|  |  |
| --- | --- |
| Capstone Project Proposal |  |

*<Erkan Hatipoglu>*

**Business Goals**

|  |  |
| --- | --- |
| **Project Overview and Goal**  What is the industry problem you are trying to solve? Why use ML/AI in solving this task? Be as specific as you can when describing how ML/AI can provide value. For example, if you’re labeling images, how will this help the business? | To be consistent with the previous projects a project about COVID-19 pneumonia detection has been selected.  The goal of this project is to make a product for doctors that distinguishes between healthy, pneumonia, and COVID-19 pneumonia chest x-ray images. The feasibility of such a work has been studied in various articles1.  The AI product will have a simple user interface. The user will be able to upload a chest x-ray image and receive an output indicating the diagnosis such as *COVID-19 Pneumonia*, *Normal Pneumonia*, or *No Pneumonia*.  It is known that the False Negative ratio for COVID-19 rapid tests2 and PCR tests3 are very high (False Positive ratio, on the other hand, is too small). As a result, doctors rely on other tools such as known symptoms or chest x-ray images to detect COVID-19. However, doctors don't have the means (yet) to determine a pneumonia image indicates a regular pneumonia or a COVID-19 pneumonia. As a result, for a patient -who has symptoms of pneumonia in her/his chest x-ray image- with a negative test (PCR or rapid) result, the treatment process may be performed assuming the patient is COVID-19, or it may be delayed till another PCR test is conducted. On the other hand, because of the high FN and low FP ratio, the second PCR test doesn’t also guarantee a better decision unless the result is positive.  In the previous projects, we have seen that differentiating pneumonia chest x-ray images from normal chest x-ray images by using AI is possible. What's more, we have seen that this can   * Help flag serious cases, * Quickly identify healthy cases, * And, generally, act as a diagnostic aid for doctors.   In addition, from other studies, we know COVID-19 chest x-ray images can also be differentiated4. As a result, such a product can help doctors   * Differentiate COVID-19 Pneumonia from Normal Pneumonia, * Conduct the right treatment at the right time, * Improve patient’s happiness, * Improve doctor's credibility.   In short, ML can help us make better decisions, enhance COVID-19 treatment procedure, and increase customer satisfaction.  1 [Liu C, Wang X, Liu C, Sun Q, Peng W. Differentiating novel coronavirus pneumonia from general pneumonia based on machine learning. *Biomed Eng Online*. 2020;19(1):66. Published 2020 Aug 19. doi:10.1186/s12938-020-00809-9](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7436068/)  2 [Are Rapid COVID-19 Test Results Reliable?](https://www.healthline.com/health/how-accurate-are-rapid-covid-tests#how-accurate-is-it)  3 [Real-life clinical sensitivity of SARS-CoV-2 RT-PCR test in symptomatic patients](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251661)  4 [Nishio, M., Noguchi, S., Matsuo, H. et al. Automatic classification between COVID-19 pneumonia, non-COVID-19 pneumonia, and the healthy on chest X-ray image: combination of data augmentation methods. Sci Rep 10, 17532 (2020). https://doi.org/10.1038/s41598-020-74539-2](https://www.nature.com/articles/s41598-020-74539-2) |
| **Business Case**  Why is this an important problem to solve? Make a case for building this product in terms of its impact on recurring revenue, market share, customer happiness, and/or other drivers of business success. | COVID-19 treatment is so stiff and has severe side effects. Differentiating normal pneumonia from COVID-19 pneumonia may change medication used and, normal pneumonia patients may get a lighter treatment and avoid unnecessary side effects.  In addition, wasted time that results from multiple PCR tests caused by negative results will be saved. This wasted time may cause irreversible damages to the patients.  So, the product will automate and speed up the COVID-19 diagnostics, improve patient’s happiness, and doctor's credibility. |
| **Application of ML/AI**  What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve? | As described above the ML model will be a multi class classifier that distinguishes *No Pneumonia, Normal Pneumonia, and COVID-19 Pneumonia* from chest x-ray images*.*  The output will be monitored using recall since we want to decrease FP and increase TP.  User satisfaction and the decision process will be monitored as outcome. |

**Success Metrics**

|  |  |
| --- | --- |
| **Success Metrics**  What business metrics will you apply to determine the success of your product? Good metrics are clearly defined and easily measurable. Specify how you will establish a baseline value to provide a point of comparison. | Since we will monitor user satisfaction and the decision process as outcome, we need to apply related success metrics. We will use the following metrics:   * Customer satisfaction   Customer surveys will be made before and after using the product. Customer surveys made before will be the baseline.   * Decision Process   Number of normal pneumonia cases diagnosed by the product will be used as a success metric. This number is very important since it directly indicates the usefulness of the product. A very small number points out the product as useless. Average pneumonia cases before the pandemic will be used as a baseline. This data will be gathered from the customer.  The difference of diagnosis time calculated with and without using the product will be another success metric. The data will be collected from the customer and the baseline will be zero. |

**Data**

|  |  |
| --- | --- |
| **Data Acquisition**  Where will you source your data from? What is the cost to acquire these data? Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome? Will data become available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed? | For the initial product, data will be collected from the open-source resources in the WEB including the dataset we have used in the first and second projects of this Nanodegree. Below you can find the links of some open-source datasets:   * [Cohen, J. P., Morrison, P. & Dao, L. COVID-19 image data collection (2020)](https://github.com/ieee8023/covid-chestxray-dataset) * [RSNA Pneumonia Detection Challenge - Kaggle](https://www.kaggle.com/c/rsna-pneumonia-detection-challenge) * [Chest X-Ray Images (Pneumonia) - Kaggle](https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia) * [COVID-19 imaging datasets](https://www.eibir.org/covid-19-imaging-datasets/)   There is no financial cost for gathering this very open-source data however, cleaning, preparing, and combining all the data into one balanced dataset will need labor.  For the initial product, since the datasets are open source, no PII or data sensitivity issues are expected.  It is possible that with every mutation of the virus the images may change form. So, it is crucial that we need to feed our model with the new data collected from the customer. This data will be gathered in small batches and the model will be retrained. While collecting those batches we also need to consider the balance of the dataset. The number of samples for each class need to be as equal as possible.  Since we will be using the data coming from the patients and the customer, we need to get the necessary permissions.  There will be no financial cost for gathering data from customer since they will be informed that this data will used for updating the product. As before, we will need labor for cleaning, preparing, and combining all the data into one balanced dataset. |
| **Data Source**  Consider the size and source of your data; what biases are built into the data and how might the data be improved? | We have several sources of data. The sizes of these datasets will be given below:   * [Cohen, J. P., Morrison, P. & Dao, L. COVID-19 image data collection (2020)](https://github.com/ieee8023/covid-chestxray-dataset)   Covid-19 Pneumonia: 468  Normal Pneumonia: 178  No Pneumonia: 0   * [RSNA Pneumonia Detection Challenge - Kaggle](https://www.kaggle.com/c/rsna-pneumonia-detection-challenge)   Covid-19 Pneumonia: 0  Normal Pneumonia: 9555  No Pneumonia: 20672   * [Chest X-Ray Images (Pneumonia) - Kaggle](https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia)   Covid-19 Pneumonia: 0  Normal Pneumonia: 3875  No Pneumonia: 1341   * [COVID-19 imaging datasets](https://www.eibir.org/covid-19-imaging-datasets/)   Covid-19 Pneumonia: 3127 (Patients)  Normal Pneumonia: 0  No Pneumonia: 0  In total:  Covid-19 Pneumonia: 3595  Normal Pneumonia: 13430  No Pneumonia: 22013  It can be easily seen that the combined dataset will be unbalanced which will introduce bias. So, we need to apply techniques to work with this imbalanced data.  Another bias may be introduced because of multisource variability1. Depending on the data collected from the customers in later stages may decrease this bias.  There may also be gender bias2 which may be reduced by diversity.  There may also be annotation bias that comes from humans annotating and generating those datasets above. We may reannotate some portion of data if needed to solve annotation bias.  1 [O. D. T. Catalá et al., "Bias Analysis on Public X-Ray Image Datasets of Pneumonia and COVID-19 Patients," in IEEE Access, vol. 9, pp. 42370-42383, 2021, doi: 10.1109/ACCESS.2021.3065456.](https://ieeexplore.ieee.org/document/9374968)  [2](https://www.pnas.org/content/117/23/12592) [Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis](https://www.pnas.org/content/117/23/12592)  [AgostinaJ. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, Enzo Ferrante](https://www.pnas.org/content/117/23/12592)  [Proceedings of the National Academy of Sciences Jun 2020, 117 (23) 12592-12594; DOI: 10.1073/pnas.1919012117](https://www.pnas.org/content/117/23/12592) |
| **Choice of Data Labels**  What labels did you decide to add to your data? And why did you decide on these labels versus any other option? | Since our goal is to differentiate COVID-19 pneumonia from normal pneumonia and there may be healthy cases also we will have 3 labels as described below.  The labels will be ***COVID-19 Pneumonia***indicating the patient has COVID-19, ***Normal Pneumonia***indicating the patient has pneumonia but no COVID 19, and ***No Pneumonia***indicating the patient has neither pneumonia nor COVID-19.  Normal pneumonia can also be classified as viral and bacterial pneumonia as we have seen in project 2. Besides bacterial and viral pneumonia can also be classified into subclasses (In fact COVID-19 pneumonia is a viral pneumonia itself). However, we do not need that complexity for our goal. Normal pneumonia label is enough for us. |

**Model**

|  |  |
| --- | --- |
| **Model Building**  How will you resource building the model that you need? Will you outsource model training and/or hosting to an external platform, or will you build the model using an in-house team, and why? | We will build the model using an in-house team since we have already an ML team capable of doing the project. We have also got enough time budget.  The training and storage of the model will be accomplished using the cloud, probably on MS Azure since our team has experience using Azure ML.  The deployment will also be on Azure. We will also use Application Insights. Application Insights is a special Azure service that provides key facts about an application. It is a very useful tool to detect anomalies and visualize performance.  The product will get the data from the WEB and consume the deployed service via an HTTP API. |
| **Evaluating Results**  Which model performance metrics are appropriate to measure the success of your model? What level of performance is required? | As stated earlier, the output will be monitored using ***recall*** since we want to decrease FP and increase TP.  It is known that False Positive ratio for COVID-19 rapid tests is between 45.1% and 55.4 %1. If we target at least a recall of 54.9 % we will beat COVID-19 rapid tests.  On the other hand, we know that by using ML we can have a sensitivity of 88.62%2. But considering that we may not want the best model for deployment, selecting a slightly less value for recall like ***85%*** will be more reasonable and attainable.  1 [Are Rapid COVID-19 Test Results Reliable?](https://www.healthline.com/health/how-accurate-are-rapid-covid-tests#how-accurate-is-it)  2 [Liu C, Wang X, Liu C, Sun Q, Peng W. Differentiating novel coronavirus pneumonia from general pneumonia based on machine learning. *Biomed Eng Online*. 2020;19(1):66. Published 2020 Aug 19. doi:10.1186/s12938-020-00809-9](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7436068/) |

**Minimum Viable Product (MVP)**

|  |  |
| --- | --- |
| **Design**  What does your minimum viable product look like? Include sketches of your product. |  |
| **Use Cases**  What persona are you designing for? Can you describe the major epic-level use cases your product addresses? How will users access this product? |  |
| **Roll-out**  How will this be adopted? What does the go-to-market plan look like? |  |

**Post-MVP-Deployment**

|  |  |
| --- | --- |
| **Designing for Longevity**  How might you improve your product in the long-term? How might real-world data be different from the training data? How will your product learn from new data? How might you employ A/B testing to improve your product? |  |
| **Monitor Bias**  How do you plan to monitor or mitigate unwanted bias in your model? |  |