidence level of the annotators for cross checking those questions with low confidence level.
<p>Contributor Satisfaction</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>First, I would make improvements in the 'Steps' section, so that annotators can understand it easily. Then I would try to include all possible scenarios in the Example section, so that annotators can get a rough idea of the upcoming task and its complexity level. For the test questions, I would try to improve the data source and quality of the images.</p>

Limitations & Improvements

<p>Data Source</p> <p>Consider the size and source of your data; what biases are built into the data and how might the data be improved?</p>	<p>Since the dataset is about pneumonia detection, there is a high chance that there can be class imbalance i.e. the count of healthy images will be far more than pneumonia images. Thus, we train our classifier on this imbalance dataset, it might provide the result as 0 (healthy image) on most of the cases. Therefore, it is important to perform up-sampling of the pneumonia class, so that we don't have this internal bias.</p>
<p>Designing for Longevity</p> <p>How might you improve your data labeling job, test questions, or product in the long-term?</p>	<ul style="list-style-type: none">• Cross check the questions which have been marked with low confidence level by the annotators to ensure quality output and prevent errors.• We should try to provide clear instructions in the 'Step' section.• We need to cover all the possible cases in the Example section to make it more clear for the annotators.• We need to ensure that high quality images are being provided for this task.• We need to provide clear description on what can be the possible reasons for the wrong labeling to the annotator and how it can be improved in the Test Questions section.