Project Proposal 

#### *Ahmed Abdelnasser Mohamed*



# 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 main goal is to help doctors quickly identify cases of pneumonia in children, Using ML here help doctors to quickly eliminate cases that do not have any pneumonia symptoms and spend more time on the cases where there are symptoms. |
| **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? | There are three labels “Pneumonia”, “Normal”, “Not defined”. The first two labels were chosen as we need to decide if there are pneumonia symptoms in the given images. The third label “Not defined” is chosen to leave room for uncertainty. We can have also chosen the labels as “Yes”, “No”, “Don’t know”. To keep it more specialty and professional, I used “Pneumonia”, “Normal” and “Not defined” labels. |

# 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? | 12 test questions were developed. 31% Pneumonia, 23% Normal and 15% Not defined |
| **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? | Take a step back and check if the rules specified are clear and unambiguous. Also provide a more detailed description so that the annotator knows why it was labeled the way it is. |
| **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.) | I will provide some Tips to help annotators, also provide more examples for all cases and for hard cases to reduce time and any negative feedback. |

# 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 size of the dataset currently that we are dealing with is not enough for a ML model to learn patterns. We might need some more data for the ML model so we can count on it, but if there are biases in the dataset, we need to account for it either by augmenting the class that does not have more labels or throwing away some data from the class that has more data. We can improve data by adding images with different views and different light condition to be more accurate. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | Questions can be improved and also providing new images with hard cases to improve the ML, and see the feedback of annotators about labels to improve it. |