



# DIAGNOSING PNEUMONIA USING AI

## CAPSTONE 3

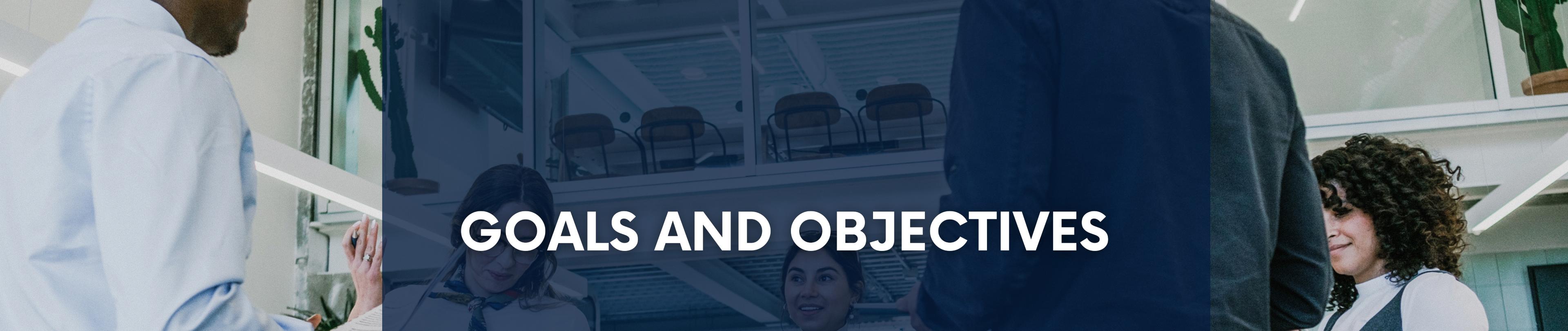
*Francisco J Torres*



# CONTENT

- 01** Loading the data and install necessary libraries:
- 02** Exploratory Data Analysis
- 03** Preprocessing: Data augmentation is performed to help model learn and be robust.
- 04** Modeling: Two Convolutional Neural Network models are built to classify X-ray images as Pneumonia or Normal.
- 05** Modeling evaluation and Metrics: Classification report, confusion matrix and test accuracy are used to evaluate model performance.
- 06** Recommendations:
- 07** Next Steps:





# GOALS AND OBJECTIVES

## Goals

The goal of this project is to build a Convolutional Neural Network that can be used to aid radiologists in diagnosing patients with Pneumonia.

---

## Target

In order to be considered useful I want the model to achieve at least a Recall score of 80%.

---

# THE DATASET

- This data set is available on Kaggle.com. The data set includes over 5,000 X-ray images. It is split into training images, testing images and validation images.
- The dataset has images of healthy lungs and images of lungs with pneumonia from bacterial and viral infections.



# 1. LOAD DATA AND INSTALL NECESSARY LIBRARIES

## LIBRARIES:

- MATPLOTLIB
- SEABORN
- KERAS
- CV2
- OS
- NUMPY
- PANDAS
- TENSORFLOW

## LOADING DATA:

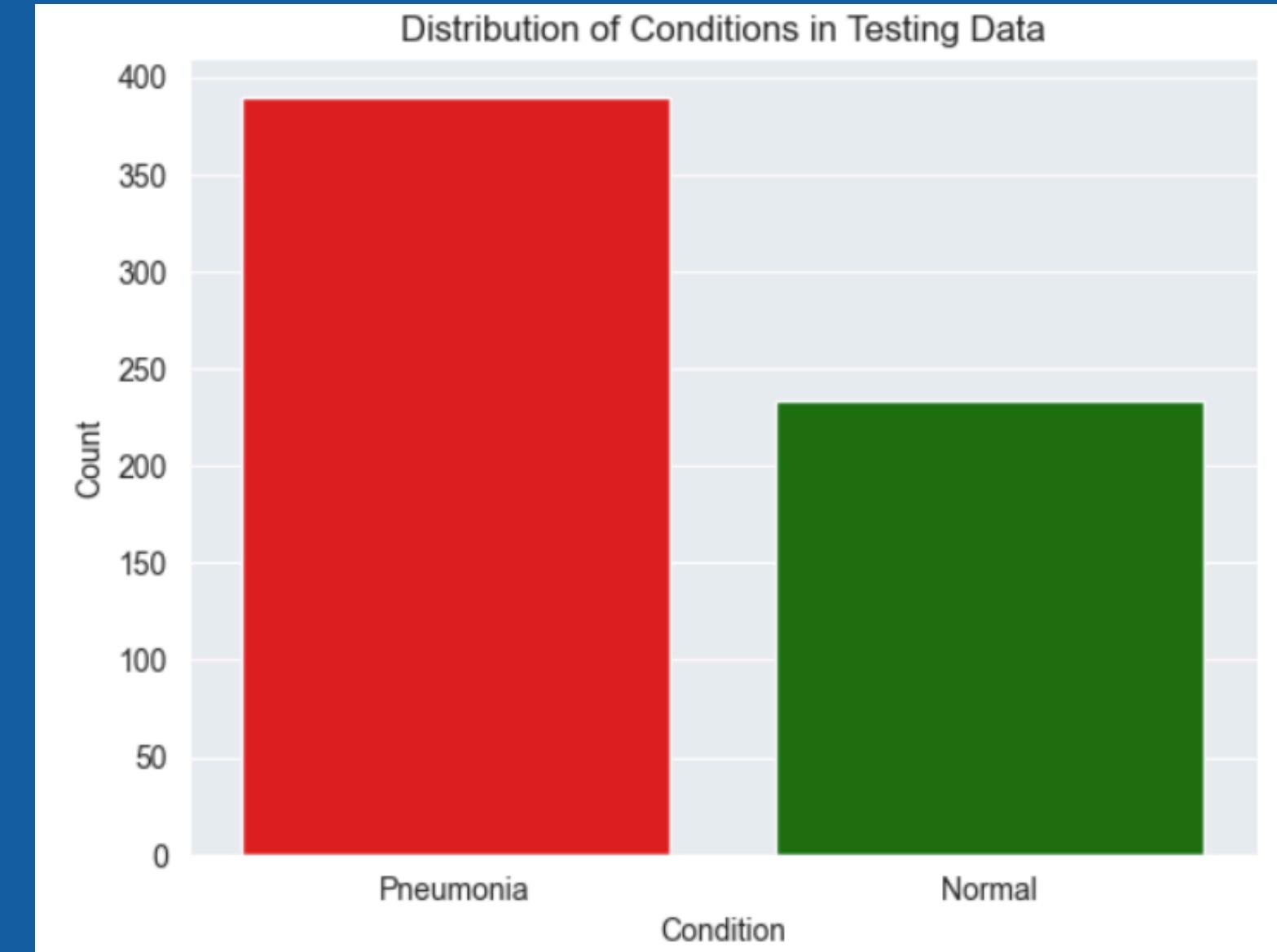
During this part of the project I experience an issue. This was my first time loading data into a notebook that is visual data (images). I had to do some exploration and look for resources online. After some time I learned that I could write a python function (**get\_training\_data**) that helped me load my X-ray images.



## 2. EXPLORATORY DATA ANALYSIS

### LOADING DATA:

- Once I successfully loaded the images into my notebook I began EDA.
- During EDA I discovered that the data set was unbalanced and the majority of images were of lungs with Pneumonia..
- This was an imbalanced dataset that could lead to overfitting.
- EDA helped me keep this in mind during model training.
- Training data was 5216 images.
- Testing data: 624 images.



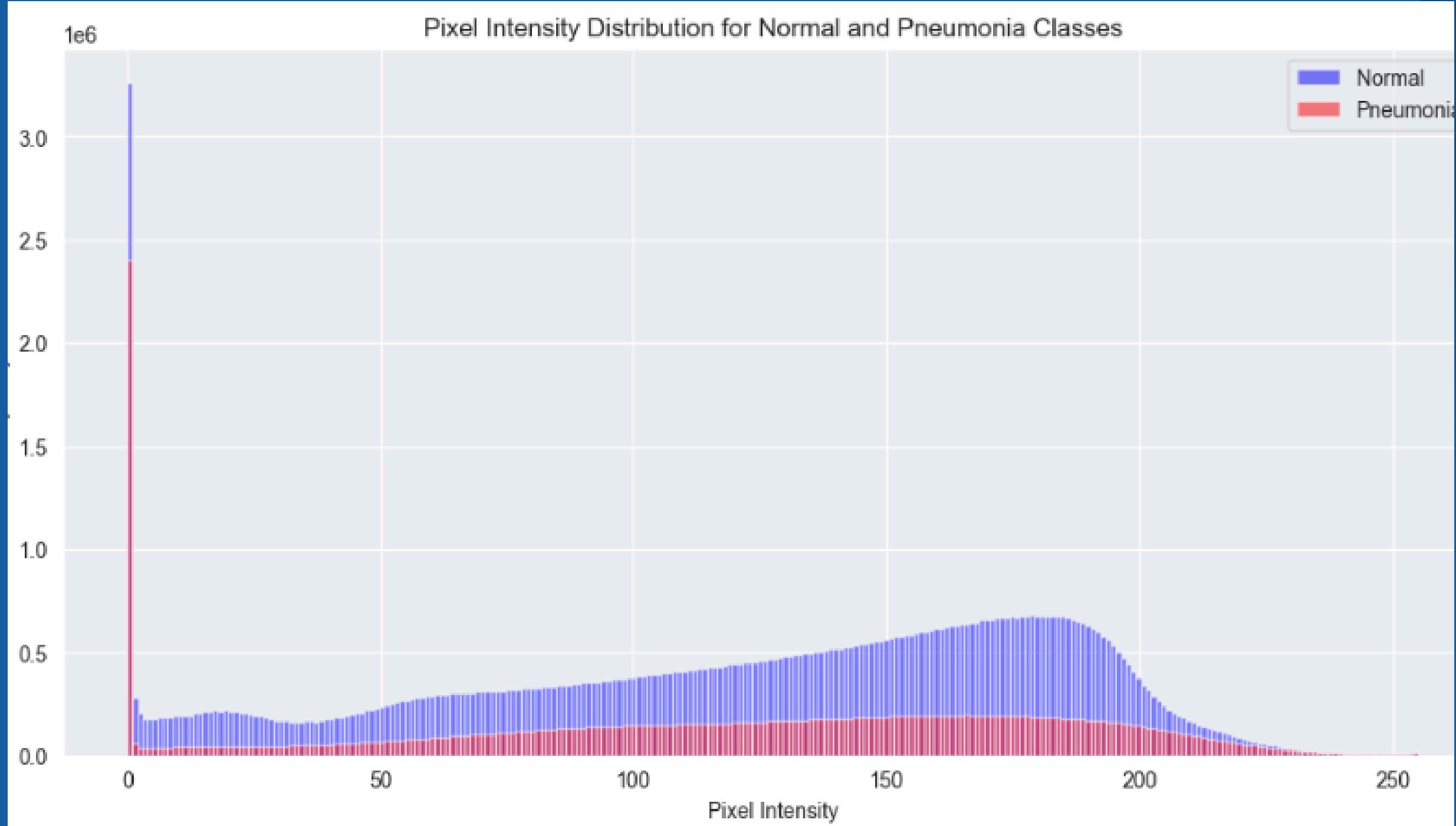
## 2. EXPLORATORY DATA ANALYSIS

### EDA CONTINUED:

During EDA I also managed to discover that the pixel intensity of the images with healthy lungs were on average greater than the images with Pneumonia.

This seemed to be an important feature that would have to be dealt with during preprocessing.

Note: This plot was created using Testing data.



## 2. EXPLORATORY DATA ANALYSIS

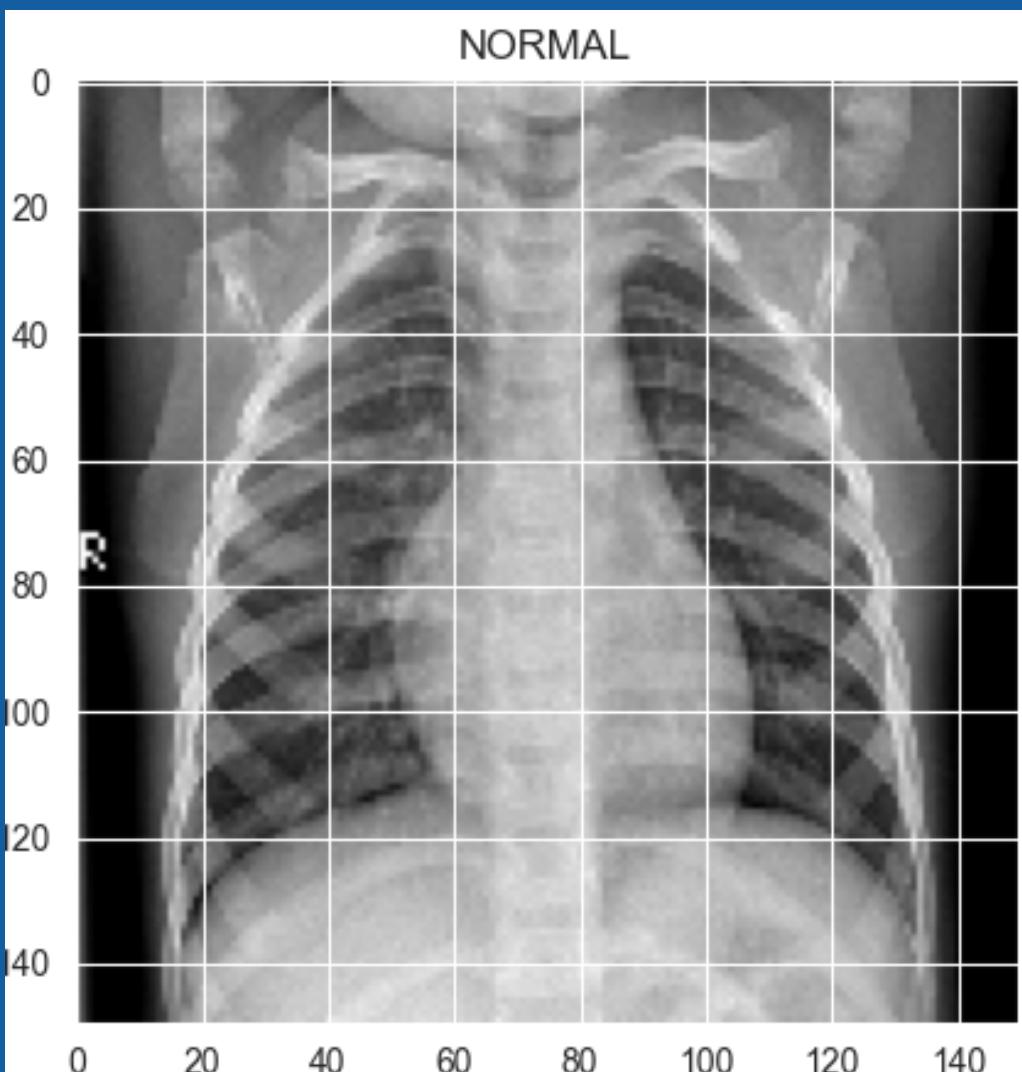
### COMPARING IMAGES BY EYE

Exploring intensity of Normal vs. Pneumonia images.

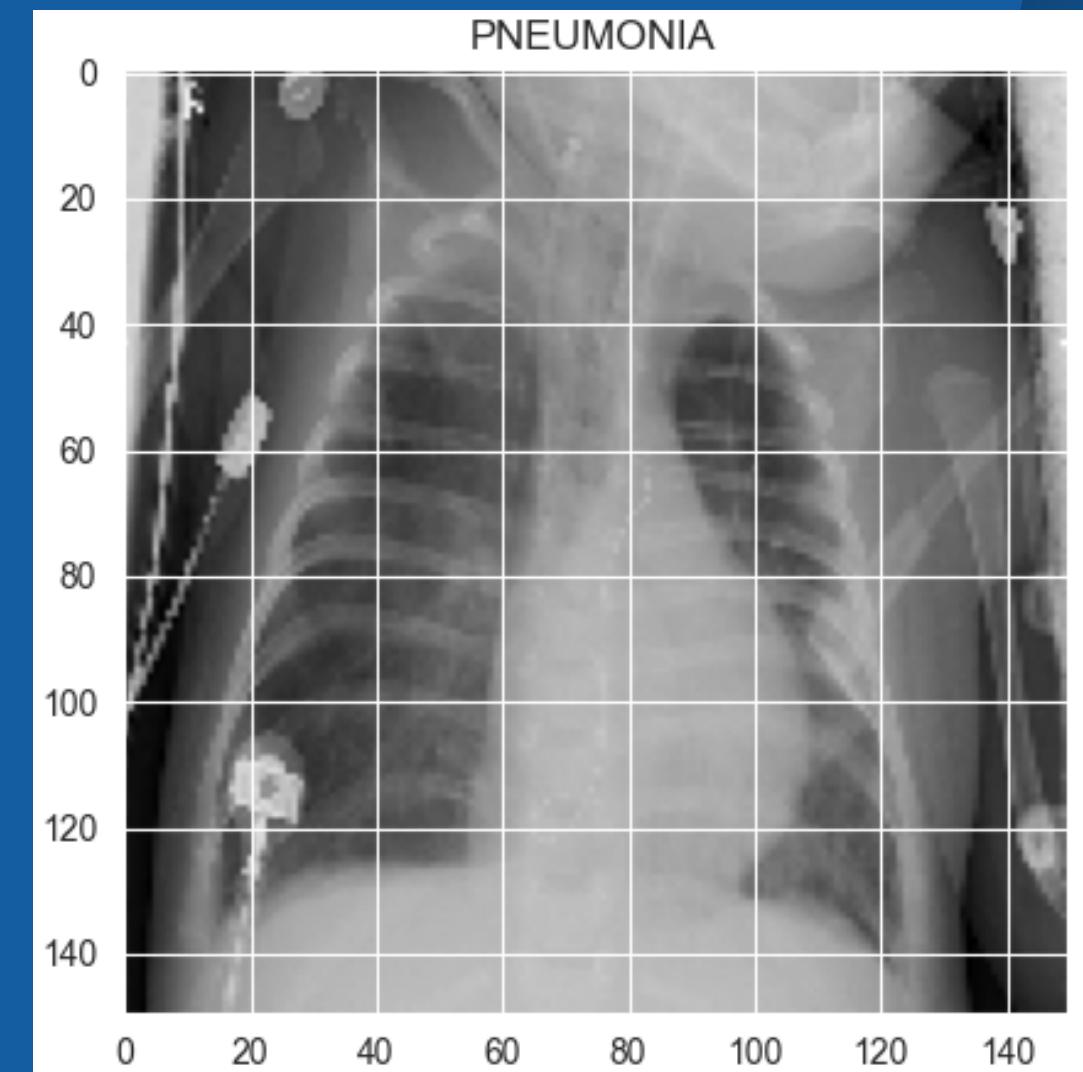
Normal image appears to be brighter and image quality is sharper.

Pneumonia image is more opaque and not as sharp.

NORMAL



PNEUMONIA



# 3. PREPROCESSING

## DATA AUGMENTATION:

Important steps:

- img\_size = 24: updated size of images to match model input.
- np.array: images and labels are extracted to prepare for CNN model.
- Normalize: data is divided by 255 to normalize it.
- Reshape: data is reshaped to match model expectation.
- Data was already split into Training, Testing and Validation so train\_test\_split was not necessary.

```
img_size = 224 # Updated size to match model input

# Extract images and labels
X_train = np.array([cv2.resize(item[0], (img_size, img_size)) for item in train])
y_train = np.array([item[1] for item in train])

X_test = np.array([cv2.resize(item[0], (img_size, img_size)) for item in test])
y_test = np.array([item[1] for item in test])

X_val = np.array([cv2.resize(item[0], (img_size, img_size)) for item in val])
y_val = np.array([item[1] for item in val])

# Normalize
X_train, X_test, X_val = X_train / 255.0, X_test / 255.0, X_val / 255.0

# Reshape to match model input shape
X_train = X_train.reshape(-1, img_size, img_size, 1)
X_test = X_test.reshape(-1, img_size, img_size, 1)
X_val = X_val.reshape(-1, img_size, img_size, 1)
```

# 4. MODELING

## MODEL 1

Convolutional Neural Network using Keras. I used Sequential model which adds layers one after the other. 9 layers were added which included filtering layers and max pooling layers. The activation functions used were ReLU and sigmoid.

The model was trained for 20 epochs.

## MODEL 1 LAYERS:

1. Conv2D – Layer 1
2. MaxPooling2D – Layer 2
3. Conv2D – Layer 3
4. MaxPooling2D – Layer 4
5. Conv2D – Layer 5
6. MaxPooling2D – Layer 6
7. Flatten – Layer 7
8. Dense (100 neurons) – Layer 8
9. Dense (1 neuron) – Layer 9

## MODEL 2

Model 2 was built as an upgrade of model 1 with some additional layers to help reduce potential overfitting and improve generalization.

The updates included:

- BatchNormalization: improve robustness of model.
- Dropout: Helps prevent overfitting.
- Kernel regularization with L2: Adds penalty to loss function to help reduce overfitting.

# 5. MODEL EVALUATION METRICS

## Evaluation Metrics:

- Test Accuracy
- Recall Score
- Confusion Matrix
- Classification Report



# MODEL 1 EVALUATION METRICS

Recall Score

84%

Accuracy

82%

	precision	recall	f1-score	support
Normal	0.89	0.80	0.84	390
Pneumonia	0.72	0.84	0.77	234
accuracy			0.82	624
macro avg	0.81	0.82	0.81	624
weighted avg	0.83	0.82	0.82	624

# MODEL 2 EVALUATION METRICS

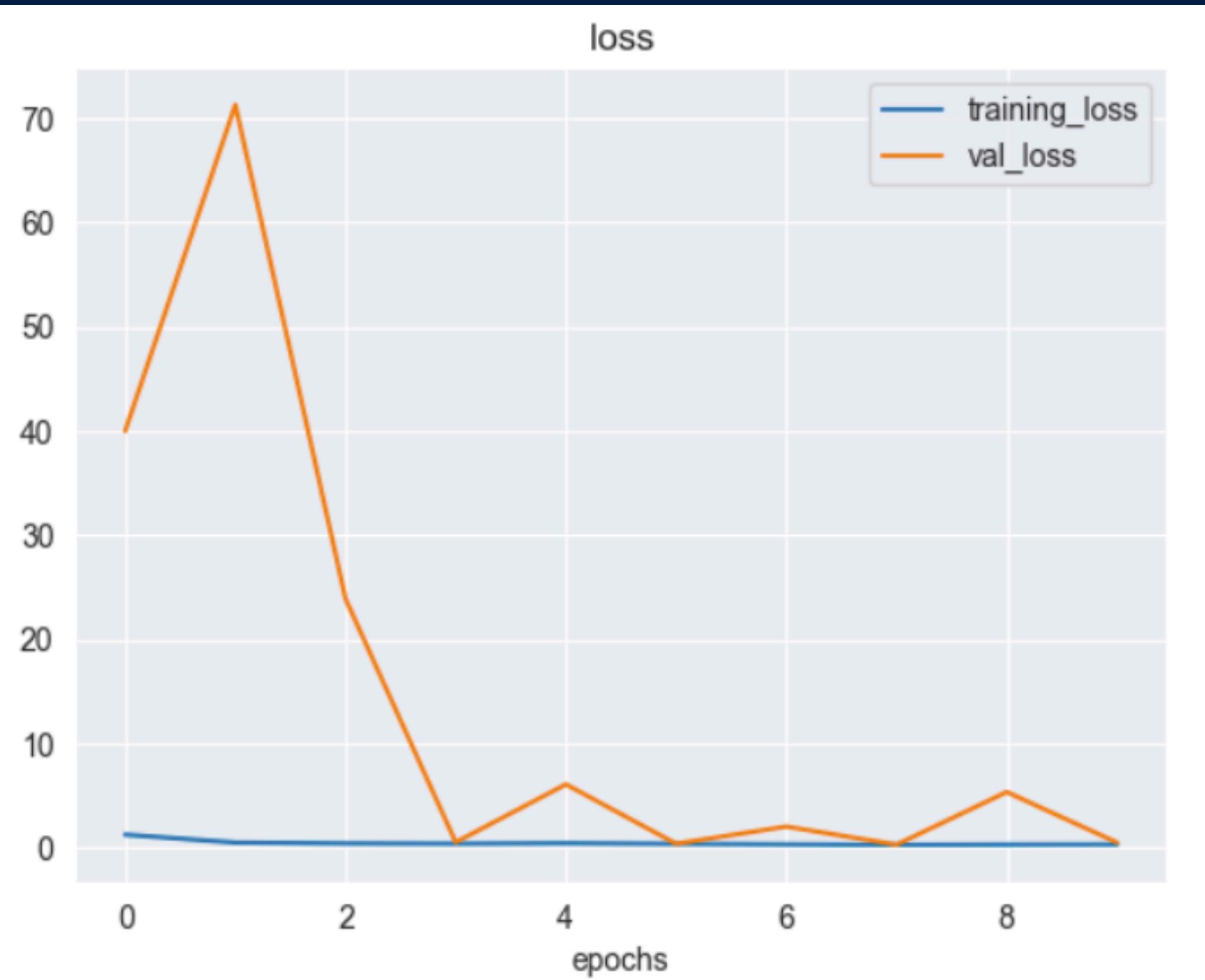
Recall Score **85%**

Accuracy **89%**

	precision	recall	f1-score	support
Normal	0.91	0.92	0.91	390
Pneumonia	0.86	0.85	0.85	234
accuracy			0.89	624
macro avg	0.88	0.88	0.88	624
weighted avg	0.89	0.89	0.89	624

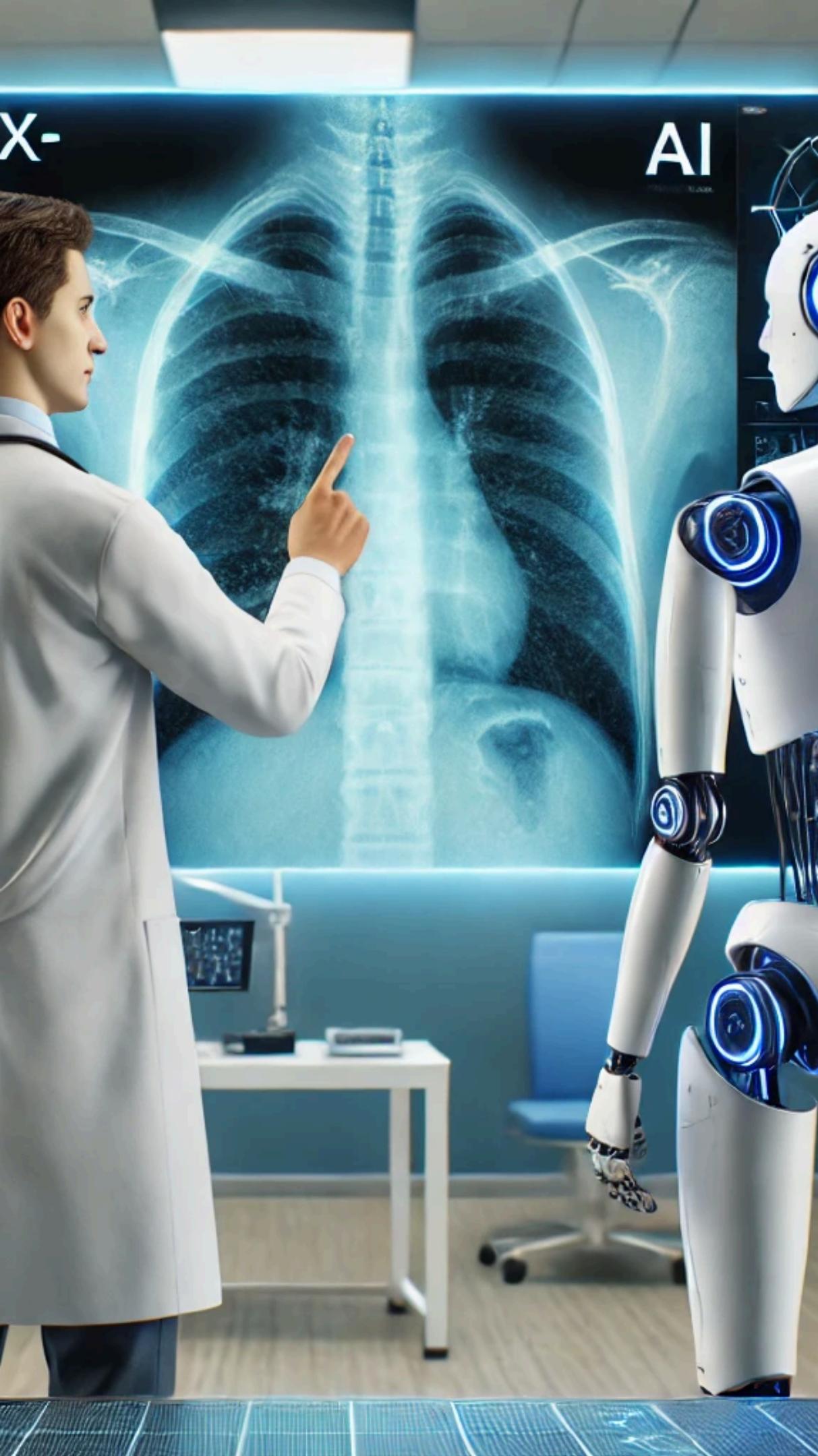
# TRAINING EPOCHS

Model was trained on 10 epochs. Loss curve shows that 5 epochs could be optimal as after this it scores stagnate.



# **FUTURE STEPS TO IMPROVE MODEL:**

- 1) Make model more complex by adding additional layers.**
- 2) Use data augmentation techniques.**
- 3) Training for a greater number of epochs.**
- 4) Use pretrained model.**



## 6. RECCOMENDATION



### Improve efficiency

Models such as these can be used to improve speed and efficiency in detection of Pneumonia. These models can help increase productivity of Radiologists.



### Prescreening Tool

Models such as this one can be used to detect early signs of Pneumonia that might not be immediately apparent to the human eye. Early diagnosis can be vital in treatment of illness.



### Address inequality

These models can be used in areas where Radiologists are scarce or unavailable. These models can be made available online and images of X-rays can be uploaded to help underdeveloped areas improve diagnosis of Pneumonia.

# NEXT STEPS AND FUTURE PROJECTS



## Improve Model Recall

I would like to keep working with this dataset and continue to improve my model to achieve a Recall score of at least 90%.



## Project 2

I would also like to create a model that can detect the difference between Pneumonia caused by a bacterial infection vs. a viral. Pneumonia caused by bacterial infection can be severe.



## Project 3

Anomaly Detection  
Train a CNN on normal images to identify abnormal structures in CT, MRI, or X-ray scans.

# ABOUT ME

I



Francisco J  
Torres

Datascience  
Career Track,  
Springboard

I AM A LIFELONG LEARNER WITH AN INSATIABLE CURIOSITY. MY DIVERSE EXPERIENCES INCLUDE WORKING AS A FOREST FIREFIGHTER, A HIGH SCHOOL TUTOR, AND AN ENGLISH TEACHER IN SPAIN. I HAVE EARNED A MASTER OF ARTS IN TEACHING WITH A CONCENTRATION IN SECONDARY MATHEMATICS. DURING THE PANDEMIC, I DISCOVERED MY PASSION FOR PROGRAMMING AND DATA SCIENCE, WHICH I PURSUED AS A HOBBY FOR FOUR YEARS BEFORE DECIDING TO TRANSITION INTO THE FIELD PROFESSIONALLY.