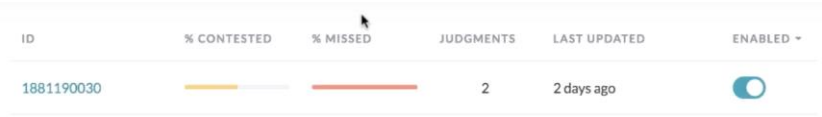
Project Proposal

Rahul Balaji

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	**Pneumonia** - is an infection in one or both lungs. Bacteria, viruses, and fungi cause it. The infection causes inflammation in the air sacs in your lungs, which are called alveoli. The alveoli fill with fluid or pus, making it difficult to breathe. Unfortunately, children and elderly are the most susceptible to pneumonia. The best method today for detecting Pneumonia is by using a chest X-ray. So, our goal is to help doctors quickly determine if there are pneumonia symptoms in the images we provide. We use Machine Learning algorithms, which will be trained on healthy and Pneumonia diagnosed Chest X-rays. So, with various classification methods of ML allows us to quickly analyze X-ray scans in an automated way. ML models trained with various inputs related to pneumonia are more accurate and with that we can save more number of children and elderly people.
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 used three labels such as “Yes”,”No”,”Unknown”. Yes label corresponds to the presence of signs of pneumonia whereas No label corresponds to the clear lungs and no signs of pneumonia, the third label “Unknown” is used for, If confused with the signs of symptoms in the image which is not known. The other options are “Positive”,”Negative”,”Maybe”, to make it simple I used “Yes”,”No”,”Unknown”.

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?	They were 8 test questions initially developed, in which 3 were "Yes" labelled, 3 were "No" labelled and 2 were "Unknown" labelled, and also Skip optioned is used to skip the images which are same to the previous labels. So, there will be no bias towards any specific label.
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?	 1. The Instructions might not be clear, so we shall provide more clear and simple instructions 2. We should also provide more examples then before to make it more understandable
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.)	 1. We can rephrase the question to remove any ambiguities 2. We should make the rules clearer and simple 3. We should also provide various examples with various test cases to make it very easy to understand

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?	X-ray Images can vary in brightness, contrast etc, this could lead to cases in which healthy lungs appear “cloudy” and lead to mislabeling. The size of the dataset currently that we are dealing with is not large enough for a machine learning model to learn patterns. We might need some more data for the ML model to be robust enough to deal with all possible scenarios. Only 16 labels are given, 8 are one label and 8 are of other. Data source could also be improved to be diverse and have more variety like images with different lighting conditions, illuminations, cropped etc.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	1. Test questions can be improved as you come across new data and with more edge cases. 2. Rules and tips also might need to be updated to reflect that. 3. Updating the dataset will large number of x-rays and every time encountering sign of Pneumonia will be helpful for obtaining more accurate model