Project Proposal 

#### *Munirah Mohammed Aldossary*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | In this project I designed a product that will help the medical industry. The product will help the doctors to identify cases of pneumonia in children faster.  I used ML to design an annotated data to identify the pneumonia in the x-ray and recognize the area of it. |
| **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 image annotation because the product needs to identify the pneumonia in an x-ray image. Then I used the Image categorization because the pneumonia could be as cloudiness in some area in the lungs.  the strength of the image categorization that the annotator could specify the cloudiness area from a full-size x-ray image, either it is on the lungs or diaphragm or other area.  On the other hand, the weaknesses of image categorization is limit the annotator somehow, so we fixed it by adding option “other” to write the other area that did not mentioned in the options. |

# 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 prepared 30 test questions, and I tried to have equal answers for each choice.  We place 47% of yes option including three different areas, And 33% for no option. Also, some other area might present in the x-ray image.  30 test questions are representing 35.1% of the 117 data image.  Graphical user interface, text, application  Description automatically generated |
| **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? | I will check the confusion that they have, then I might add some tips to the rule’s tips in the begging of the question. Also, I could add reasons to read the right answer after they submit the question. |
| **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 would improve the examples to add more complicated x-ray images. Also, improving the test and add more clear reasons to it. |

# 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? | One of the biases is the unknown images. I could improve these images by image processing and make it more clear to the annotators to identify the Pneumonia and the area of it. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | I might include more complex x-ray images and add more area choices that will help to categorize more specific could area and more accurate Pneumonia results. Also, I could add a feature where the product conducts the type of the Pneumonia, and according to John Hopkins medicine there are 4 main types. |