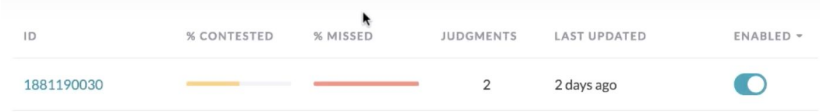
Project Proposal

Odili Charles Opute

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	About 6 billion of the worlds 7.8 billion people live in Asia and Africa, which is home to populated countries like Nigeria where the doctor:patient ratio stands at 1:6000 [1] against WHO's recommendation of about 1:1000. This means there is a fundamental problem of access to basic diagnosis and health care provision to most people living in these countries, especially for illnesses that early detection can be the difference between life and death. According to the WHO, Pneumonia (inflammation in the lungs) accounts for 15% of all deaths of children under 5 years of age, killing more than 800K children in 2017[3]. With limited access to qualified health personnel needed to carry out diagnosis of Pneumonia, Machine Learning can democratise early detection for even the world's poor by delivering a portable model that semi-skilled personnel or volunteers in health service centers can use to quickly detect diagnose the infection and identify those at high risk or those not infected. Using Machine Learning, the same or even better quality of service from a fully trained and experienced doctor, can be closer and more accessible to millions of people suffering from Pneumonia. 1. https://nimedhealth.com.ng/2019/09/14/disturbing-doctor-patient-ratio-in-nigeria-it-is-1-doctor-to-6000-patients/ 2. https://www.who.int/gho/health_workforce/physicians_density/en/ 3. https://www.who.int/news-room/fact-sheets/detail/pneumonia
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?	The labels for annotating the x-ray images in this project includes: 1. Yes - when there is evidence of Pneumonia infection 2. No - when there is no evidence of Pneumonia infection 3. Maybe - when there isn't certainty to mark and image as either a Yes or No.

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?	There are 11 test questions
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?	 1. Ensure the Steps, Rules & Tips sections create clarity for what the annotator needs to do and how. 2. Re-phrase the question making sure it can be represented under one of the labels
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.)	 Seeing that "Ease Of Job" has the lowest satisfaction rating and it could be tightly related to the rating on instructions clarity, I will optimise annotation job to improve instruction clarity, notably the Steps, Rules & Tips, and Examples sections. I will also review sequence and improve the clarity of the explanations provided with the test questions so that annotators feel they have the right level of information to prepare (through the tests) for the annotation job

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 data is not large enough nor is there an indication of how diverse it is, even for racial or gender diversity. We also do not know the age distribution of the children whose x-ray images we are using. These and more could introduce biases that might negatively impact the quality of inference we are able to make from model we intend to build. To improve our dataset and ultimately our model, we want to introduce a number of additional fields as metadata to the data. E.g nationality, age, gender. This allows the product manager and his/her team to evaluate the features in the dataset and only proceed to annotate the data if there is a healthy mix.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	1. By reviewing the test questions and improving the most failed question(s) 2. By reviewing the contributor satisfaction and improving relevant section of the annotation job 3. By getting and applying feedback on prediction accuracy from actual users of the deployed model