

# Project Proposal



MICHAEL KWONG HIN SANG

---

## Data Labeling Approach

<b>Project Overview and Goal</b>  What is the industry problem you are trying to solve? Why use ML in solving this task?	<p>This project is to build a product as a diagnostic aid that can help doctors quickly identify cases of pneumonia in children.</p> <p>I am using ML in this task as there are inputs pairs of x-ray images and the desired outputs that are supervised by the data annotators. The ML algorithm/model can learn from this dataset of inputs and outputs, and be able to create an output for an input it has even never seen before without any help from a human.</p>
<b>Choice of Data Labels</b>  What labels did you decide to add to your data? And why did you decide on these labels vs any other option?	<p>I am using a binary classification supervised model with a yes/no labelling. The yes class is for an x-ray image that indicates pneumonia. The no class is for an x-ray image without pneumonia. I also added an unknown option if the annotator is unsure if the x-ray image shows pneumonia or not and need another opinion.</p>

## 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 developed 16 test questions to prepare for launching the data annotation job. This represents more than 13% test questions given the size of the dataset of 119 x-ray images. The test questions allow the annotator to provide reasons. This helps to improve the data annotation quality.</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>If multiple annotators contest a test question, I will have a confidence level set for each. The confidence level will then indicate the probability of someone choosing the binary yes/no classes and thus give a better idea of how the question was answered, whether with confidence or not.</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>In my Instruction document, I would improve the data source. The x-ray images may be taken from different machines in which case they may not show the same clarity and resolution. Also, I would ask for better x-ray images that show clearly the lungs and diaphragm.</p> <p>I would also improve the test questions by adding simple and accurate instructions and make it easier for annotators.</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?	<p>In the data, the answer distribution has a balance between yes and no to avoid biasing distribution. Too many test questions in one class would create bias in the data.</p>
<b>Designing for Longevity</b>  How might you improve your data labeling job, test questions, or product in the long-term?	<p>In the long term, I would continuously look at the results from annotators and continuously have the AI model train with more data.</p> <p>I can create test questions for each image with the annotator having ability to provide reasons to improve data annotation quality when they are unknown. Also, annotators can discard x-ray images that appear obviously unclear. In this way, the next data to train can be more accurate. The algorithm of Appen platform will learn, and in the long term, the model will improve as it goes through more samples of these improved new data.</p>