Project Proposal 

#### *Mohammed Tayor*

# Data Labeling Approach

| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | According to WHO, Pneumonia is the single largest infectious cause of death in children worldwide. As any other disease, the early diagnosis is key to preventing any complications. The mostly used medical tool to sport such cases is through x-ray scans. While medical professionals have rules to confirm or deny the presence of pneumonia based on known rules, it is still not an exact science and can’t be quantifiable by normal algorithms. Hence the use of a learning-based approach to distill the knowledge of pneumonia detection from multiple scans including both normal and sick cases. |
| --- | --- |
| **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? | As the task is a classification task primarily between healthy and pneumonia cases, choosing two outcomes of “yes” and “no” respectively for presence of pneumonia and its absence seems to be the logical path. To have an insurance against the quality of annotators we add a third class of “maybe” in case they have confusions of a scan and don’t have a clear decision according to the given rules. |

# 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? | 9 test questions are included which represent more than 7% from the 117 total cases. |
| --- | --- |
| **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? | First thing that comes to mind is the unclear description of the problem and the rules. The increase of test examples also is another way to give more information to annotators. Gathering feedback from annotators or looking for more expert medical help can also solve the issue. |
| **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.) | From the individual scores, “ease of job” and “test questions fair” are the worst contributors to the overall bad score. Adding more detailed description, rules , and examples can clear a lot of confusion to annotators and make the job more fair and easy. |

# 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? | First the samples are too small for a computer vision task. We need to collect more samples. Also the more diverse scenarios the better. We can use augmentation techniques as a quick fix. Some scans include noise pixels , so it also needs denoising. Balance of classes can also be a challenge, any data source skewed to one category will give an unsatisfactory distribution curve of the real data. So a pre-selection by an expert might enhance data quality. |
| --- | --- |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | A process of continuous feedback loop needs to be implemented. Selection, balancing, denoising of more new data should be directly influenced by the feedback of annotators from previous labeling tasks. |