

# Medical Image Annotation: Image Classification Using Appen's Figure-Eight Platform



This is a joint project between Udacity University and Appen.



Study & analysis conducted by: Frederick Zoreta

# Image Detection & Classification: Finding Pneumonia

Frederick Andaya Zoreta



## Data Labeling Approach

<b>Project Overview and Goal</b>  What is the industry problem you are trying to solve? Why use ML in solving this task?	Udacity's R&D team is thoroughly involved in a Machine-Learning based project. The medical imaging industry is the main audience. Data sets that involves responses from a wide variety of target audience was given. ML was utilized in order to improve the chances of determining the presence of pneumonia in a given data set of x-ray images.
<b>Choice of Data Labels</b>  What labels did you decide to add to your data? And why did you decide on these labels vs any other option?	The chosen data set mainly contains 3 images of chest X-rays. These images are as follows: <ol style="list-style-type: none"><li>1. Normal</li><li>2. Bacterial Pneumonia</li><li>3. Viral Pneumonia</li></ol> Chosen data set is deemed viable for a wide target audience that represents 'every day, ordinary people'. Majority of the population are not too familiar in determining pneumonia. The 2 types pneumonia were included in order to make the dataset more realistic.

## Test Questions & Quality Assurance

<p><b>Number of Test Questions</b></p> <p>Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?</p>	<p>This initial survey consisted of only 12 questions. The rationale for this are the following:</p> <ol style="list-style-type: none"> <li>1. Looking X-ray images may confuse the audience</li> <li>2. X-ray images are not known to be 'pleasant to look at'</li> <li>3. Majority of the public are not experts in classifying such images</li> </ol> <p>** It is highly important to note that this data annotation serves as a 'launching pad' for further studies.</p>
<p><b>Improving a Test Question</b></p> <p>Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?</p>	<p>*It would be very vital to make the statements simpler. The simpler, the better it is for the audience.</p> <p>* The use of another survey platform would also be an option. The over-all 'feel' or Usability of the platform's outcome is a HUGE ISSUE &amp; CONCERN for us researchers.</p> <p>* It is highly recommended to look into other similar studies and observe the types of KPIs, metrics and even the analytics models that they utilized.</p>
<p><b>Contributor Satisfaction</b></p> <p>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.)</p>	<p>I would highly focus on both the examples and test questions.</p> <p>Why examples? The target audience may several biases and even be confused regarding x-ray images. A large part of the audience simply reported 'no fuzzy images' or 'no smoke like images', and they automatically marked it as normal.</p> <p>Having 2 types of pneumonia X-rays is good. It is a great way to make the survey close to reality.</p>

# Limitations & Improvements

Additional photos that involves outside source:

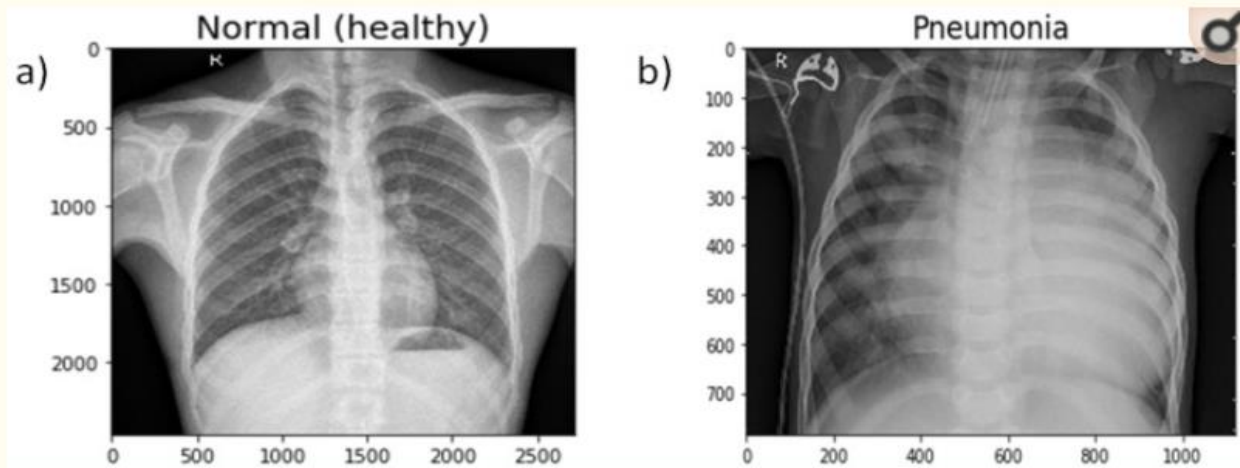


Figure 2

Chest Xray of (a) a healthy person and (b) a person suffering from pneumonia.

The images above yields the result of an extensive research that uses 'Data Augmentation' in order to tackle the issue of a limited dataset. This study mentions that deep learning models were also used in order to fine tune pneumonia classification survey.

Table 1

Description of the experimental dataset.

Category	Training Set	Test Set
Normal (Healthy)	1283	300
Pneumonia (Viral + Bacteria)	3873	400
Total	5156	700
Percentage	88.05%	11.95%

Above image and table were taken from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7345724/>

<p><b>Data Source</b></p> <p>Consider the size and source of your data; what biases are built into the data and how might the data be improved?</p>	<p>The main bias is x rays being 'fuzzy' or having some 'blurry' figures. Although labeling these photos as pneumonia is correct, it somewhat gives the audience an 'unfair' advantage of some sort.</p> <p>It is also common for the audience to simply answer 'unknown'. It appears to be an acceptable response based on the fact that majority of the audience are not experts on this field.</p>
<p><b>Designing for Longevity</b></p> <p>How might you improve your data labeling job, test questions, or product in the long-term?</p>	<p>I was a former digital marketing analyst. We were ALWAYS using A/B testing as a means of using 2 types of data and research design. After a certain period of time, whichever among the 2 types of methods, or in digital marketing realm 'ADS' yields higher conversion and activity would be chosen.</p>

\*\*The image below shows several x-rays that has various respiratory conditions. Although or specific study is a lot simpler, it should be noted that certain images may appear as 'the same' to certain audiences. From my personal perspective, photo b (COVID-19) and photo c (SARS) appears the same.

\*\*A vital part of improving this over-all data driven survey is to also consider several factors:

- The over-all usability or 'feel' of the survey. This could range from the accuracy & simplicity of the statements, the clarity of the x-ray images shown and 'ease of use' of the platform being used.
- It could be a possibility that having 2 types of image classification tests ( as mentioned above using 'split tests or A/B tests') may give a good comparison among 2 types of surveys.

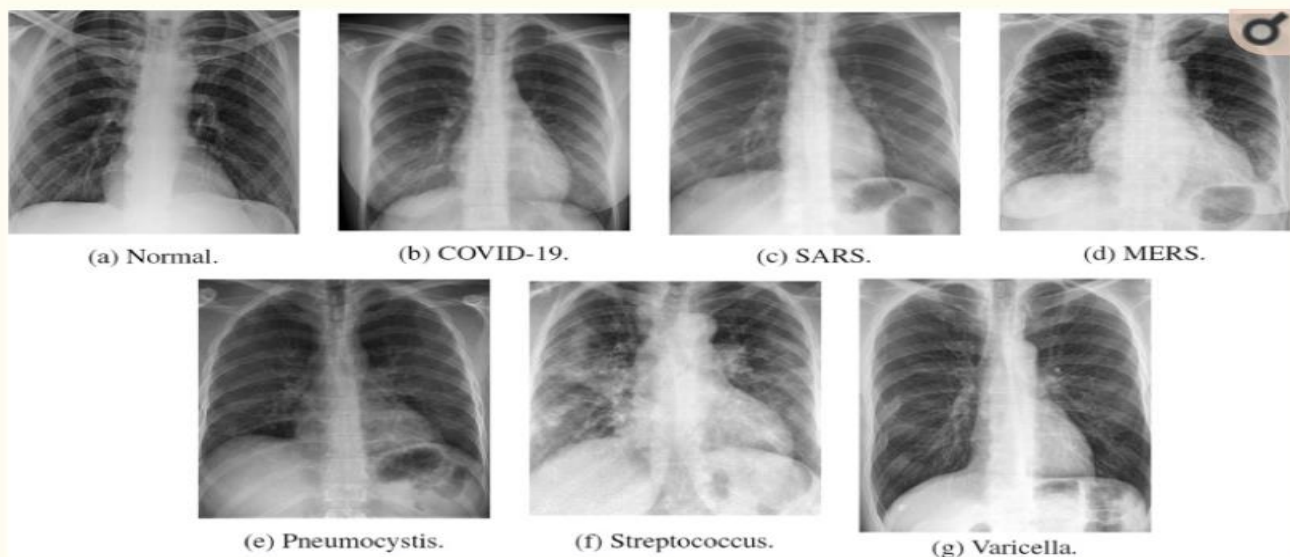


Fig. 6

RYDLS-20 image samples.

Source of image above: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7207172/>

## Additional Observations Regarding Limitations


- a. Design labels for all scenarios; the best annotations and test questions should be designed to handle failure cases.

The questionnaire created with the premise of using it as a ‘test scenario’ or a ‘warm up’ for a more defined and organized survey. All the possible corrections and learning points from this survey could be applied to the next version.




That specific version would have more defined questions that would be more in-line with the over-all goal of this project.

### Answer Distribution (10 Test Questions)

Cover all answers and roughly approximate your dataset's answer distribution to avoid biasing contributors. [Learn more](#)

 We recommend every answer to be represented at least once in your first 8 Test Questions. See below for missing answer(s).

#### Does this xray indicate pneumonia ?

Yes		40%
No		30%
Unknown		30%

#### How confident are you with your assessment?

High		60%
Low		40%

The image above is the ‘actual’ result of my test scenario. This clearly indicates that I may need to also increase the number of questions in the survey.

## b. The Figure Eight Platform

As an AI Product Manager, my main issue with the study is the actual platform itself. Although the platform appears to be 'very intuitive' and is considered a 'no-code' platform, the lack of proper, detailed & organized video tutorials makes it difficult.

Other than the actual understanding of ML & AI fundamentals, the 'backbone' of this project lies in the over-all usability of the platform.

The link below shows the platform's video tutorials. The specific scenario given is totally different from this survey. There was no image classification involved among given tutorials. Even their YouTube channel does not include the actual editing of the sample photos and changing it with the given dataset.

<https://success.appen.com/hc/en-us/sections/201955376-Video-Tutorials>

As a product manager, I am also employing my over-all knowledge of UX design and research. The platform appears very intuitive, but the lack of quality video tutorials is a huge factor in the outcome of this initial research.

### Appendix:

1. Efficiency Pneumonia Detection in Chest Xray Images Using Deep Transfer Learning  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7345724/>
2. COVID-19 Identification in Chest-Xray Images on Flat & Hierarchical Classification Scenarios  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7207172/>
3. The Art of A/B Testing: Walk through the beautiful math of statistical significance  
<https://towardsdatascience.com/the-art-of-a-b-testing-5a10c9bb70a4>
4. Complementing A B Testing with Machine Learning & Feature Importance  
<https://towardsdatascience.com/complementing-a-b-testing-with-machine-learning-6c5c92baa162>
5. Best Use of Train/Val/Test Splits for Medical Data  
<https://glassboxmedicine.com/2019/09/15/best-use-of-train-val-test-splits-with-tips-for-medical-data/>
6. Computer-Aided Detection in Chest Radiography Based on Artificial Intelligence: A Survey  
<https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-018-0544-y>