Project Proposal 

#### *AYENI TRUST OLAMILEKAN*



# 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 project can be used to help doctors quickly identify cases of pneumonia in children.  As such, this project classification system distinguishes between healthy and pneumonia x-ray images.  Therefore ML (Machine learning) can be used to solve this task because there can be ‘n’ number of classes into which an image can be classified, manual classification (when there are thousands of images) is difficult for humans to do, that is why automating the process of object identification and classification with machines is gaining ground. |
| **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 chose to use the labels:   * YES * NO * UNKNOWN   These labels prove to be precise in answering the questions I have provided to the annotators where:   1. YES denotes Pneumonic cases 2. NO denotes Healthy cases 3. UNKNOWN denotes “ I’m not sure whether its pneumonic or not”. |

# 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 created 8 test questions before launching the data annotation job. |
| **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 would create more test questions in other to eliminate bias and increase the confidence level of the annotators and my model.  I can also involve multiple possible answers, make my test questions indistinguishable from the overall data set. |
| **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 focus on both areas so as to help the contributors look out on what they missed previously and to train them to look out for everything that might appear in my dataset. |

# 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? | With respect to the size of the data, the quality of a model’s output relies on the data on which it’s trained. Datasets can contain inherent bias due to the nature of their collection, measurement and annotation — all of which are rooted in human decision-making. Thus, smaller sample sizes of data compared to larger groups do not outshine the previous.  The data can be improved by **amplifying or boosting specific intersectional data inputs (getting more data from various sources)**. Doing this early on will inform the model’s training formula and help the system stay as objective as possible otherwise, the training formula may be unintentionally optimized to produce irrelevant results. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | I can do that by updating my data anytime a new case is found to include more relevant definitions and examples. I would also include more test questions that fits the changes in the newly updated model. |