Medical Image Annotation Project Proposal 

#### *Joyce Chidiadi*



# 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 goal of this project is to solve the problem of correctly identifying pneumonia in children from medical diagnostics images based on symptoms presented in the images provided.  The labeled images resulting from this project will be used by machine learning (ML) engineers to build models that will help doctors to quickly diagnose cases of pneumonia. The ML models will also help them to validate their medical decisions for pneumonia or possible pneumonia 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? | I have chosen binary labels (Yes or No) for classifying the pneumonia status and another label (Other), to represent images that are not clear or are totally different from the pattern of the majority of images in the dataset.  When a label has “Yes” it means that pneumonia is present.  “No” means the image does it show pneumonia and so it is a healthy image.  “Other” is for images that neither indicate presence or absence of pneumonia. This will help identify uncertain images.  Other labels that can be chosen are “Normal”, “Pneumonia”, and the third class, “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? | 12 questions were developed for this data annotation job. This number is about 10% of the total number of images in the dataset which is 117 images. |
| **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? | If there is or are question(s) missed by most of the annotators, it means that the question was not clear.  To improve the question(s),   * it needs to be rephrased. * more detailed information and examples will be added to simplify 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 will provide an example or tips for each test questions to improve the instructions in order to reduce ambiguity. |

# 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 number of images (117) in this dataset is very little to be used in building a reliable machine learning model.  The labels provided in the original data is not enough for training the data. It is not clear what kind of bias might be in the data but future inferences using the build model built will identify such biases.  **Improvement:** More data need to be collected to improve the outcome and usage of the machine learning model that will be built with the annotated data from this project. To build a good model, there needs to be data large enough for the model to learn the pattern in that data during training.  To avoid bias, the data source will need to be taken from a diverse source including children of different origins and those living in different types of countries and climates. This will help reduce bias.  The data should not have equal or nearly equal number of the three labels to avoid overfitting the model. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | To improve the design of the data labeling job for long term usage,   * more data from reliable sources will need to be added * the test questions will need to be reviewed from time to time to reflect or correct ambiguities and account for changes in the society and technologies * Medical professionals such as doctors, sonographers, medical laboratory analyst, need to be consulted to leverage on their experiences and expertise to improve the product and make it more reliable. * The machine learning model will be continuously trained and updated to improve the image classification results. |