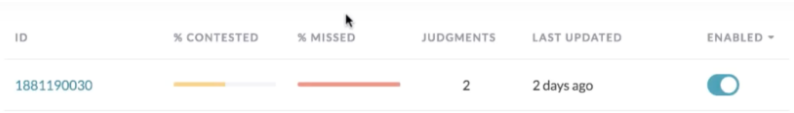
Project Proposal

Raqiya Al Maqbali

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 industry problem I am solving is the need for early and accurate identification of pneumonia in children. Pneumonia is a life-threatening condition, and delays in diagnosis can result in severe consequences. By leveraging machine learning (ML), I can create a system that aids doctors in quickly distinguishing healthy lung X-rays from those showing signs of pneumonia, thereby improving diagnosis speed and accuracy. The machine learning solution will allow for high throughput screening of X-ray images in healthcare settings, assisting doctors in prioritizing cases that need urgent attention.
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 have decided to use two primary labels in our dataset: (Healthy) and (Pneumonia). The decision to use these labels comes from the fact that this binary classification will help doctors easily differentiate between normal lung function and pneumonia presence. The goal is to streamline the identification process, focusing on distinguishing between healthy and diseased lung images rather than introducing more complexity by adding intermediary or unrelated conditions. Additionally, adding more detailed labels could increase the risk of confusion and errors among annotators.

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?	Given the size of the dataset, which consists of hundreds of X-ray images, I have decided to create 10 test questions. This number will ensure that the annotators understand the guidelines and can accurately distinguish between healthy and pneumonia-infected lungs. These test questions will be designed to cover a variety of image qualities, including borderline and obvious cases, to better evaluate the annotators' ability to correctly label the images.
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?	 If a particular test question was missed by almost 100% of the annotators, I would consider redesigning it by providing additional visual examples and clearer guidelines in the instruction document. For instance, we would clarify common areas of confusion, such as differentiating between pneumonia-infected lungs that may not present obvious cloudiness and those with clear, visible issues. I would also add more specific details about subtle visual markers of pneumonia, such as small localized opacities or diaphragm obscuration. Finally, I would include an "uncertain" option for annotators, allowing them to flag cases they find particularly difficult.

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.)

If the results from the test launch show annotators rating the instructions and test questions below 3.5, we would prioritize improving the clarity of the instruction document. Specifically, we would:

1. Add more examples of both healthy and pneumonia-labeled X-rays.
2. Refine the wording in the guidelines to reduce ambiguity, ensuring that annotators understand how to handle edge cases.
3. Include more detailed explanations for each label to reduce the chance of misclassification, especially for less obvious cases.
4. Offer more practice questions before the real test questions to help annotators become familiar with the labeling task.

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 comes from a set of X-ray images that have been taken under different conditions and varying quality levels. One potential bias in the dataset is that all images may not equally represent various age groups, races, or genders. Another limitation could be that the images are not always taken at the same resolution or angle, which might skew the results. To improve the dataset, we could: - Increase the diversity of the images by including more X-rays from different hospitals and demographic groups. - Standardize the X-ray image collection process as much as possible to minimize variability in image quality.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	To ensure that this project remains relevant in the long term, we would regularly review the labeling guidelines based on feedback from annotators and healthcare professionals. Additionally, we could: - Incorporate periodic retraining of the machine learning model with new data to ensure that the system continues to improve its accuracy over time. - Introduce more nuanced labels as the system becomes more capable, allowing for finer distinctions between different types of pneumonia and other lung diseases. - Allow for updates to test questions and guidelines as the medical understanding of pneumonia evolves, ensuring that the labeling process remains up-to-date with current diagnostic standards.