Project Proposal

Mohamed Fazal Mustafa

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 a common condition that hospitalizes more than 5,000,000 Indians per year, leading to the deaths of more than 50,000 deaths[MoHFW].Currently due to Covid there has been an 60%-80% increase in Pneumonia Cases. However, pneumonia is both preventable and treatable if detected early enough. Unfortunately, children and the elderly are the most susceptible to pneumonia. Our goal is to quickly identify and distinguish between healthy, mild, and severe cases while also providing a diagnostic aid for doctors and medical technicians. Using classifications methods with machine learning will allow for an automated method that can quickly analyze x-ray scans. The accuracy of machine learning models can even exceed the detection accuracy of humans.
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 a binary classification with "yes", "no", and "unknown" as options. "Yes" corresponds to the presence of pneumonia where "no" is the classification for nominal lungs. I chose this option over a variable confidence method (e.g. likelihood) as the testers are likely not trained professionals the weighted results of their confidence values would imply the same result (binary result with unknown). This allows for testers to quickly respond in clear cases (yes and no) and have an option (unknown) for mild cases or false positives. I also considered implementing a labeled bounding box model but the goal is to detect whether or not the ailment is present and not particularly where it is located.

Test Questions & Quality Assurance

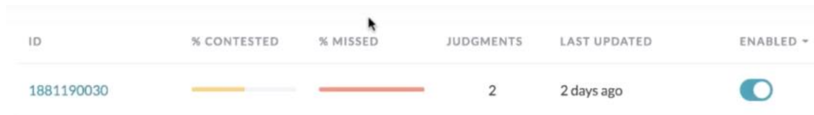
Number of Test Questions

Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?

Due to the challenging classification problem, having more examples will lead to a higher accuracy and mitigate ambiguity in edge cases. I provided 25 test questions in preparation for launching the data annotation job. This should provide testers with the means to evaluate cases that are not so clear as well as provide confidence in their selections.

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?



With a high failure rate, either the instructions are not clear, or the examples are not sufficient. I would improve both the instruction with more precision (if possible) while also providing more test examples. I would likely include more examples of "unknown" cases. Additionally, consulting with a medical professional might allow for some insights regarding addition cues related to the presence of pneumonia. If the results are still failing, then it would be worth considering redesigning the test questions.

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.)

Given that an overall value of 3.2 is quite poor, I would need to analyze areas for improvement. From the metrics results, the "test question fairness" and "ease of job" values are quite low. However, this is a challenging task and reducing the "ease" will be quite challenging. However, we can improve the fairness and instructions. I would prioritize the fairness of the questions by providing more examples. Furthermore, I would attempt to improve the instructions such that resulting value is closer to 5.0

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, opaqueness, along with other optical distortions and occlusions. This could lead to cases in which healthy lungs appear "cloudy" and lead to mislabeling. Furthermore, there are other cues and conditions that occur with pneumonia other than cloudiness (like spots, clusters, and speckles). Finally, other lung conditions also lead to "cloudy" lungs. If possible, I would add more data (images) for testing and validation. Furthermore, I would consider two cases: a case in which all x-ray images has been acquired with the same lighting/imaging characteristics (consistency) and cases in which x ray images had been acquired with vastly different imaging characteristics. If all future x-ray images could be taken with the same specifications as the test images, then the number of unknowns would decrease. If future images were known to be acquired from devices with varying imaging specifications, then having variety in the images would prevent bias (i.e. like an offset).
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Naturally, the inclusion of medical staff, x-ray technicians, and medical researchers would assist in the labeling of x-ray images. I would attempt to consult with a specialist to help accurately determine cases of pneumonia as well as ask for their advice on providing instructions. Additional cues might be able to improve the accuracy (e.g. spots). Furthermore, I would attempt to use properly classified images that "clearly" provide examples of clear cases for future examples (feedback). I may even run a user study to gain insights on the quality of the instructions and examples if budget and time allowed for it.