Project Proposal

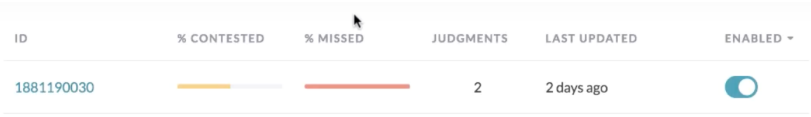
April Marie Canillo

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 key purpose of this project is to quickly distinguish between x-rays of serious cases of pneumonia in children, or healthy cases. This can also then be used as a tool that acts as a diagnostic aid for doctors. This project was designed on Appen's platform using the Image Categorization project template, in order to build a labeled dataset that distinguishes between healthy and pneumonia x-ray images. This dataset can then be used by Machine Learning engineers in the future to build a classification product. Machine Learning is a good tool to use to solve this task because an algorithm's ability to quickly learn from good data, will enable medical professionals to make decisions quickly and confidently.

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 Custom Markup Language (CML) is designed with radio buttons indicating a scale between 0 to 4. This design automatically includes room for low confidence and uncertainty; which may be observed in this dataset since we are unable to train the model on all possible data points (see Figure 1). All possible labels are shown with examples in the Instructions_Preview.html document, and also include clarifying details and reasonings. The inclusion of all cases, including ambiguous or tricky annotation cases, can ensure that the results minimize the likelihood of the contributors or model missing these details. Figure 1: Data Labelling Table 1 - Healthy Indicates a visibly health x-ray. Clear lung area, visible ribs and heart, and clear diaphragm shadow. 2 - Likely healthy, definitely not pneumonia The x-ray is mostly clear however there is a certain area of the x-ray that indicates it's not exactly healthy. The image definitely does not show normal characteristics of pneumonia. 3 - Likely pneumonia, definitely not healthy Some visibility of organs, however thick clouds that obscure certain areas. Definitely not typical of healthy x-rays. 4 - Pneumonia Clearly distinct cloudy areas that obscure the lungs, and/or several small cloudy areas, and/or little (if any) visible diaphragm shadow. 0 - Cannot Determine Visual features on the x-ray that are characteristic of typical pneumonia or healthy patients, therefore cannot be determined.

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 were 117 x-ray images in this data sample, and 7% were used as test questions. The Answer Distribution was distributed evenly across all nine test questions to avoid biasing contributors (see Figure 2). The question and reasoning quality is also defined to ensure quality from the contributors. Figure 2: Answer Distribution *On a scale of 1 (definitely healthy lungs) to 4 (definitely pneumonia lungs) what does this x-ray image show?* 1 - Healthy 22% 4 - Pneumonia 22% 3 - Likely pneumonia, definitely not healthy 22% 2 - Likely healthy, definitely not pneumonia 22% 0 - Cannot determine 11%
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?	 It is important to identify where the contributors had misunderstood the job or failed according to the statistics, in order to update the instructions, update the design, or create more test questions. Results like these further emphasize the need for continual auditing of results. I would review the failed question to see if there are any details that can be added, and use the same question as part of the Examples page in order to better prepare the contributors.

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.)	 The problem appears to stem from unclear instructions and possibly the weak connection between the instructions and the test questions. In this case, I would add more details in the instructions and ensure the examples include all possible test questions that the contributor will come across. Thereafter I would observe the contributor satisfaction results for changes, and will continually adjust either the Instructions, Examples or Test Questions, depending on these results.

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?	This model is region-based, and at this time not designed for expansion. However, the project outline can be similarly replicated across different regions and made to adopt each region's nuances. Though the distribution of the answers to the Test Questions were designed to reduce bias, the questions were selected at random and therefore may limit the scope of complexity of the project. There may be greater likelihood of success for the contributors and the results if the Product Manager had manually selected and created the appropriate Test Questions and answers, rather than selecting these at random. Additionally, the Test Questions' answers were provided by the Product Manager, rather than the medical professional. Therefore there is a bias towards reading the data through untrained eyes. To ensure that the results of the data are situated toward solving user problems, the answers should be provided by the user (medical professionals).
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	The systematic approach of this project ensures that there are plans for longevity. The ever-evolving data being introduced from the medical industry means that this dynamic model will need to be continuously trained on new data in order to stay relevant.

Additional areas for improvement & involvement of stakeholders	The radio scale could be improved as more complex data is added. In this case, rather than a 0 to 4 scale, a 1-10 scale may be more appropriate, with more detailed descriptors with percentages such as “20% of lungs covered with pneumonia” or “80% chance of pneumonia based on ____”. A more detailed scale will allow doctors to combine their experience with the data provided by the model, in order to make their best quick judgements. • In this case, user interviews with medical professionals would be necessary, in order to ensure that the labeling of the data is as accurate as possible. • Prototype usability testing with medical professionals will also be mandatory to ensure that the product is simple and intuitive enough to not hinder their performance - particularly in high-stress environments / situations. Engineers would need to be informed of results from the usability testing of medical professionals. For example, they may need to be tasked with ensuring the product includes a hands-off functionality, and the ability to add user notes and update information live. Designers would be tasked with a minimalist approach to the design, for example to ensure the user the navigation time between login and x-ray information is less than 5 seconds, and for ease of use for the medical professionals’ use. Designers would need to be closely updated by the UX Foundational Research team to ensure that the product will integrate well into the medical environment.

How can the Figure Eight platform be improved?	1. **Edit / Delete Test Questions** An option to edit or delete the test questions should be an implemented feature under the "Quality" tab. Unfortunately as I was updating the "Design" tab, I noticed that when x-ray images were placed next to each other on a grid, I was able to more definitively distinguish whether or not the image shown was of a child with pneumonia or healthy; thereby changing the results of my original answers to the test questions. This then prompted me to want to review the test questions for edit or deletion, however I could not find a visible option for this, and had to settle on launching with outdated answers. 2. **Auto-fill Design Examples using Quality data** It would be more time-efficient to include the option to Auto-Fill the Examples portion of the Design tab with the Test Questions and Answers. Normally, I would need to answer the test questions and remember to take screenshots and record my answers for each one. Once I'm satisfied with the number of test questions, I would save, navigate back to the Design Tab, and fill the Examples portion with the screenshots and notes I had established. This back-tracking slows down the process of completing the launch. 3. **Option to manually add Test Questions** Rather than having the system randomly select Test Questions to answers, the option for the Appen Client to manually add images, questions, and answers (which they had personally selected before-hand) would be less time consuming than the current process of skipping through data images until the appropriate example is found.