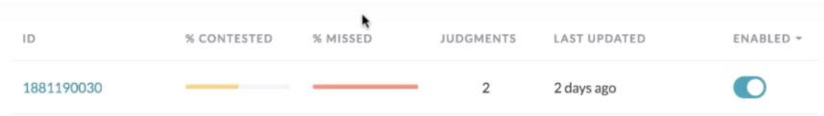
Project Proposal

Abdulelah Almajed

Data Labeling Approach

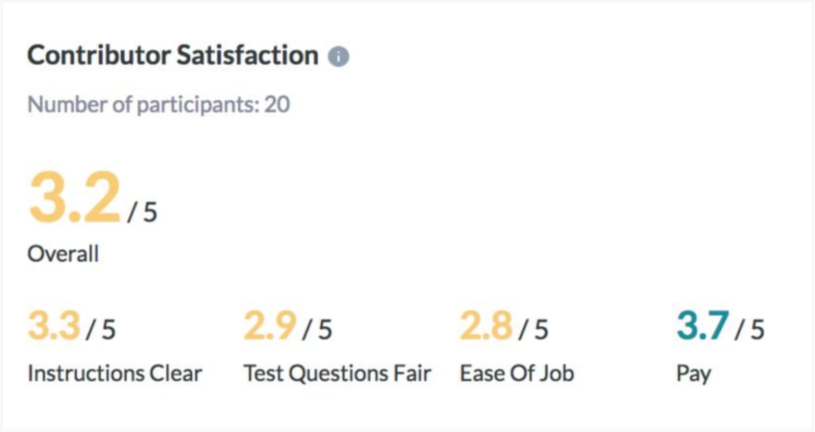
Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	The industry problem we are addressing is the accurate and timely diagnosis of pneumonia in children using chest X-rays. Pneumonia is a leading cause of childhood mortality, and early detection is crucial for effective treatment. However, accurately interpreting chest X-rays can be challenging, especially in resource-limited settings. By using machine learning to analyze chest X-rays, we aim to develop an automated system that can assist healthcare professionals in detecting signs of pneumonia quickly and accurately, improving the efficiency of diagnosis and ultimately leading to better patient outcomes.
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?	We have chosen the following labels for the pneumonia detection task: Presence of pneumonia: "**Yes**", "**No**", or "**Not Sure**". Observed pneumonia symptoms: "**Concentrated opaque area in the lungs**", "**Multiple smaller opaque areas throughout the lung area**", and "**Obscured diaphragm shadow**". Likelihood of pneumonia (when "**Not Sure**"): A 5-point scale from **"Not at all likely**" to "**Extremely likely**". These labels capture the key visual indicators of pneumonia, the annotators' level of certainty, and provide relevant data for training the ML model while minimizing ambiguity.

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?	Considering the dataset size of 117 samples (101 unlabeled and 16 labeled), I developed 8 test questions using the labeled data. This aligns with Appen's recommendation of having more than 5% of the unlabeled data as test questions. These carefully selected test questions will help assess the annotators' understanding and maintain the quality of annotations throughout the data annotation job, ensuring reliable results.
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?	 ID | % CONTESTED | % MISSED | JUDGMENTS | LAST UPDATED | ENABLED 1881190030 | | | 2 | 2 days ago | To improve a test question that almost all annotators missed, I would: - Analyze the question for ambiguity, complexity, and alignment with the task. - Review statistics to identify patterns and common misunderstandings. - Gather feedback from annotators to understand their challenges and suggestions. - Revise the question based on insights, simplifying or clarifying as needed. - Update guidelines and training materials to address gaps in understanding. - Conduct a pilot test with a small group to ensure the effectiveness of improvements. - Monitor annotators' performance and iterate if necessary. By following these steps, I aim to identify the root cause, gather insights, and make targeted improvements to enhance clarity, align the question with task requirements, and support annotators in answering accurately, ultimately ensuring data quality.

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.)



Contributor Satisfaction

Number of participants: 20

3.2 / 5
Overall

3.3 / 5	2.9 / 5	2.8 / 5	3.7 / 5
Instructions Clear	Test Questions Fair	Ease Of Job	Pay

I would focus on improving the following areas:

- Instructions: Clarify and simplify the language, add more examples, and break down complex tasks into smaller steps.
- Test Questions: Ensure they align with the task, are unambiguous, and cover a range of scenarios. Revise questions based on annotator feedback.
- Examples: Provide more diverse and representative examples to better illustrate the task and expected outputs.

By targeting these areas, I aim to enhance the clarity and comprehensiveness of the instructions, test questions, and examples, ultimately improving annotator understanding and performance.

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 dataset, consisting of only 117 samples (101 unlabeled and 16 labeled), is prone to sampling bias, which could significantly impact the final predictions. Additionally, variations in image size and exposure times across the dataset may introduce measurement bias. To improve the data and mitigate these biases, we should: 1. Increase the dataset size significantly. 2. Ensure consistent image size and exposure times for all new data.
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, test questions, and product in the long run, I suggest: 1. Continuously update the model by training it on new data to adapt to emerging imaging technologies, symptoms, and diseases. 2. Regularly review and update the annotation job, incorporating more relevant definitions, examples, and test questions based on the evolving dataset. This dynamic approach ensures the model stays up-to-date and the annotation process remains aligned with the latest developments in the field.