Project Proposal 

#### *Abdulrahman Jamman Almalki*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | This data annotation job will help determine if there are cases of pneumonia in the images we provide, it will help doctors quickly identify cases of pneumonia in children.  The ML will help detect diseases faster to help doctors to react fast and save time and lives. |
| **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 decided to use (yes),(no),(Unknown),and(other) as a select options because it is easy for people to understand where I chose (yes) and (no) to allow the participant to determine whether there was an infection or not and (unknown) if they are unsure, and if they noticed something unusual they can select (other) and specify it in the box below it. |

# 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 have included 5 test questions that is 5% test questions to mix into the training set or about 1 test question for every 19 data points. |
| **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? | By designing an option with the least negative impact like (unknown) or (other) so when the participant don’t know the answer they don’t have to select (yes) or (no), I also might move the Question to the example section. |
| **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 audit the result check the feedback then go back and update the instructions or the design, create more test questions and improve on what they failed at the most, after that I will improve the examples and explain them more. |

# 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? | I think the data have **Sample bias** where the collected data have more **normal** images than the infected ones, I can be improve the data by balancing the data between **normal** or **infected** by add more images of infection from multiple sources or removing some of the normal images. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | I might use a dynamic model and train it so if it encounters a new infection it will detect it as a new one to avoid it being labeled as pneumonia or other infection. |