Project Proposal

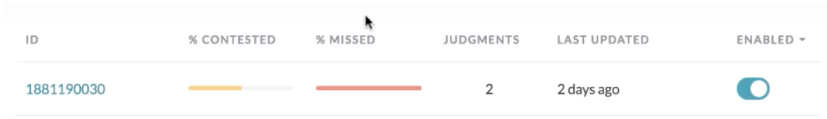
pneumonia in children x-rays – medical data annotation

Block Marie-Lynne

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	Pneumonia is a leading cause of morbidity and mortality worldwide, especially in children. Early detection can lead to timely treatment and better patient outcomes. However, the interpretation of chest X-ray images requires specialized training and is time-consuming. By providing medical personnel support in quickly identifying pneumonia based on x-rays, they can rule out certain cases quickly, start treatments in time, and focus on more complex cases. The primary goal of this project is to facilitate the early detection of pneumonia through chest X-ray images. By utilizing machine learning, we aim to assist healthcare professionals in quickly identifying healthy cases, flagging potential pneumonia cases, and prioritizing patient care. Machine learning can automatically analyze vast amounts of X-ray images, highlighting potential pneumonia cases for further review. This not only speeds up the diagnostic process but also reduces human error, ensuring that subtle signs of pneumonia or when symptoms might be unclear are not overlooked.
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:** Indicating the presence of pneumonia in the X-ray image. **No:** Indicating a clear X-ray without signs of pneumonia. **Unknown:** For cases where the annotator is uncertain about the presence of pneumonia. The choice of these labels allows for a clear binary classification (Yes/No) for machine learning models, while also accounting for any ambiguities or uncertainties (Unknown). This ensures that uncertain cases can be flagged for further expert review, minimizing the risk of misdiagnosis.

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?	Considering the dataset's size, it is recommended to include at least 5% test questions for quality assurance. This means that for every 20 data points, there should be at least 1 test question provided. I added 9 test cases because this also allowed for a balanced distribution over all the answers, which is a strategy to avoid bias with annotators. My distribution: Yes - 44% No - 44% Unknown - 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?	 There are a few steps we can take in case of test cases being missed: **Re-Evaluate the Question:** - Ambiguity: Check if the question is ambiguous or can be interpreted in multiple ways. - Complexity: Check if the question is too complex. **Review the Image/Data:** - Clarity: Ensure the image or data accompanying the question is of high quality and clear. - Relevance: Make sure the image/data is relevant to the question asked. **Re-Examine Instructions:** - Consistency: Ensure that the test question is consistent with the instructions provided. - Examples: Consider providing a similar example in the instruction section to guide annotators.

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.)

There are some areas we can look to improve:

Instructions themselves:

- Simplicity: Ensure that the language used is simple and straightforward.
  (for example only using jargon or technical terms unless absolutely necessary)
- Step-by-Step: breaking down the instructions into clear, step-by-step guidelines.

Examples:

- Relevance: Ensure the examples are directly relevant to the task at hand.
- Diversity: Include a diverse set of examples that cover various scenarios the annotators might encounter. This can help them understand edge cases.
- Annotations: For each example, provide clear annotations or explanations. Highlight why a particular label was chosen or an action was taken.

Test Questions:

- Alignment: Ensure that the test questions align well with the instructions and examples provided.
- Feedback: Provide feedback on test questions, especially if they are missed, so annotators understand their mistakes.
- Complexity: Are test questions ambiguous or complex. Simplify or clarify if necessary.

Rules & Tips:

- Highlight: Emphasize the most critical rules or common pitfalls to avoid.
- Readable lay-out: bullet points or numbered lists
- Examples: Provide examples for each rule or tip, if possible, to give a practical context.

	Visual Aids: - Diagrams & Images: If possible, include diagrams, flowcharts, or annotated images to visually explain complex concepts. - Pictures: add pictures to illustrate the cases - Formatting: Use bold, italics, and underlines judiciously to emphasize key points. For complex tasks it can be good to first run a small pilot with a few annotators to avoid issues like this and adjust before launching the full 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 dataset provided is very small and certainly not large enough for a machine learning algorithm to pick up patterns and edge cases. However, for medical imaging task such as pneumonia detection in chest X-ray images, several potential biases and considerations come to mind: **Demographic Biases**: If the X-ray images are predominantly from a particular demographic (e.g., a specific age group, gender, or ethnicity), the model might not perform as well on a diverse set of patients. ***Improvement:*** Acquire X-ray images from diverse demographics to ensure the model's applicability to a broader patient population. **Equipment Biases**: X-ray machines from different manufacturers or different models can produce images with slight variations, quality might differ. ***Improvement:*** Incorporate X-ray images taken from various machines and settings. Ensure that the dataset contains images with different resolutions, contrasts, and exposure settings. **Geographical Biases**: Pneumonia's appearance might vary depending on the region or country due to factors like prevalent strains of pathogens or common co-morbidities. ***Improvement:*** Ensure the dataset includes X-ray images from various geographic regions. **Annotation Biases**: If the data was labeled by a small group of radiologists or even a single individual, their subjective judgments might introduce biases. ***Improvement:*** Have multiple experts and other people (like the job we created) annotate the images and consider using consensus or majority voting for the final label or another way to make decisions based on multiple inputs.

	Size of the Dataset: A small dataset might not capture the variability and nuances in X-ray images related to pneumonia, as said above, this dataset certainly isn't large enough ***Improvement:*** Increase the dataset size. More data often leads to better generalization by the model. **Severity of Cases**: If the dataset primarily consists of extreme cases (either very mild or very severe), it might not be representative of the general patient population. ***Improvement:*** Ensure a balanced representation of mild, moderate, and severe cases of pneumonia. **Co-morbidities and Other Diseases**: X-ray images of patients with other lung-related diseases or conditions might be included, which can introduce noise. ***Improvement:*** Clearly label or categorize images with co-morbidities and consider their influence when training the model. **Time Period**: If the dataset consists of older X-ray images, it might not represent current medical conditions or the quality of modern X-ray machines. ***Improvement:*** Regularly update the dataset with recent X-ray images. In conclusion, understanding the potential biases in your dataset is crucial for building a robust and generalizable model. Regularly reviewing and updating the dataset and considering the aforementioned points can help in improving the data quality and the resulting model's performance. Other steps that could be considered to keep your data representative for the task is for example taking into account a Feedback-loop when the data is used/integrated and corrections are done by medical personnel.

Designing for Longevity

How might you improve your data labeling job, test questions, or product in the long-term?

In essence, the key to longevity is continuous monitoring, feedback, and iterative improvement. Regularly revisiting and updating the data labeling job ensures it remains effective and efficient in the face of evolving technology, user needs, and in this case also medical knowledge.

A few things that are important to provide longevity for data:

Regularly Update the Dataset:

- New Data Sources: in this case, one of the most obvious is, as medical imaging technology advances, incorporate newer images from state-of-the-art equipment.
- Diverse Data: Continually diversify the dataset with images from different demographics, geographical regions, and equipment (see above).

Iterative Feedback Loop:
as also said above, a feedback loop can provide a good solution for data quality and longevity. You can have 2 different types of feedback loops:

- Annotator Feedback: Regularly collect feedback from annotators on the clarity of instructions, test questions, and any challenges they face.
- Model Feedback: As your ML model gets trained and deployed, gather feedback on its predictions to refine the labeling job further.

Periodic Review of Test Questions:

- Relevance: Ensure that test questions remain relevant and aligned with the latest medical knowledge in this case.
- Performance Metrics: Monitor the accuracy and confidence scores of test questions and refine those that consistently pose challenges.

Quality Assurance Mechanisms:

- **Multiple Annotations**: For ambiguous images, get them annotated by multiple experts and use consensus or majority voting, also mentioned above to ensure data quality.

Stakeholder Engagement:

- Medical Expertise: Engage with medical professionals to ensure the labeling job aligns with medical best practices.
- User Feedback: If the end product is an application or tool for doctors, gather user feedback for continuous improvement.