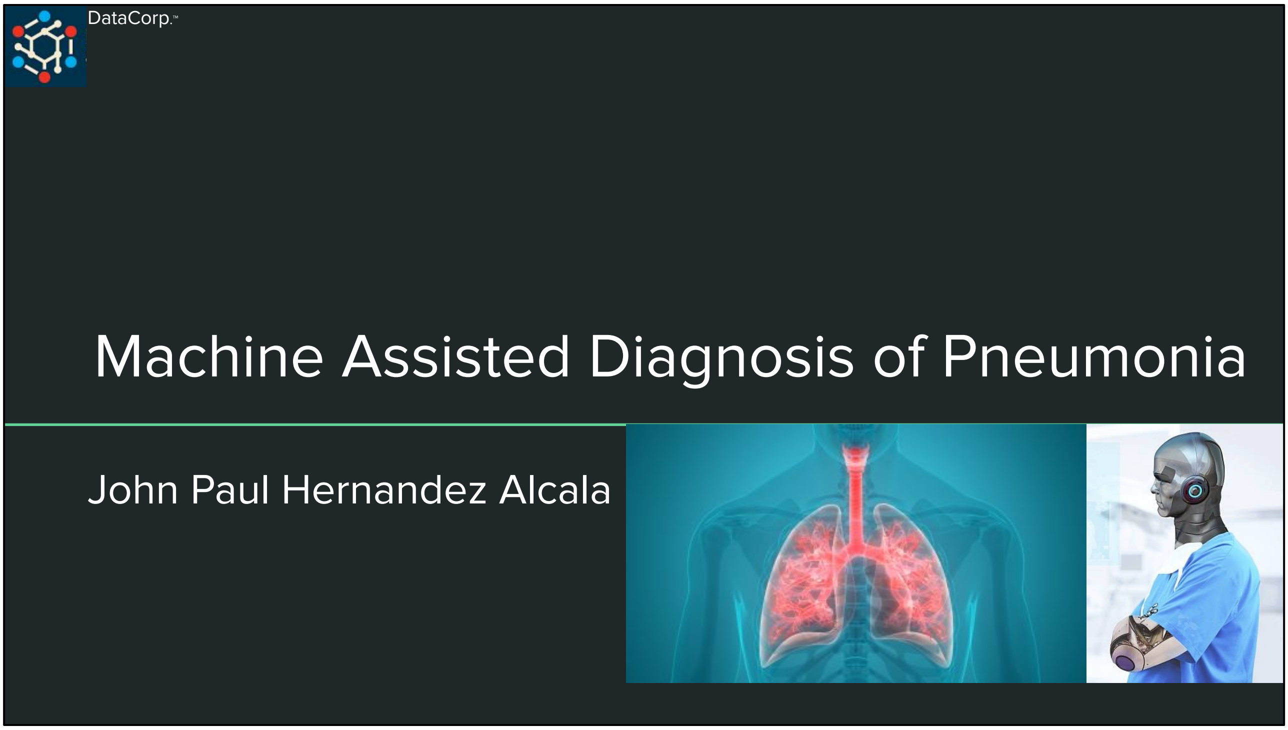


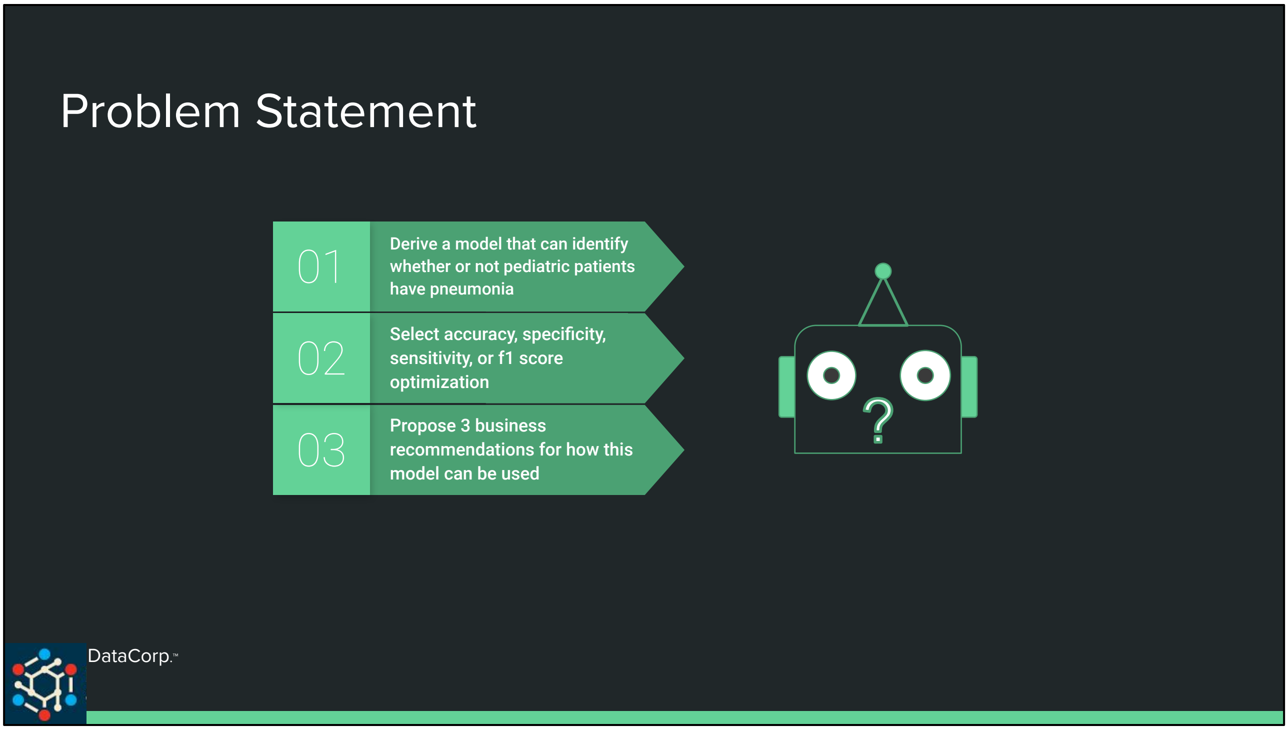
DataCorp.™

Machine Assisted Diagnosis of Pneumonia

John Paul Hernandez Alcala

Hello, my name is John Paul, and I am a data scientist with DataCorp. Today my goal is to present a model that can assist with diagnosing pneumonia. Let’s get started!

Images: https://www.the-cma.org.uk/cma_images/lungs.jpeg
https://www.google.com/url?sa=i&url=https%3A%2F%2Fmedium.com%2Fthe-research-nest%2Fwould-you-trust-your-doctor-or-an-ai-machine-2ae02dda9e80&psig=AOvVaw2hggvF9yfjeFVn1w5MaPnV&ust=1604450267801000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCLCC8q-R5ewCFQAAAAAdAAAAABAD



Problem Statement

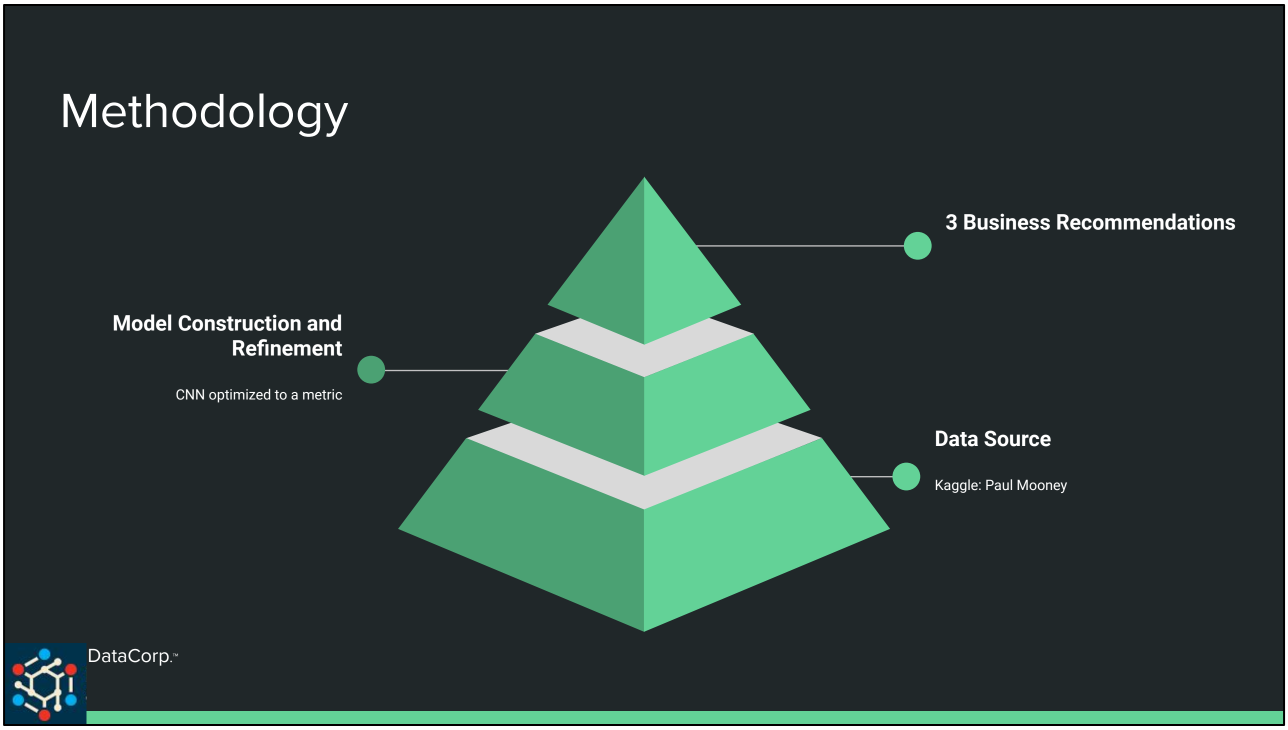
01 Derive a model that can identify whether or not pediatric patients have pneumonia

02 Select accuracy, specificity, sensitivity, or f1 score optimization

03 Propose 3 business recommendations for how this model can be used

DataCorp.™

Before we dive into this, we have to present ourselves with three tasks that make up our problem statement: derive a models that can predict pneumonia, figure out why high accuracy is not always the be all metric, and how our model can change modern medicine



Methodology

3 Business Recommendations

Model Construction and Refinement

CNN optimized to a metric

Data Source

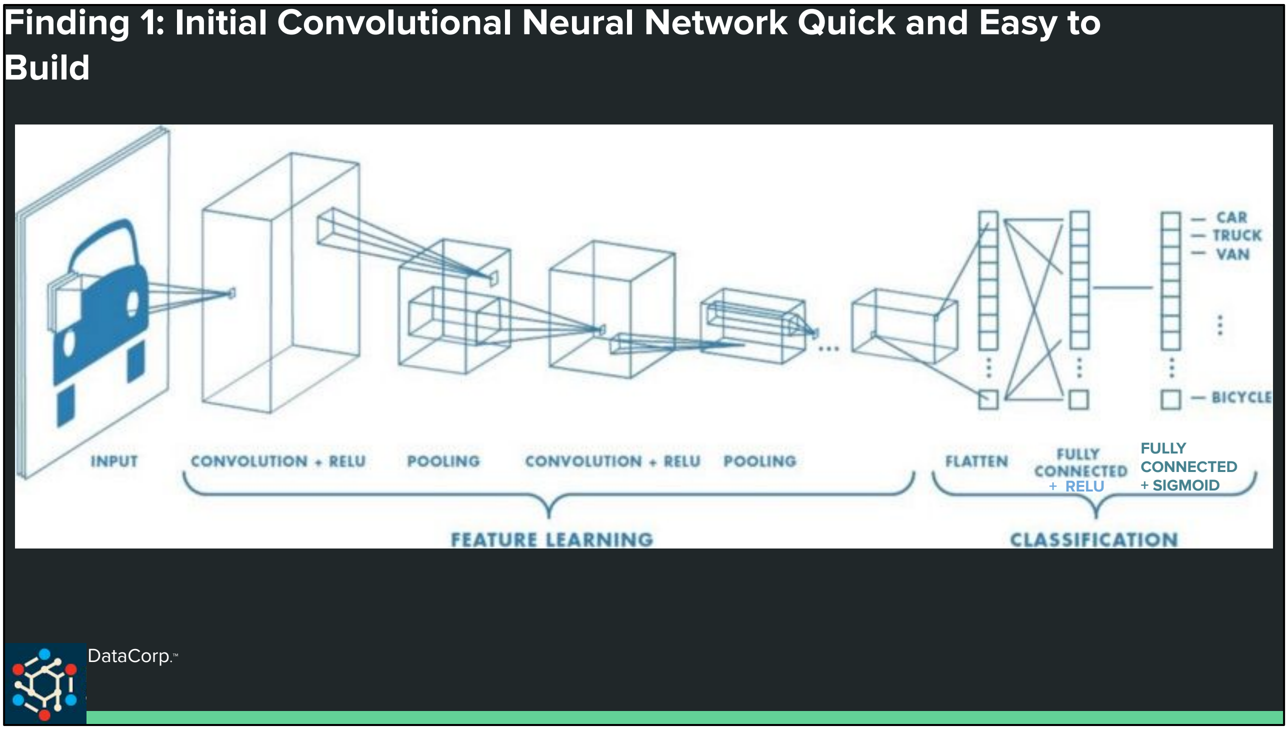
Kaggle: Paul Mooney

DataCorp™

We accomplish these tasks, so we can get a comprehensive idea what Convolutional Neural Networks can do in medicine. For the 1st task, we will derive a model that is better than a random guess which is 50% accuracy.

For the 2nd task, we will determine why one metric would be better to optimize over others. Finally, based on the results of our final model and analysis, we can propose 3 business recommendations for how this model can be used

To start on our journey of completing our tasks, we first start off by collecting data from Kaggle, work our way up with building a CNN model and refining it. Finally, we will arrive to our three business recommendations.



Finding 1: Initial Convolutional Neural Network Quick and Easy to Build

DataCorp.™

Image: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Convolution layer or Kernel/Filter reduces dimensionality of our input image while extracting the high-level features

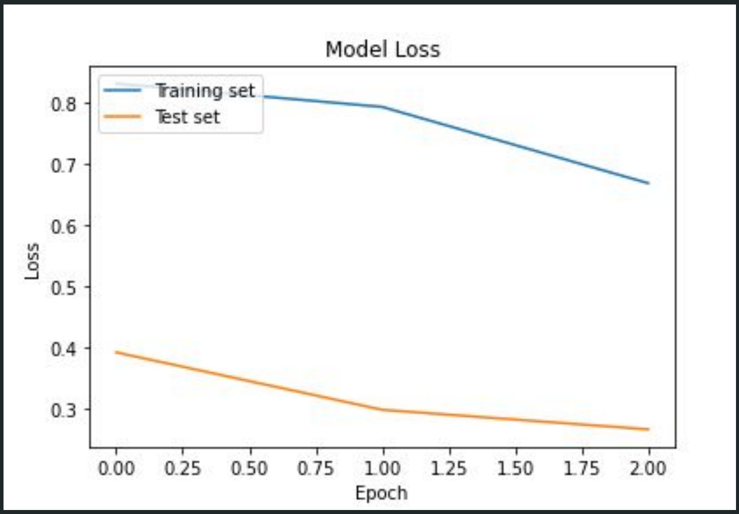
Pooling layer reduces the spatial size of the convolved feature which in turn decreases the computational power required to process the data. Additionally extracts dominant features.

After going through the above process, we have successfully enabled the model to understand the features.

Flatten the layer, so we can feed it to a regular Neural Network for classification purposes

Fully Connected Layers are a cheap way of learning non-linear combinations of the high-level features as represented by the output of the convolutional layer

Finding 1: Initial Convolutional Neural Network Quick and Easy to Build, but may need more “passes” with the data set

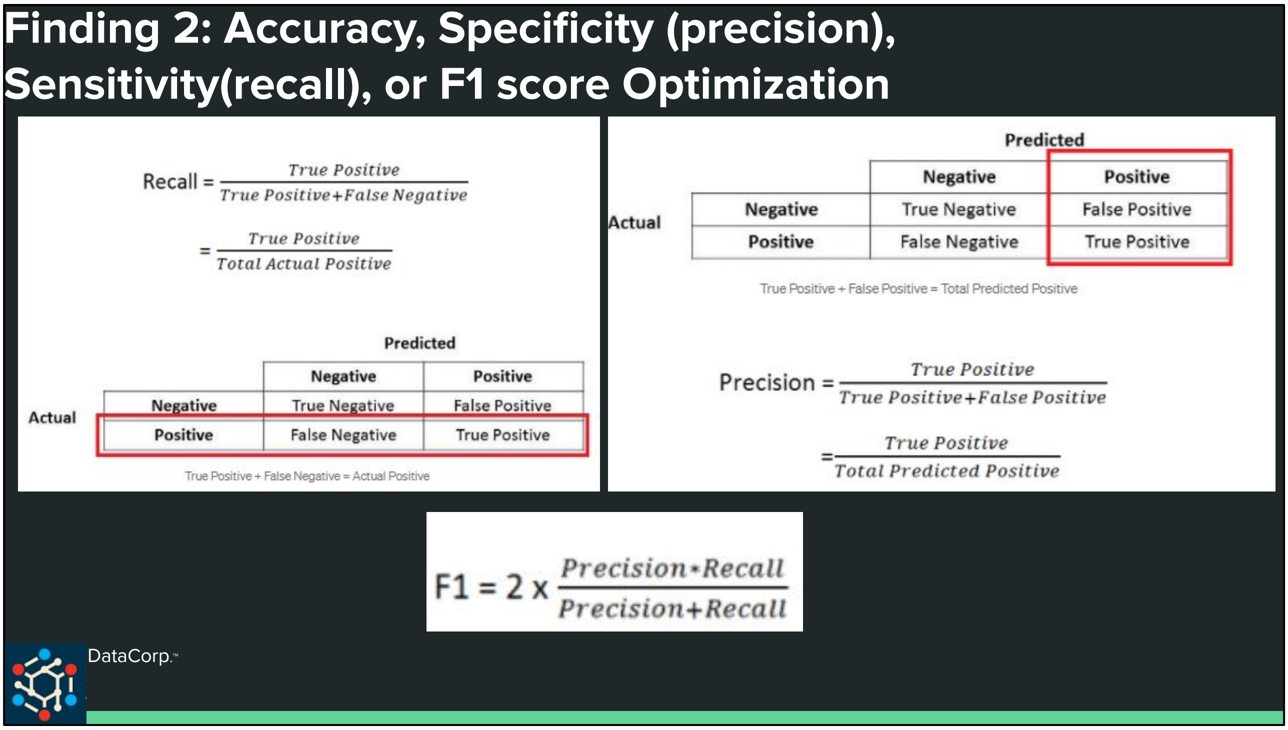


DataCorp™

Image:
https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Epoch is a term used in machine learning and indicates the number of passes of the entire training dataset the machine learning algorithm has completed

What is loss? loss is a number indicating how bad the model's prediction was on a single example. If the model's prediction is perfect, the loss is zero; otherwise, the loss is greater.



Finding 2: Accuracy, Specificity (precision), Sensitivity(recall), or F1 score Optimization

$$\text{Recall} = \frac{\textit{True Positive}}{\textit{True Positive}+\textit{False Negative}}$$

$$= \frac{\textit{True Positive}}{\textit{Total Actual Positive}}$$

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

True Positive + False Negative = Actual Positive

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

True Positive + False Positive = Total Predicted Positive

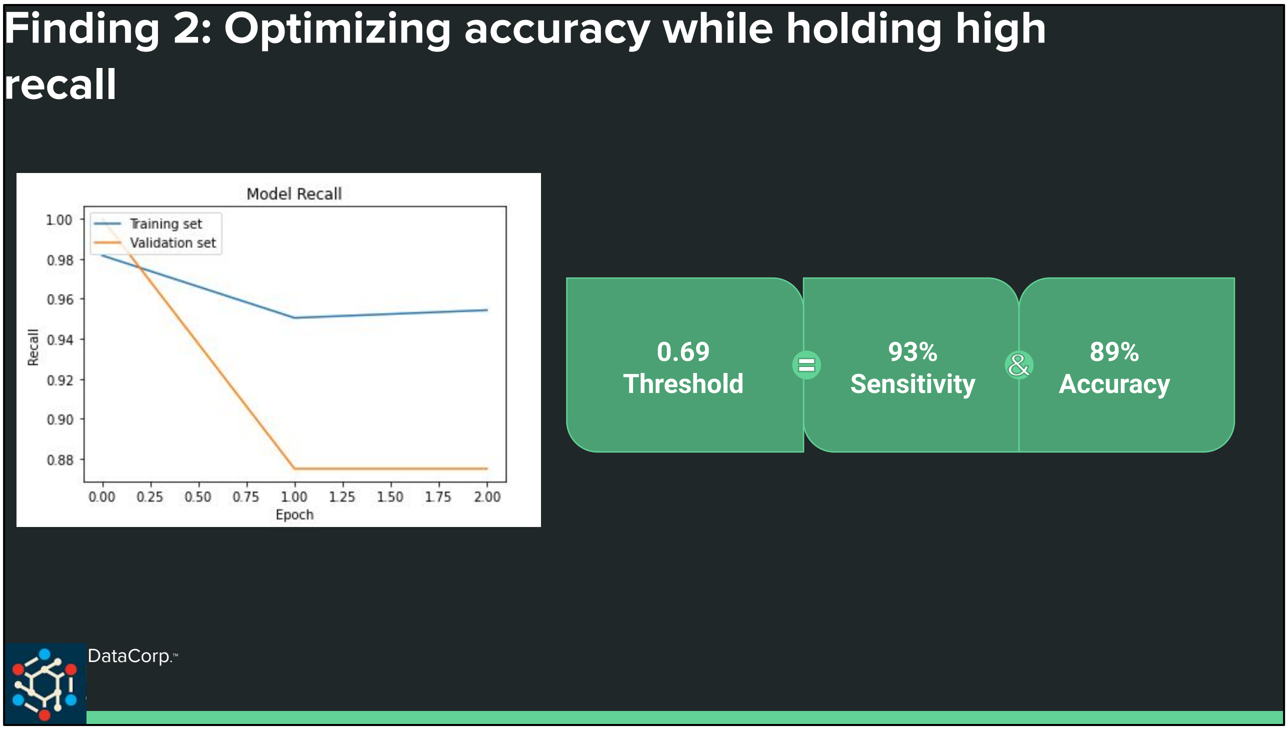
$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive}+\textit{False Positive}}$$

$$= \frac{\textit{True Positive}}{\textit{Total Predicted Positive}}$$

$$\text{F1} = 2 \times \frac{\textit{Precision} * \textit{Recall}}{\textit{Precision} + \textit{Recall}}$$

DataCorp™

It is necessary that pneumonia is caught with every positive patient, so we need a high sensitivity; however, we still want the model to be more accurate than a coin toss (50%)

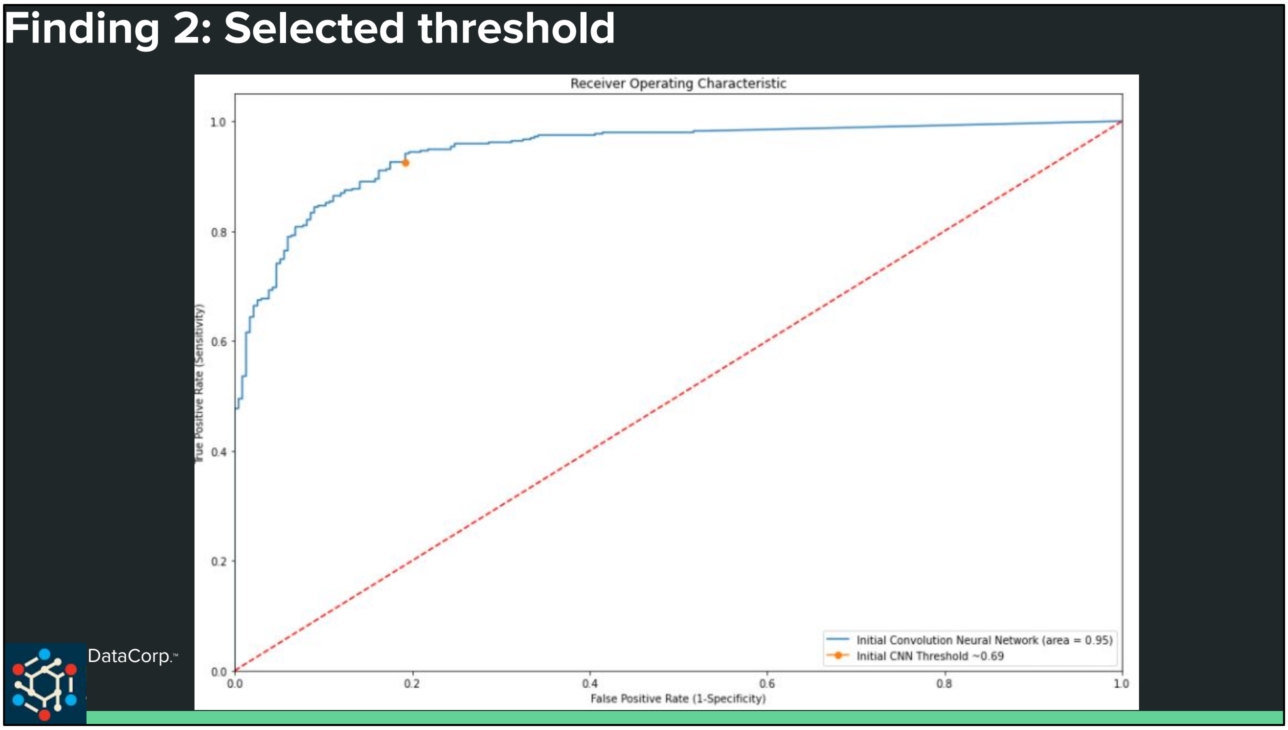


Finding 2: Optimizing accuracy while holding high recall

0.69 Threshold = 93% Sensitivity & 89% Accuracy

DataCorp.™

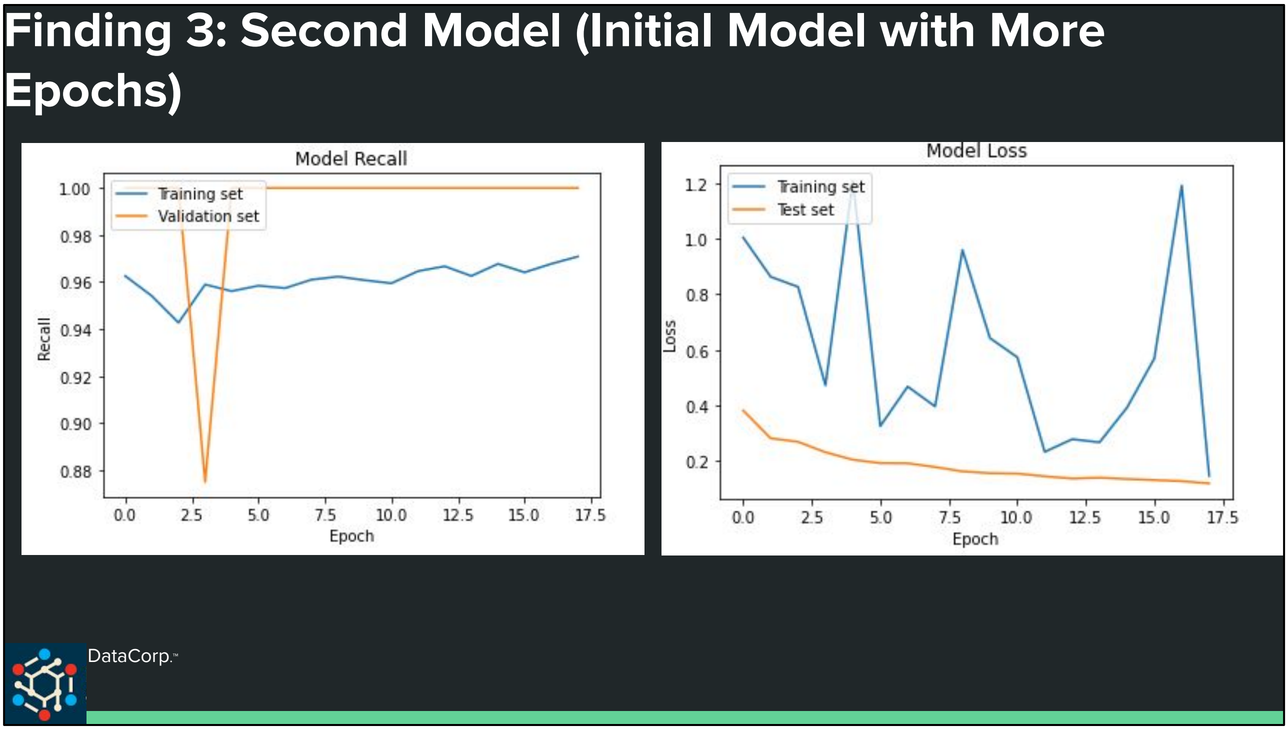
The model without threshold select results in a high recall score with the training set and oscillating high recall score with the validation set. With choosing the threshold at 0.69, we achieve a sensitivity percentage of 93% with an accuracy of 89% which is way better than a coin flip while maintaining a high sensitivity.



Finding 2: Selected threshold

Generally, the higher the AUC, the better the model is at distinguishing between clients that will subscribe and will not. So an AUC of 0.95 means that there is 95% chance that the model will be able to distinguish between 'pneumonia' class and 'normal' class. Additionally, it provides an aggregate measure of performance across all possible classification thresholds.

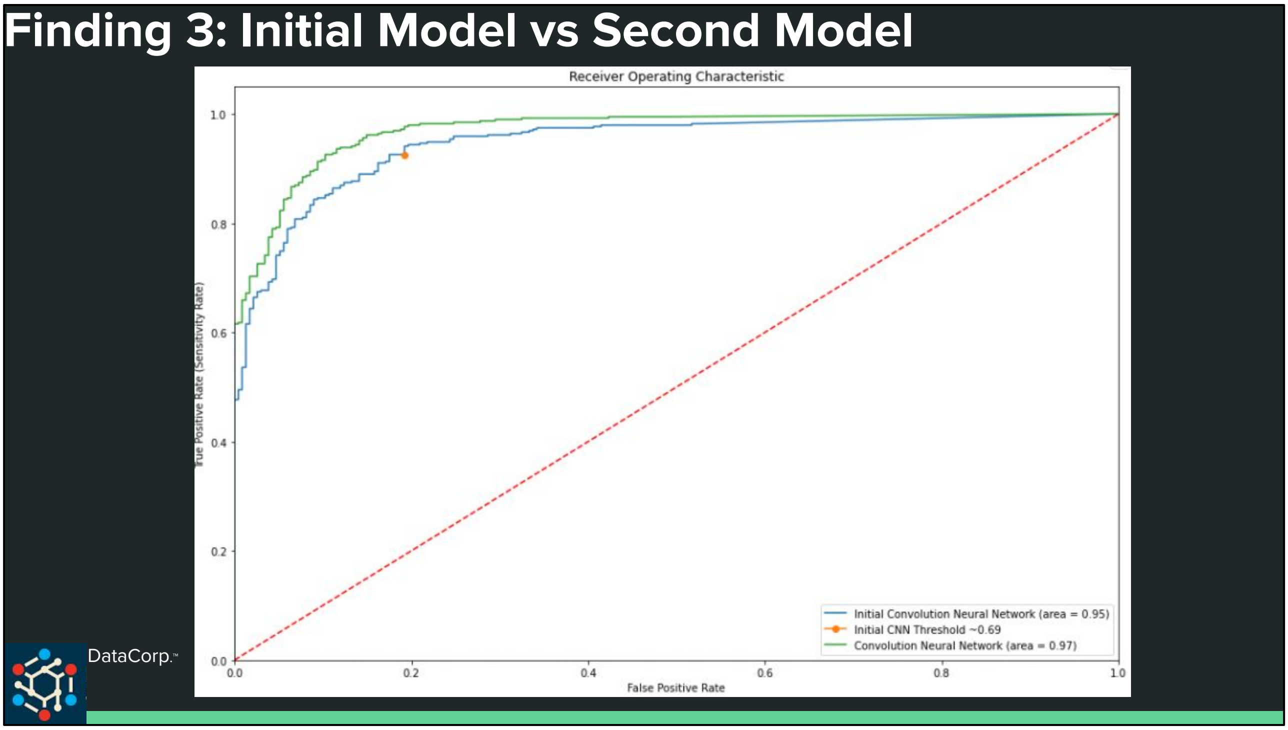
This threshold makes our model more specific, but less sensitive than the default threshold of 0.5; that is, our threshold model allows for more false negatives in order to reduce false positives so an optimal accuracy can be achieved.



Finding 3: Second Model (Initial Model with More Epochs)

DataCorp.™

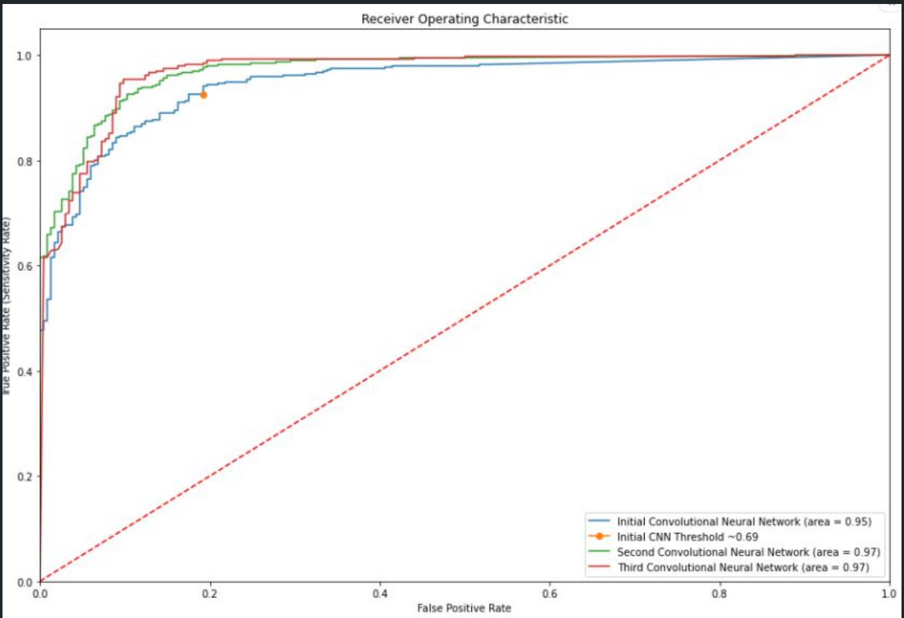
After many epochs, we see that the second model obtains a higher recall and lower loss than the initial model



Finding 3: Initial Model vs Second Model

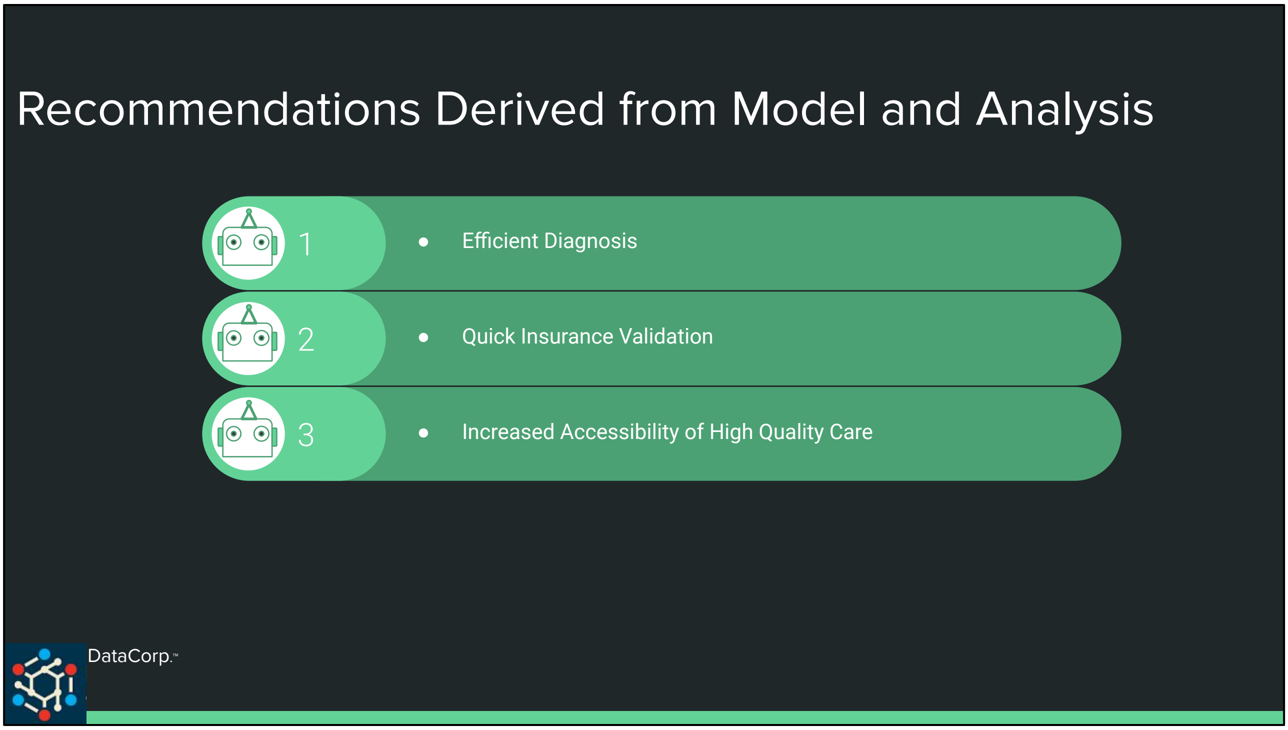
Here we can see that our second model has a better sensitivity rate at the initial threshold

Finding 3: All Models Compared



DataCorp™

Here we can see that our final model performs better than previous models at the set threshold without setting a threshold, and it has a higher AUC of 0.97



Recommendations Derived from Model and Analysis

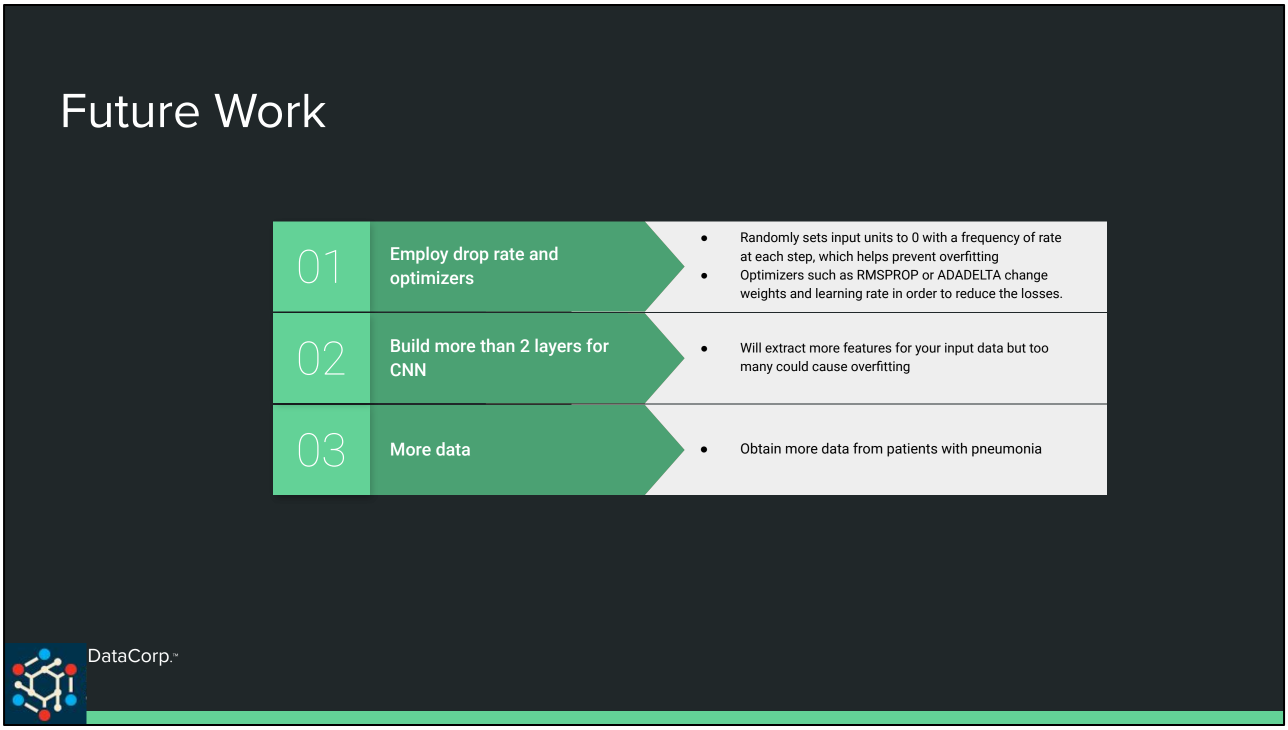
1
- Efficient Diagnosis

2
- Quick Insurance Validation

3
- Increased Accessibility of High Quality Care

DataCorp.™

With our model, we could see more time efficient diagnosis where doctors only look at patients flagged as negative and positive patients are moved quickly to treatment. Additionally we could see where preauthorizations are more quickly obtained for patients, so the financial aspect of treatment is resolved.
Finally, our model could increase accessibility of high quality care to rural or less funded medical facilities.



Future Work

01 Employ drop rate and optimizers

- Randomly sets input units to 0 with a frequency of rate at each step, which helps prevent overfitting
- Optimizers such as RMSPROP or ADADELTA change weights and learning rate in order to reduce the losses.

02 Build more than 2 layers for CNN

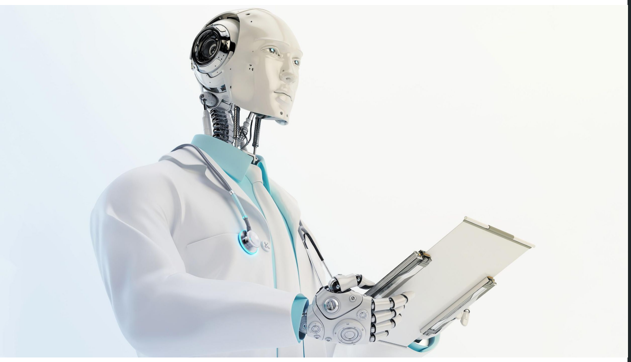
- Will extract more features for your input data but too many could cause overfitting

03 More data

- Obtain more data from patients with pneumonia

DataCorp.™

Now that we have proposed our three business recommendations, we know there are ways we can improve our current model. We can use drop rate and different optimizers which change weights and learning rate in order to reduce the losses. We can build a model that has more than 2 layers so more features can be extracted. Finally, we can get more data to train and test against our current model which will allows us to work towards a better model.



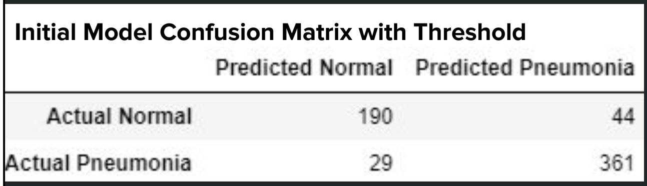
https://www.thetimes.co.uk/imageserver/image/%2Fmethode%2Fsundaytimes%2Fprod%2Fweb%2Fbin%2F20549920-bb0f-11e7-8b2e-f28d30e9c9fd.jpg?crop=2667%2C1500%2C0%2C0

DataCorp.™

Thank You

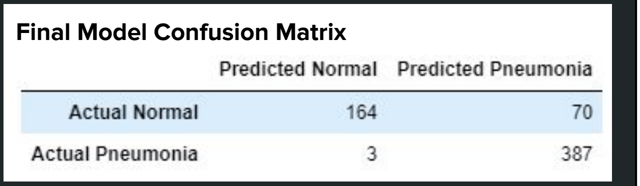
Thank you for your time! Please, feel free to ask me any questions at this time.

Appendix 1a. Confusion Matrix for Initial Model and Final Model



Initial Model Confusion Matrix with Threshold

	Predicted Normal	Predicted Pneumonia
Actual Normal	190	44
Actual Pneumonia	29	361



Final Model Confusion Matrix

	Predicted Normal	Predicted Pneumonia
Actual Normal	164	70
Actual Pneumonia	3	387