Medical Image Annotation 

#### *Ognen Loparski*

# Data Labelling 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 for the medical professionals dealing with pneumonia in children to spend more of their time on cases that need their attention and thus provide better care for children.  ML has the goal of quickly identifying clear cases that require no medical treatment as well as identifying ones that need to be attended by medical professionals immediately. |
| **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 2 labels:   1. Something is wrong, this could be a serious case 2. It seems to be a healthy case   After choosing a label, a drop-down appears to choose how confident the annotator is. The labels are good for categorisation, while the confidence level is giving me a way to measure the probability of errors.  I was toying with the idea of having a third label as “Unknown”, but after receiving feedback decided to take it out in order to increase clarity for the annotators. That option would have had an additional text box in order to give some reasoning as to why it was not possible to make.  This would have been the CML for the unknown label:  <cml:radio value="unknown" label="Unknown, I am not sure"/>  <cml:text name="unknown\_reason" label="Please explain why it was impossible to determine whether pneumonia is present or not." validates="required" only-if="pneumonia\_present:[unknown]" single="true"/>  In the real world I would do an A/B test, once with the drop-downs and once with the unknown label, to see which one gives better results. |

# 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 8 test questions, half for healthy cases and half for ones with pneumonia. |
| **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? | There is probably something missing in the instructions. This particular test case might be looking at a part of the image that the annotators haven’t been instructed to look at. Or it could be a subset of instructions that need to be added. This is all just guessing though, would need to see the actual image to determine possible causes.  For this case I might try the “unknown” label and see what kind of results will come back.  I would also check again how relevant this test question is and if it is really helping the annotators in the desired way. |
| **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 definitely run my A/B tests to compare. If they still get such results, I would try improving all sections one by one. For example I would add a few ambiguous examples and test questions and see if things improve. In a later iteration I would add more visuals in the instructions, maybe even a small video (if the platform allows). If there is a lot of text, would emphasise important things/words, with bolding, different colours, font size, grouping, etc.  If there is concrete information in the feedback from the contributors, I would definitely try to implement that into an updated version of everything really: instructions, examples and test questions. The fact they found it difficult probably means that the dataset has a lot more complicated cases than the test questions. |

# 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? | I guess if it is coming from some general children screening, the bias will be leaning more towards healthy cases.  If the data comes from patients coming to hospital, then it could be opposite, having substantially more unhealthy cases. |
| **Designing for Longevity**How might you improve your data labelling job, test questions, or product in the long-term? | First learn more about pneumonia and x rays, especially about the indicators medical professionals look into. I am sure there is more to it than just a bit of cloudiness. As I learn more about the problem and there are more edge cases, I would add them in the instructions and test questions.  I would also re-balance the dataset to include a much higher proportion of edge cases and level them with the more standard ones. |