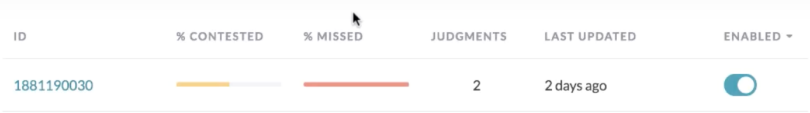
Project Proposal

Sukanya Rammohan

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	According to the Center for Disease Control, pneumonia, which is an infection of the lungs, kills children below 5 years. As per their report, "In the United States, 1.3 million people were diagnosed with pneumonia in an emergency department during 2017" (cdc.gov). With the advancement in Artificial Intelligence, it is possible to identify images with disease indicators. Utilizing AI for Healthcare would greatly help the doctors and the patients in reducing wait times, easy follow-up and timely deduction of disease. The goal of the project is to help doctors identify possible pneumonia cases in children by quickly identifying serious cases from healthy ones.
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?	The labels added for the images are • **Yes** for images showing indicators • **No** for images not showing pneumonia indicators • **Unsure** for images which overlap between the two classes. For this project, the deduction only falls under these three categories, since the option 'unsure' ensures that no image is discarded only on grounds of the perceptions of the moderators.

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?	The number of questions developed are 35 so as to have better results. Out of 117 data points, having 35 as the sample improves performance as it is about 30% of the total data
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?	 • Optimizing the instructions with more examples would improve performance. • Instead of giving just the generalized indicators of pneumonia, adding more questions on the visibility of diaphragm, its shadow and presence of cloudy visualization will help annotators. • Adding follow up questions to the existing questions with drop down pre-determined explanations options with Likert scale response of their perceptions of these options would strengthen the confidence level of the answer choices. Then adding a threshold of 70% would ensure resolving the 3 classes into binary.

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.)	 • From the above results, the Ease of Job option has the lowest rating, which implies the complexity. Providing more comparable examples would help the annotators find the job easier. • If there are multiple test questions, ordering the sequence of the questions and enabling questions based on validation will ensure ease of job as well as fairness of test questions. • Enabling the instructions to display next to every questions will make the job easier that the annotators do not need to scroll the page several times in case they run into unsure options.

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 data provided is not complete, as it did not account for information about the age, gender, race, ethnicity of the patients, whose X-ray samples are provided. Also, the data provided is very limited. For any data to produce accurate results, more samples are required. Moreover, the quality of the images plays a huge role as some of the images are blurred and have the potential of being misclassified.

Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	To improve the data labeling job, the following improvisation techniques can be adopted • Increasing the quantity of data • Providing more information on attributes and including more information on the age, gender, ethnicity of the patients • Improving image quality or standardizing the samples for uniformity in terms of the size of the image, pixel quality and so on. • Continuous re-evaluation of the failed questions and optimizing those questions for better quality • Getting feedback on test questions from annotators will help improvise questions and having the annotators as stakeholders in the job would improve ratings