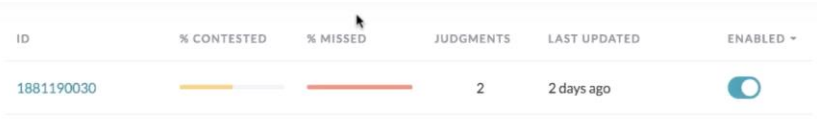
# Project Proposal

*Nouf Almutairi*

## Data Labeling Approach

| | |
|---|---|
| **Project Overview and Goal**<br><br>What is the industry problem you are trying to solve? Why use ML in solving this task? | The industry problem to be resolved here is to build a product that helps doctors identify cases of pneumonia in children. Using machine learning in the project helps expedite the process of getting accurate diagnosis. After completing the project time and effort will be saved to focus on more patient clinical care. |
| **Choice of Data Labels**<br><br>What labels did you decide to add to your data? And why did you decide on these labels vs. any other option? | I choose to have 3 labels for the contributors to pick from for this project, as it's required to identify the pneumonia. Since X-Rays are not clear most of the time it's best to have a 3rd option (other) to decrease the false negative/false positive cases and ambiguity in the results. |

# Test Questions & Quality Assurance

| | |
|---|---|
| **Number of Test Questions**<br><br>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 is small. I provided 8 test questions with 50% healthy cases, 25% pneumonia and 25% other. To have good representation of the cases in the dataset. |
| **Improving a Test Question**<br><br>Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question? |  <br><br>In this case, the question should be paraphrased to focus more on the details needed to spot the illness. We can change the data for clearer details to help the contributor spot the illness and classify it accordingly. |
| **Contributor Satisfaction**<br><br>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.) |  <br><br>In such cases I'll correct the instructions to be more detailed and precise and explain the purpose of said step to enlighten the contributor on the importance of said step. As for the test question I'll review them to ensure they're related to the topic and I'll with clear answers to help the contributor choose their answer. As for the ease of the job I'll review the examples and ensure that we have a representation for all the possible cases with prober explanation and clear labels, I'll also review the questions to make them more precise and detailed enough that we need as less as possible to represent the data size and problem. |

# Limitations & Improvements

| | |
|---|---|
| **Data Source**<br><br>Consider the size and source of your data; what biases are built into the data and how might the data be improved? | I think if we increase the size because 117 is smaller than average need for such problems. Also the dataset is relatively outdated for the topic it's 4 years ago and i think the best practice would be o use the current month dataset or in the better practice use the current year dataset collection. Also for biases in the data we only have two cases so there might be biases since the data is smaller and we might face generalization in the model with high false positive because we don't have enough data to represent the main characteristics of pneumonia. |
| **Designing for Longevity**<br><br>How might you improve your data labeling job, test questions, or product in the long-term? | For the long term purpose I'll add more test questions and varied examples to cover the main characteristics of the pneumonia. I'll increase the data size and choose more recent dataset to benefit from the enhanced X-ray technology in getting crisper image to represent the disease. I'll also enhance and increase the quality of the test question to make them shorter intuitive and precise. |