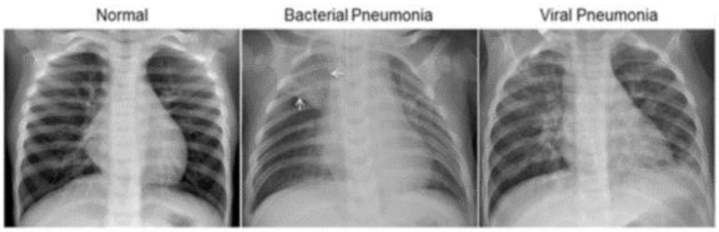
Capstone Project Proposal

Eda AYDIN

Business Goals

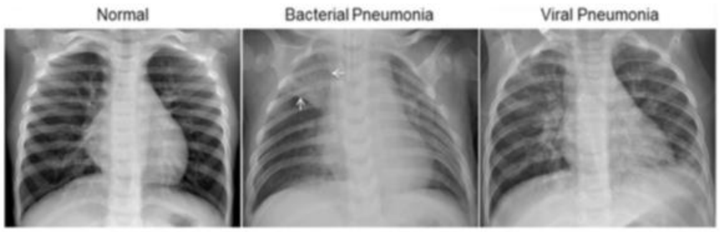
Project Overview and Goal What is the industry problem you are trying to solve? Why use ML/AI in solving this task? Be as specific as you can when describing how ML/AI can provide value. For example, if you're labeling images, how will this help the business?	Any bacteria or virus that enters the body must be found in a short time so that the diagnosis and treatment process of the disease can be found as soon as possible. The most important of these is that the correct diagnosis takes time. There are different methods for the diagnosis of each virus and bacteria. Blood tests, X-ray images, etc. Different types of viruses have shown different symptoms on the human body until today. A virus such as Ebola can be detected externally, such as high fever, headache, and vomiting. Cholera virus, on the other hand, can be detected after a long time, sometimes without any symptoms, only by giving symptoms such as vomiting and diarrhea. The Covid-19 virus, which is the most important we live in now and produces new variations against the vaccine, can be detected with the help of X-ray images. In this project, the diagnosis of Pneumonia, which is seen as one of the most important permanent effects of the Covid-19 virus, will be detected with the help of X-ray images.  At this stage, there are 3 different categories in determining the images: Normal, Bacterial Pneumonia, Viral Pneumonia.

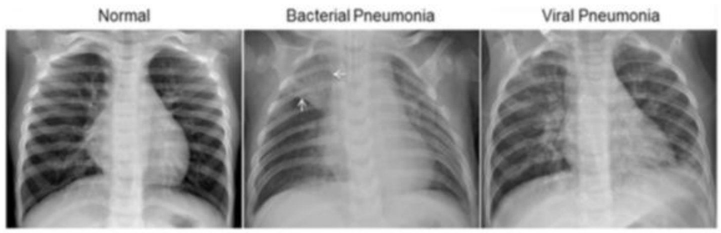
	Separating images in 3 different categories will help us determine what caused pneumonia.
Business Case Why is this an important problem to solve? Make a case for building this product in terms of its impact on recurring revenue, market share, customer happiness and/or other drivers of business success.	Shortening the diagnosis and treatment time in the field of health is one of the most important tasks of artificial intelligence. Diagnosing whether the pneumonia is caused by bacteria or viruses in a short time will help eliminate the 50% probability. Another reason is that the causes of pneumonia are an area that will actively progress even after many years due to the increasing genetically modified diet, different variations of viruses and bacteria.
Application of ML/AI What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve?	Deep Learning, Convolutional Neural Network and PyTorch will be used. Data Augmentation will be used to produce images to ensure a high level of accuracy and because the existing images are not sufficient for some situations.

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.	The business metrics that I’m applying here is the pneumonia correct predictions and pneumonia errors. These two metrics can be clearly defined and easily measured by using optimization. The baseline value is established with help of the accuracy of the pneumonia prediction, compared with the other competitive applications in the market.

Data

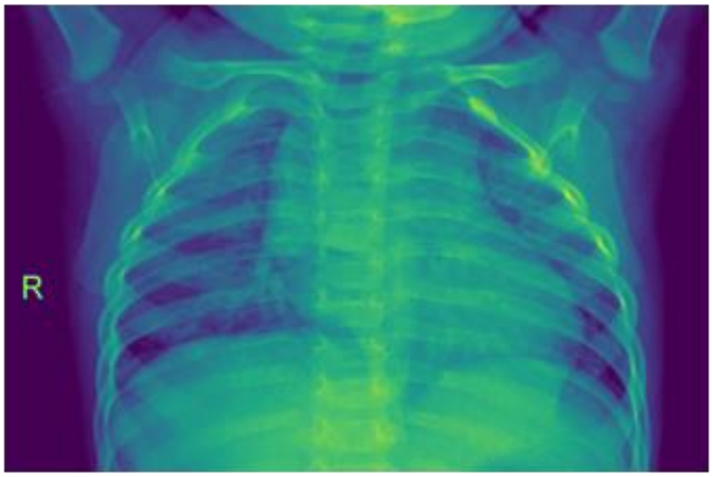
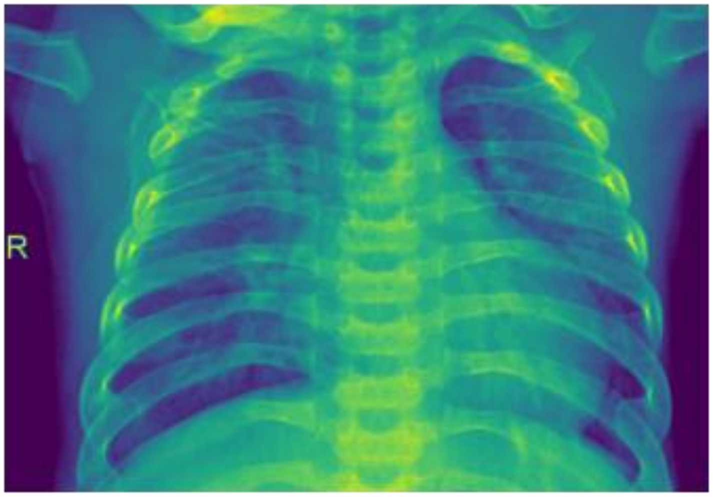
Data Acquisition Where will you source your data from? What is the cost to acquire these data? Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome? Will data become available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed?	Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. All chest X-ray imaging was performed as part of patients' routine clinical care. For the analysis of chest x-ray images, all chest radiographs were initially screened for quality control by removing all low quality or unreadable scans. The diagnoses for the images were then graded by two expert physicians before being cleared for training the AI system. In order to account for any grading errors, the evaluation set was also checked by a third expert.  Normal Bacterial Pneumonia Viral Pneumonia The normal chest X-ray(left panel) depicts clear lungs without any areas of abnormal opacification in the image. Bacterial pneumonia (middle panel) typically exhibits a focal lobar consolidation, in this case in the right upper lobe (white arrows), whereas viral pneumonia(right panel) manifests with a more diffuse "interstitial" pattern in both lungs. Always having an expert opinion in health-related applications are highly recommended. Since the data are collected within the framework of the hospital's ethical rules, the model will be established without specifying who the data belong to. So, there will be no sensitivity issues. For the diagnosis of pneumonia, a model will be established from the data shared in the places determined by the hospitals. Data should be kept up to date as new variations are constantly being released. So, the large batch of data can be acquired from licenses online resources.
Data Source Consider the size and source of	The dataset is organized into 3 folders (train, test, val) and contains subfolders of each image category (Pneumonia / Normal). There are 5,863 X-Ray images (JPEG) and 2 categories(Pneumonia/Normal)

your data; what biases are built into the data and how might the data be improved?	Because the images are remarkably similar to each other, the model can be improved by using optimization to detect only minor differences. By creating multiple layers on the images, the smallest detail can be considered. This also happens with the Convolutional Neural Network.
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?	 The normal chest X-ray(left panel) depicts clear lungs without any areas of abnormal opacification in the image. Bacterial pneumonia (middle panel) typically exhibits a focal lobar consolidation, in this case in the right upper lobe (white arrows), whereas viral pneumonia(right panel) manifests with a more diffuse "interstitial" pattern in both lungs. When data were collected, they were classified as probability of having pneumonia and probability of not having pneumonia. Then, patients with pneumonia were divided into two different classes to determine whether it was bacterial-based or virus-based. The general formation of diseases is based on two basic principles in the field of bioinformatics. It is a bacterial or virus-based formation. When the formation of some diseases is bacterial based, the diagnosis and treatment time varies accordingly. Likewise, when it is virus-based, diagnosis and treatment change accordingly. This is the main reason why I identified it as both bacterial pneumonia and viral pneumonia. It is also the opposite of this, that the patient does not have pneumonia. The reason why I did not choose the other options is that the source of pneumonia is the two points that the main causes connect. If the strength of label expressions is the duration of the treatment, we have a chance to get faster results. In other words, if it is bacterial-based, the probability of 50% viral-based will be extremely low. In this way, it will be easier for us to get fast and clear results during the treatment. The weakness is that some microorganisms, although very rarely, can be both bacterial and viral based. In this case, the system will be difficult to qualify and compare. In order to improve this step, a separate classification for these organisms may be made in the future.

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 need two things to create the model. One to a strong data and the other to a strong model. Since the raw format of the data will be obtained from the hospital, a software team must be established in the company for the strong establishment of the model. The computers that these people will have must work on the GPU base in order to set up the model. In addition, this team should include Radiologists in the hospital who are responsible for generating this data. They will also guide how the model should be created.
Evaluating Results Which model performance metrics are appropriate to measure the success of your model? What level of performance is required?	The success of the model is based on the Model Accuracy – Estimates of the Doctor. While the doctor presents a result for the diagnosis of the disease, it also draws conclusions from the model. According to the comparison of these results, the results of diagnosis satisfaction are explained. If the performance of the model should be close to the performance of the doctor so that we can get a satisfactory result. The performance metric I would use to assess the performance of the model is accuracy (Precision – the number of instances it predicts correctly. Recall – the number of relevant instances that are retrieved.(if the model is perfectly accurate and the diagnosis of disease is compatible with the prediction result of the doctor, the next stage which is the treatment of disease can be passed. Based on the Precision and Recall, the F1 score is predicted, which defines the model accuracy. The baseline value beyond which the model the model is considered a success is 94% of accuracy.

Minimum Viable Product (MVP)

Design What does your minimum viable product look like? Include sketches of your product.	The product will be managed by authorized persons, namely doctors and software developers. Patients' information will be entered by the physician's assistants. According to the patient's history, the procedures to be performed in the system (MRI, X-ray, Analysis, etc.) will be determined. The images taken as a result of these processes will be uploaded to the system by doctors (radiologists, analysts). In another step, image classification will be made according to the model created by the software developers.  The patient who has not pneumonia  The patient who has pneumonia

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?	The persona I’m designing for is the User persona, the users can be adult age group with any medical records. The major use cases would be addressing diagnosis of the pneumonia based on the medical records like blood testing, lab results, blood pressure, lung structure, etc.) In addition, there may be other underlying diseases. The users access this product via website. Radiologists : Role-based persona – They are the people who will be responsible for data entry into the system. Any patient's MRI or X-ray images must be uploaded to the system by an authorized person. Images such as MRI and X-ray taken by the radiologist are uploaded to the system. Doctors: Administrative assistant persona - After the MRI and X-ray images uploaded to the system, the doctor enters his own detection information into the system during the determination of the definitive diagnosis. At the same time in the background, the artificial intelligence model determines the accuracy result according to the X-ray images without seeing the result determined by the doctor. In this section, the detections of the artificial intelligence model are not visible to the doctor in the system. If both information is compatible with each other, the next stage, the treatment stage, is started. If there is too much difference between the two determinations at the stage of determining this diagnosis, then the results of the determination are sent to the radiologist and the initial images are taken again. That's why this is the important user-case in the system. The epic use case of the model would be lung cancer patients and asthma and bronchitis patients. The patient's history and susceptibility to diseases are extremely helpful in diagnosis. The model observes the detection of more details in the images obtained with the increase of factors that may cause deterioration of the lung structure by the patient forewarning of important diseases that may impair the normal appearance of the lung, apart from pneumonia. In this case, data is needed for multiple models, not just a single model, and the diagnosis of multiple doctors, not just a single doctor's diagnosis. Therefore, the patient's history and diseases directly affect the system and change its size.

Roll-out How will this be adopted? What does the go-to-market plan look like?	This will be adopted as a health care product and the go-to-market plan is finding the “Arterys: Medical Imaging Cloud AI”, developing the product in a unique way, and releasing it on the website. The pricing strategy of the GTM plan will be, launching it as free at the beginning, as the product gets more reach and successful, the price can be slowly increased with respect to the size of the data and features of the data. (data size and modeling can be increased in the future to diagnose other diseases not only in the pneumonia area but also in the lung area for better results) The distribution strategy would be finding marketing people all over the country to promote the application and also targeting various hospitals and clinics to get more users for the product.

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?	It may be possible by adding different diseases to improve the system. In other words, as a result of pneumonia caused by a new virus or bacteria, it must be in a separate category. Another development system is that it creates a separate system for children and babies. The lung structure of children and infants is different from the lung structure of adults. As we age, our lungs grow accordingly. Sometimes the same type of virus shows different symptoms depending on age. The system can be improved by adding these states. The A/B testing can be done until we achieve high “Performance metrics” testing against statistically significant sample size and running tests long enough to capture any seasonality effects.
Monitor Bias How do you plan to monitor or mitigate unwanted bias in your model?	The bias can be monitored by carefully annotating the new training images / datasets. Convolutional Neural Network, namely Data Augmentation and Data Loader are used for the lack of data. In addition, different kinds of optimizations can be added. (Adam optimization, early stopping)