Project Proposal 

#### *Manal Almehmadi*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | My goal is to build a product that helps doctors quickly identify cases of pneumonia in children. It will act as a diagnostic aid for doctors.  These classified images will be used by ML engineers to develop a classification system. With this system, we can analysis x ray images to quickly identify healthy patients and surfacing potential cases of pneumonia. |
| **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’ve chosen straightforward labels which is Yes, No and Unknown. It's included an unknown option to handle failure cases or unable to answer.  I decided to make three questions to get detailed information.  First question to identify pneumonia cases and has three labels:   * Yes: to indicate pneumonia case. * No: to indicate healthy case. * Unknown: to indicate undetermined case.   Next question to know where the area of cloudiness/opacity and has two labels:   * several concentrated areas * one large area   Last one is about which side of lung is affected:   * Right * Left * Unknown |

# 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? | I did 24 test questions. It includes an example for all labels to avoid bias towards any specific label. |
| **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? | * Make questions more descriptive. * Provide 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’ll try to clarify instruction and make steps more descriptive. |

# 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 data is not enough to train model especially a model to serve medical industry. Also, data source could also be improved to be diverse like images with different lighting conditions or x ray machine. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | Make the job detects other symptoms of lung-based diseases such as Enlarged heart, Breast cancer and Blocked blood vessels. |