Project Proposal

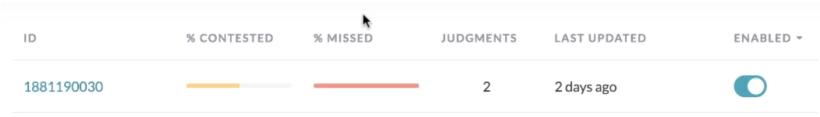
Chukwuemeka Ezumezu

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 is to build a product that will help doctors to quickly identify cases of pneumonia in children, by classifying an x-ray image as having pneumonia or not. It will serve as a diagnostic aid to doctors and never as a replacement for the real diagnosis. ML is used to build a model that can accurately/precisely identify pneumonia or healthy x-ray images after preparing and annotating ground-truth images.
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?	"Yes" for pneumonia images, "No" for healthy images, and "Unknown" to capture uncertainties. It is to make a simple binary annotation job, i.e a situation whereby the image is marked as a healthy or pneumonia image. But also being that the images are not being labeled by experts. The third option "UnKnown", is meant to capture the uncertainties, a situation whereby the annotators can not properly identify an image so that it can properly be labeled by experts later. The downside of this approach is can it can not identify different types of pneumonia and how simple or severe every case is.

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?	Having small datasets of 117 images, 9 equal answers distributed test questions are enough samples to get the annotators started and understand how to label the x-ray images. This also accounts for 5% of the whole dataset and also for 1 in every 19 data points. Making the job design effective enough to capture different 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?	 I will first need to know where things went wrong and why almost all of them missed the answer. In this case, I will see if the instructions need to be better clarified or include more samples to better understand the case. Or in some cases redesign the job. It's essential to capture all cases because what the contributor missed our model will also miss, affecting our model's entire performance.
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 try to improve the instructions and test questions, rephrasing them, making sure I use simpler, straightforward, and easy-to-understand grammar. I will also include more examples in the instructions, most failed questions can be used as examples in institutions, to make the job as simple and easy to understand as possible.

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 image model is a deep learning model, having only 117 images without being split into train, test, and validation set will not be enough for our model to converge. It will build inaccuracy in the resulting model, classifying all data into one class or making irrelevant predictions. Deep learning models need a much larger and well-distributed dataset of all samples to learn properly. A much larger and well-distributed dataset needs to be used to make sure it solves a real-life problem than just a random dataset that is likely not customizable enough to solve the specific problem.

Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	A static model can be used for a model that the dataset does not change over time, otherwise, a dynamic model is commonly used which requires regular updates of the model on new data to learn and meet up with the demand problem being solved. In this case, we need to regularly redesign our job by updating data and instructions to include more regular samples.