

# **Pneumonia Detection in X-Rays Using Deep Learning**

---

Tan Yue Hang

Data Science and AI (Institute of Data)

Capstone Project



## Author's Bio

- **Name:** Tan Yue Hang
- **Education:**
  - Master of Engineering Science (M.Eng.Sc.)
  - Bachelor of Mechanical Engineering
  - Professional Certificate in Data Science and Artificial Intelligence
- **Certification & Achievements:**
  - IBM - Databases and SQL for Data Science with Python
  - Six Sigma Green Belt
  - IEEE academic publications
- **Profession:**
  - Senior Process R&D Engineer
  - Aspired Data Scientist
- **Email:** tan\_hang2003@yahoo.com
- **LinkedIn:** [Yue Hang Tan](#)





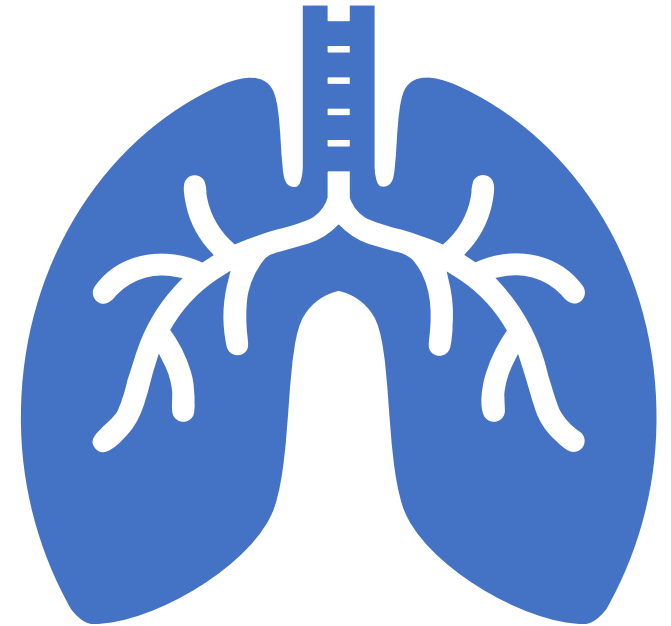
# Disclaimers:

- The medical information gathered and compiled in this project is only meant for supporting deep learning model building and **should not be taken as medical advice** by any means.
- The objective of this project is to build a Deep Learning model that classifies chest X-Ray images for educational purposes only. It is not approved by any certified medical professional for real-world application. Please **seek advice from certified medical professionals** should you have any medical-related questions.

---

# Abstract

Pneumonia, a severe lung condition, has impacted countless individuals globally. Some fatalities have resulted from misdiagnoses due to the complexities of image interpretation that require skilled radiologists. This issue is further complicated by the large volume of patients and X-Ray images that need to be analyzed daily. The objective of this project is to develop a deep learning model that employs transfer learning with a pre-trained CNN model to distinguish between healthy lungs and those infected with pneumonia. The findings indicate that a recall score of 98% can be attained by using DenseNet169 for feature extraction from the chest X-Ray images, coupled with multiple dense layers with an optimized number of neurons for classification.



## **Section 1:**

Challenges in Healthcare

## **Section 2:**

Data Cleaning and EDA

## **Section 3:**

Transfer Learning with Pre-trained  
CNN Model

# **Presentation Contents**

# Section 1: Challenges in Healthcare

## Business Problems:

- Diagnosis Accuracy
- Time and resources





# The Disease

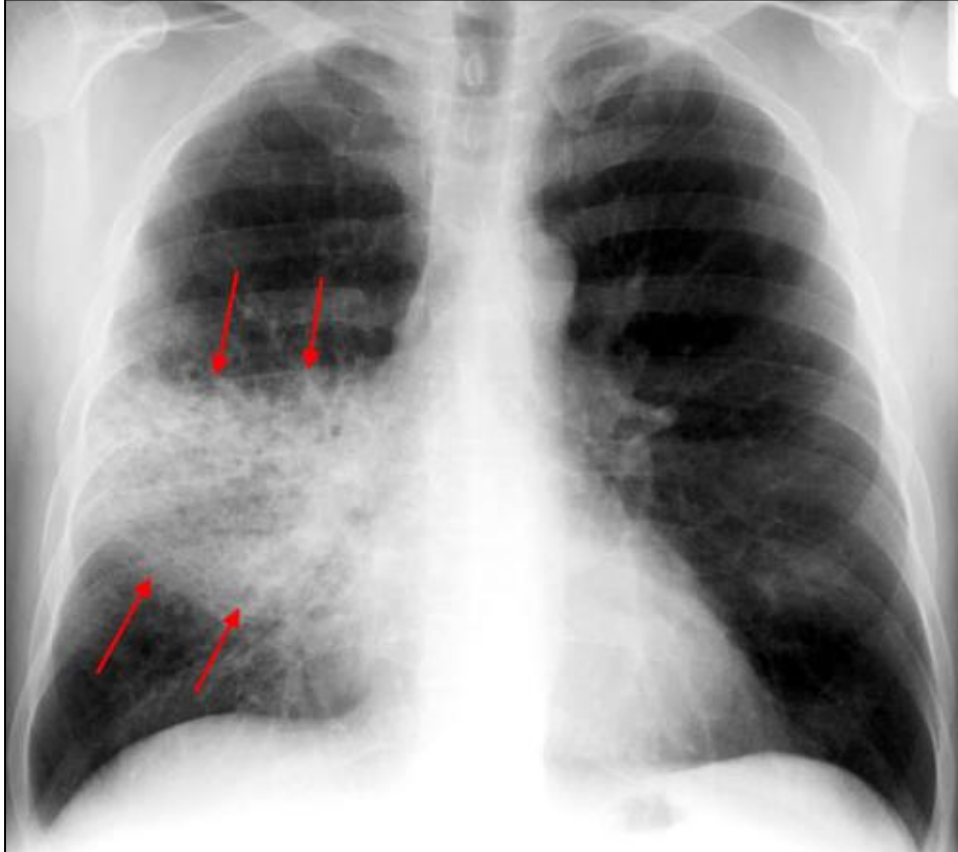
- Pneumonia is an infection that inflames the air sacs in one or both lungs. It can be caused by bacteria, viruses (influenza or COVID-19), or fungi. The air sacs may fill with fluid or pus, causing discomfort and suffering.
  - Symptoms: Cough with phlegm, fever, chill and difficulty in breathing.
  - Severity: This depends on factors such as age, overall health, and whether the infection is bacterial or viral.
- Mortality Rate:

Demography	Mortality Rate
Hospitalized Patients	<a href="#">5 – 10%</a>
Patients in ICU	<a href="#">Up to 30%</a>

- Treatment: Antibiotics (bacterial), cough suppressants, pain medications. → There may be more tolerance for having more **false positives** in the confusion matrix. ◀
- Diagnosis methods: **Chest X-Ray**, breathing test, sputum test, urine test, blood test, bronchoscopy, etc.



# Human Diagnosis



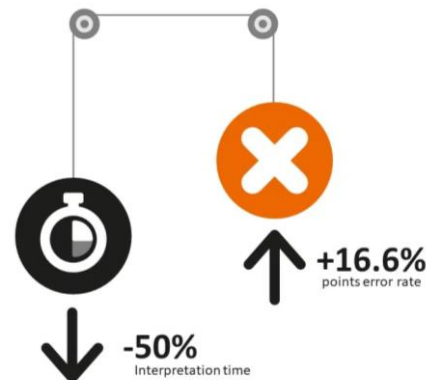
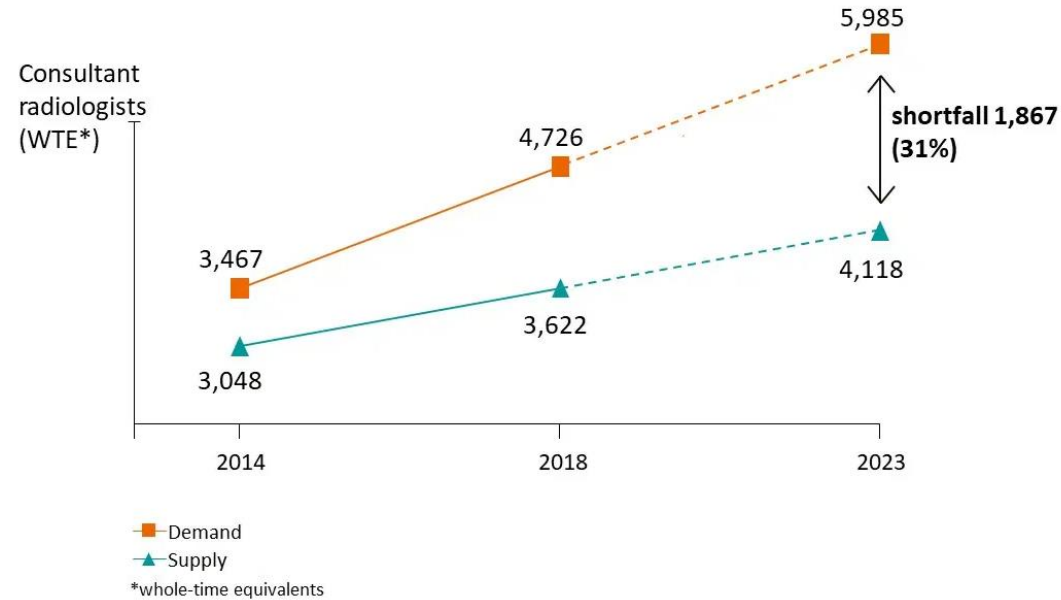
An X-Ray image of pneumonic lung. [2]

- H. J. Koo et al. (2018) mentioned that various factors could affect how a pneumonic chest X-Ray image appears: [1]
  - Viral or bacterial infection
  - Family of the viruses
  - Age
  - Immune status
  - Seasonal variation & community outbreaks
- The observation of pneumonia in lungs is generally complex. Depending on the type of infection, the observation could come in various sizes. [2]
- An **example** of how a radiologist interprets a chest X-Ray image on the left:
  - “The trachea is centrally located with no mediastinal shift. The costophrenic angles are sharp and unobscured. A non-uniform opacity is present in the right middle zone, characterized by visible air bronchograms and blurry borders. The right cardiac border is obscured, indicating right middle lobe pneumonia. The size of the heart appears to be normal.” [2]



# Challenges in Healthcare Sector

Radiology is grappling with a data surge from more exams and a staff shortage. The result is a heavy workload for radiologists as exam numbers outpace specialists. [3]



- Increased workload risks more interpretation errors. Cutting interpretation time in half ups the error rate by 16.6%. [4]

**This is where AI can improve diagnostics. [3]**

# Deep Learning Imaging: Intersection of AI and Healthcare




- Early detection and treatment are key. Therefore, fast and effective diagnosis plays a crucial role in the early detection of pneumonia.
- In the study by L. L. Plesner et al. (2023), AI was able to correctly identify about 99.1% of abnormal X-rays and about 99.8% of critical X-rays. On the other hand, human radiologists were able to correctly identify about 72.3% of abnormal X-rays and about 93.5% of critical X-rays. [5]
- AI can identify features that are not easily detected by the human eyes, thereby transforming radiology from a subjective perceptual skill to a more objective science. [6]
- The enhanced efficiency offered by Artificial Intelligence empowers radiologists to undertake tasks of greater value, thereby increasing their visibility to patients. [6]

Comparisons of AI Accuracy with Human Radiologists' Accuracy [5]

	Abnormal X-Rays Accuracy	Critical X-Rays Accuracy
AI	99.1%	99.8%
Human Radiologists	72.3%	93.5%

# Source of Dataset

- The dataset was acquired from [Kaggle](#), and originally came from [Mendeley](#).
- The dataset is organized into three folders: train, val, and test. Each folder contains two sub-folders, **Normal** and **Pneumonia**, which represent the two classes of the X-Ray images.




Mendeley Data

You are viewing a previous version of this dataset

## Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification

Published: 6 January 2018 | Version 2 | DOI: 10.17632/rscbjbr9sj.2  
Contributors: Daniel Kermany, [Kang Zhang](#), [Michael Goldbaum](#)



Search

Sign In Register

PAUL MOONEY · UPDATED 6 YEARS AGO 6207 New Notebook Download (2 GB)

## Chest X-Ray Images (Pneumonia)

5,863 images, 2 categories


Data Card Code (2376) Discussion (59) Suggestions (0)

### About Dataset

**Context**  
[http://www.cell.com/cell/fulltext/S0092-8674\(18\)30154-5](http://www.cell.com/cell/fulltext/S0092-8674(18)30154-5)

**Usability** 7.50

**License**  
Other (specified in description)



# Section 2:

## Data Cleaning & EDA

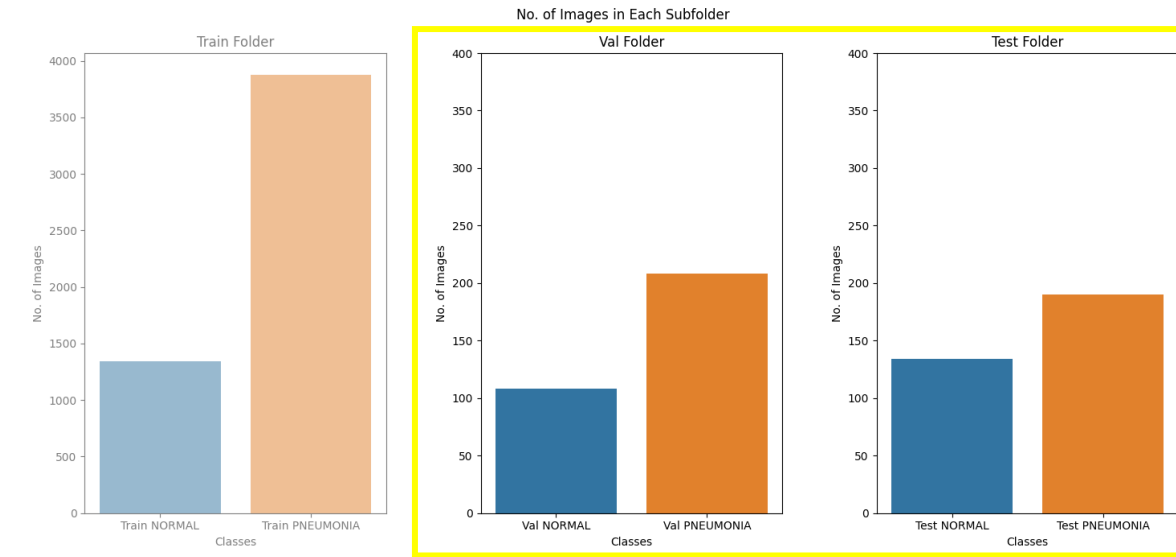
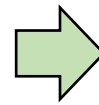
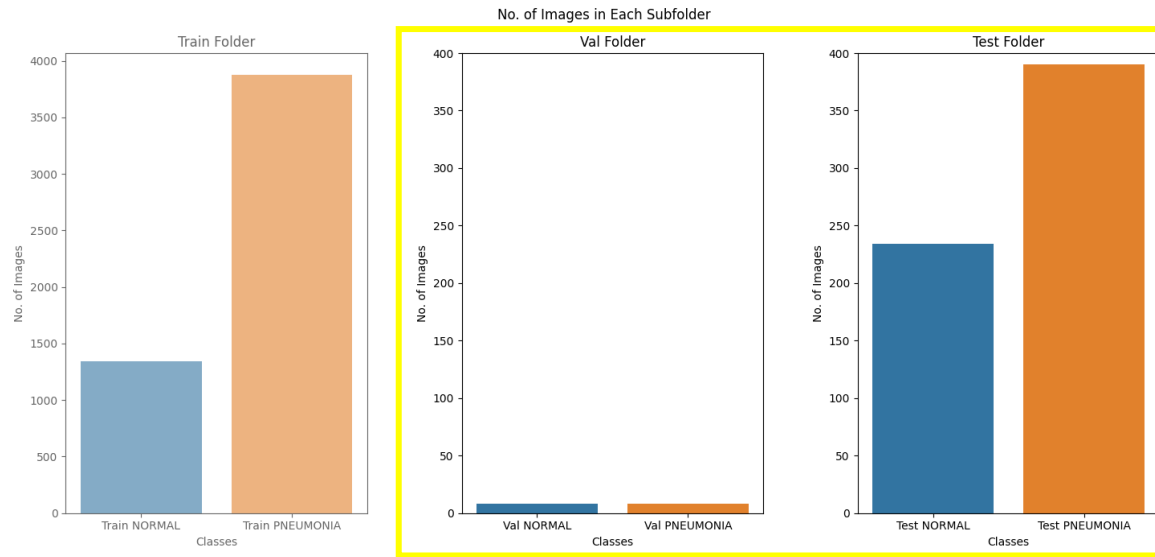
### Data Science Problems:

- Sample size of data
  - Cleanliness of data
- 



# Re-arranging the Images

- There are two issues with the dataset:
  - The training images are imbalanced. The Pneumonic class is about 2.9 times larger than the Normal class.
  - There are too few validation images.
- Apply more class weights to the Normal class than the Pneumonic class during model training (3:1).
- Move some of the image data from the test folder to the validation folder.



	Train		Val		Test	
Classes	Normal	Pneumonia	Normal	Pneumonia	Normal	Pneumonia
No. of Images	1341	3875	8	8	234	390

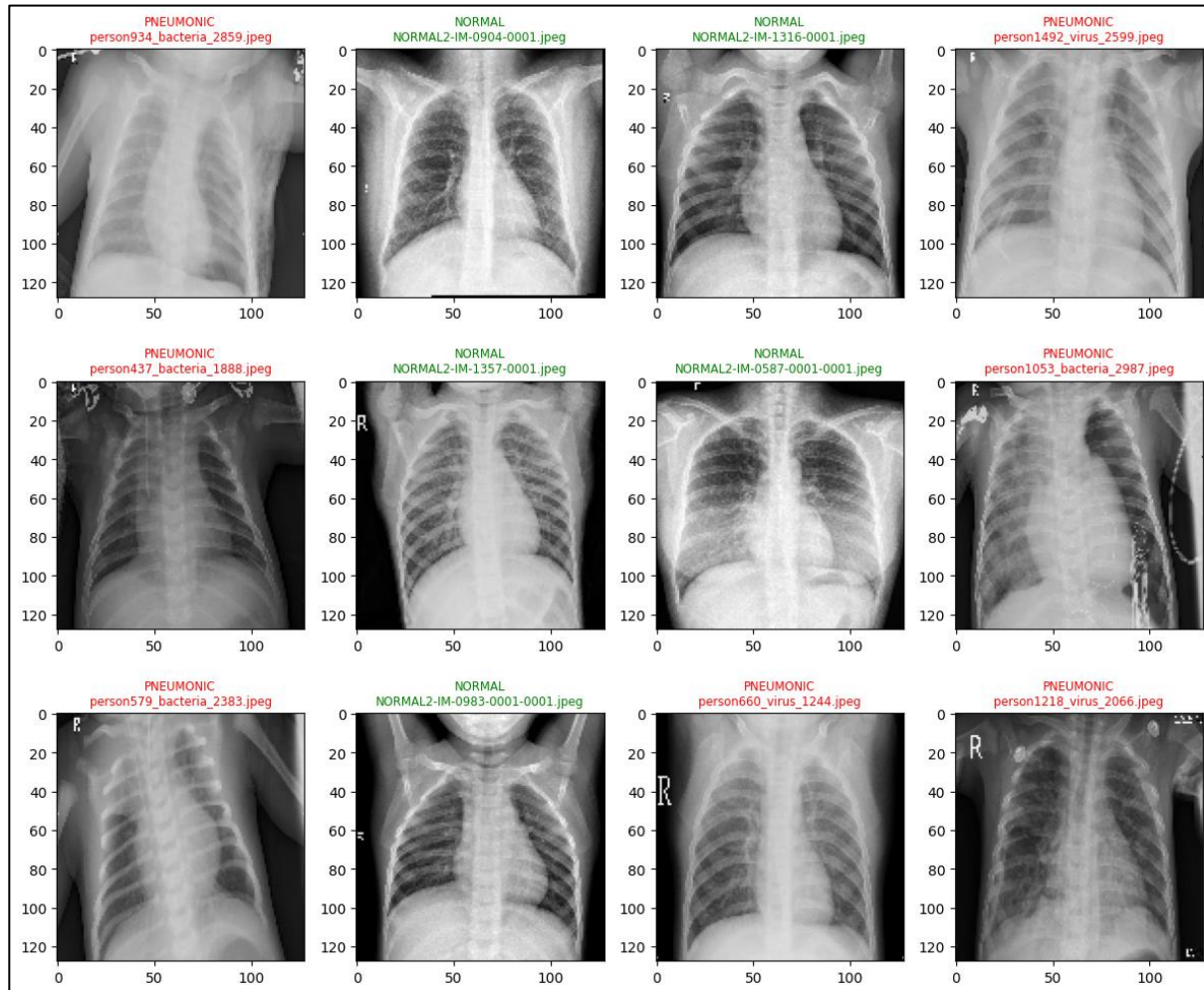
Before moving

	Train		Val		Test	
Classes	Normal	Pneumonia	Normal	Pneumonia	Normal	Pneumonia
No. of Images	1341	3875	108	208	134	190

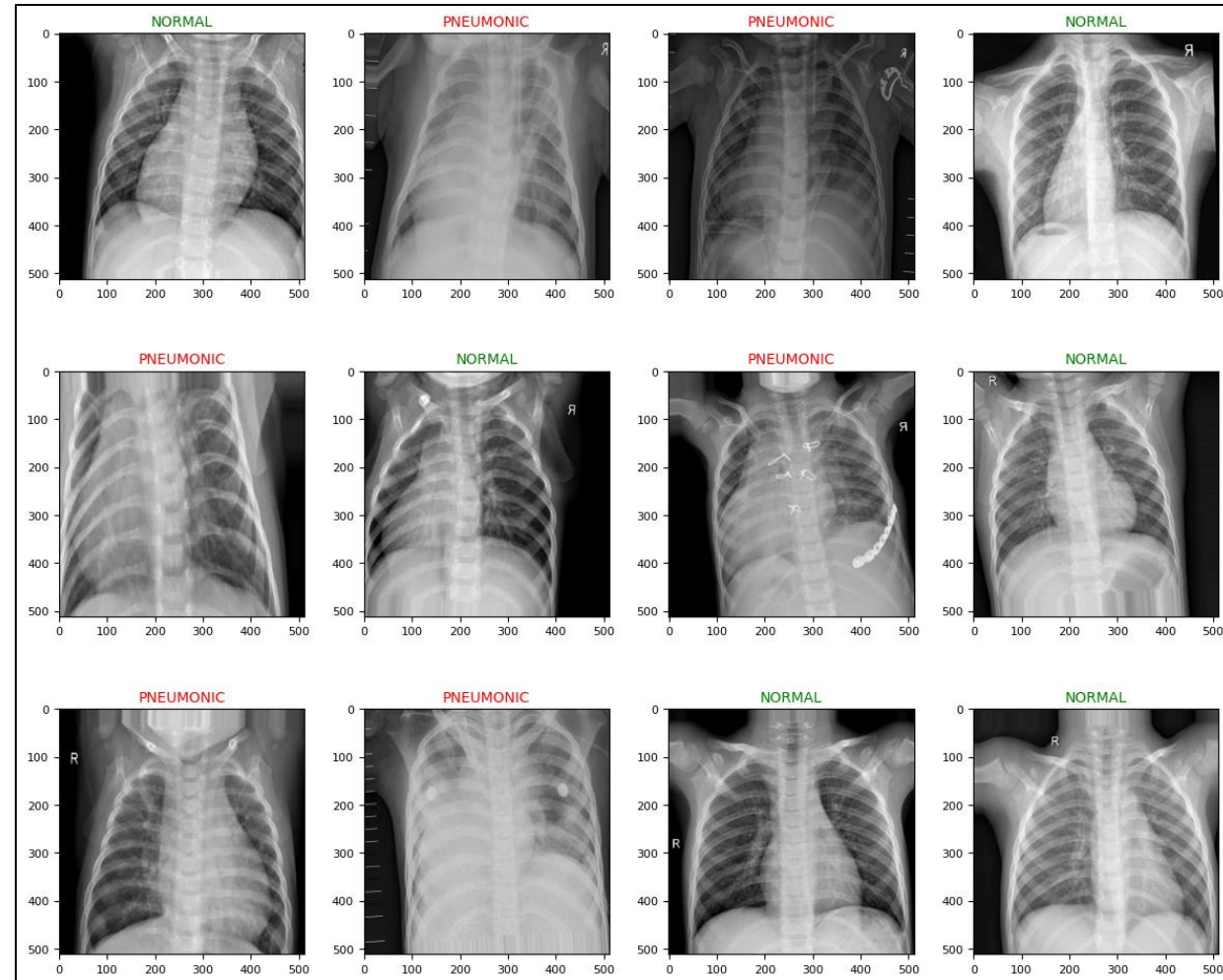
After moving some images from test to val folder



# Chest X-Ray Images Preview




Without Augmentation



With Augmentations

- rotation\_range=5, width\_shift\_range=0.1, height\_shift\_range=0.1, horizontal\_flip=True
- All the images are randomly loaded and plotted. Therefore, the images on the right are **not** the same as the ones on the left.



# Section 3:

## Transfer Learning with Pre-trained CNN Model

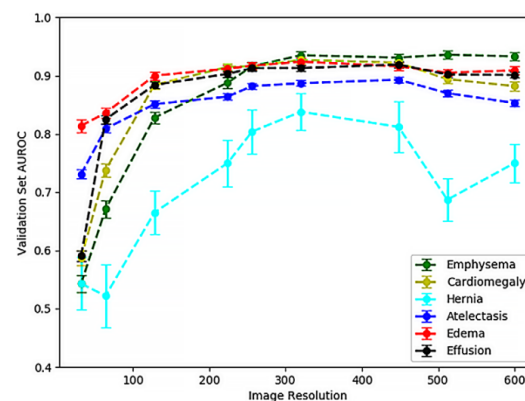
### Data Science Problems:

- Metric
- CNN model
- Hyperparameters tuning

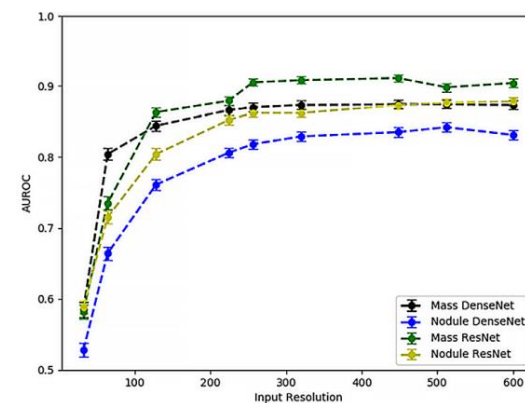


# Literature Reviews

- **CNN model:** Campos-Lopez et al. (2023) demonstrated that DenseNet201 and ResNet50 work well on X-Ray images of pneumonia [7]. This aligns with the findings from Sabottke and Spieler (2020) as well. [8] VGG16 and VGG19 were used in Campos-Lopez et al. (2023) study as well, but they did not offer comparable performance. [7]
- **Image resolution:** Sabottke and Spieler (2020) discovered that lower resolution (64 x 64) works better for observations that is generally larger in size, whereas a higher resolution (256 x 256, 448 x 448) could work better for subtler observations and binary classification. [8]
- **Full stage data augmentation:** In the study by Q. Zheng et al. (2020), more than an 8% improvement in accuracy was observed with a full stage data augmentation framework on both CIFAR-10 (coarse-grained) and CIFAR-100 (fine-grained) datasets. Among these, translation, horizontal flip, and scale transformation seem to contribute more to the improvement. [9]
- Similarly, M. Nagaraju et al. (2022) also discovered a similar outcome: the Full Stage Data Augmentation Framework shows better performance than the existing method by up to 28.31%. [10]
- **Batch normalization:** This is a regularization method that standardizes the activations within a layer. This standardization is achieved by subtracting the mean of the batch from each activation and then dividing the result by the batch's standard deviation. [11]



AUC with different illnesses at various image resolutions. [8]



AUC with different CNN models at various image resolutions. [8]

Algorithms	CIFAR-10 (%)
Dropout [41]	84.40
Probout [42]	88.65
NIN + dropout [43]	89.59
Maxout + dropout [44]	88.32
Stochastic pooling [45]	84.86
Probabilistic weighted pooling [46]	88.71
Our method	93.41

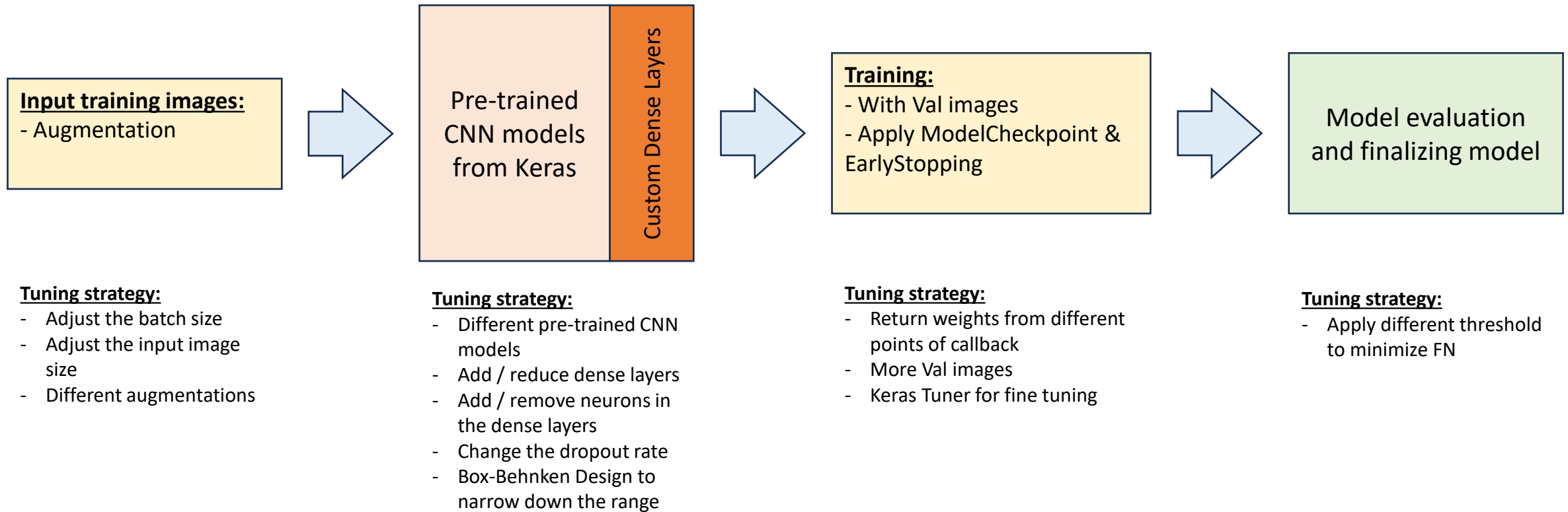
Methods	CIFAR-100 (%)
No data augmentation	61.85
Augmentation in training stage	66.49
Augmentation in testing stage	63.84
Full stage augmentation	70.22

Methods	CIFAR-10 (%)
Translation	91.41
Horizontal flip	90.27
Rotation	88.78
Scale transformation	90.70
Noise disturbance	87.54

Methods	CIFAR-100 (%)
Translation	63.73
Horizontal flip	64.11
Rotation	62.20
Scale transformation	64.83
Noise disturbance	60.47

Improvement with full stage data augmentation framework [9]

# Modelling Strategy



- Model building and tuning will follow the flow illustrated above.

## Pilot: Model Selection

Model	Accuracy	Precision	Recall	F1	AUC	True Normal	False Pneumonia	False Normal	True Pneumonia	Wall Time
<b>DenseNet169</b>	0.910494	0.885167	<b>0.973684</b>	0.927318	0.969	110	24	<b>5</b>	185	21min 16s
DenseNet201	0.873457	0.846512	0.957895	0.898765	0.933	101	33	8	182	16min 43s
VGG16	0.864198	0.837963	0.952632	0.891626	0.924	99	35	9	181	19min 26s
VGG19	0.858025	0.849515	0.921053	0.883838	0.917	103	31	15	175	21min 21s
ResNet50V2	0.839506	0.8	0.968421	0.87619	0.940	88	46	6	184	18min 17s
ResNet152V2	0.882716	0.85514	0.963158	0.905941	0.958	103	31	7	183	20min 35s
MobileNetV3Large	0.756173	0.751131	0.873684	0.807786	0.772	79	55	24	166	14min 27s

- **Recall** is the target metric to be maximized, as the goal is to reduce the number of False Negatives in the diagnosis as much as possible.
- Evidently, DenseNet169 appears to perform the best among the seven selected models.
- The performance results of the models generally align with other studies, where DenseNet and ResNet50 architectures perform better than VGG16 and VGG19 on X-Ray images. [7][8]

### Fixed parameters:

BATCH\_SIZE = 6  
 EPOCHS = 5  
 NEURONS = 512  
 img\_input\_size = 512  
 img\_size = (img\_input\_size, img\_input\_size)  
 img\_shape = (img\_input\_size, img\_input\_size, 3)  
 Threshold = 0.5

### Fixed augmentations (training data only):

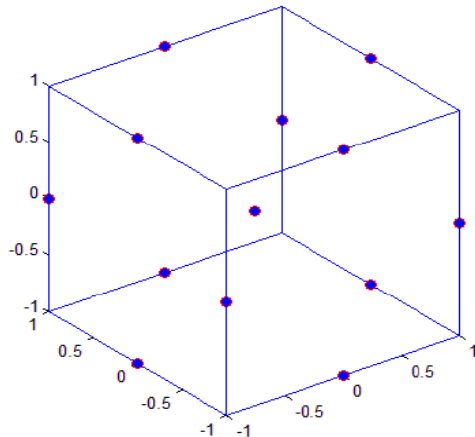
rotation\_range=5,  
 width\_shift\_range=0.1,  
 height\_shift\_range=0.1,  
 horizontal\_flip=True



# Hyperparameters Correlation: Box-Behnken Design

- Deep learning models operate like “black boxes”.
- The input parameters and output responses are modeled with the Design of Experiments (DoE) approach.
  - Minitab software.
  - Understand the correlations between the factors and responses.
  - To obtain a ballpark range of the selected hyperparameters to start with.
- Box-Behnken Design is selected.
  - Popular Response Surface Methodology (RSM).
  - More efficient in terms of the number of experiments required.

	Image Size	Threshold	Neurons
LL	128	0.3	128
HH	640	0.7	1024

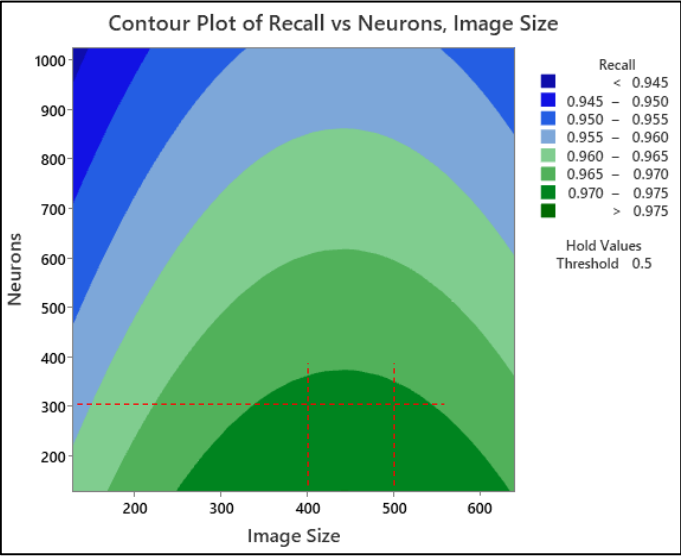
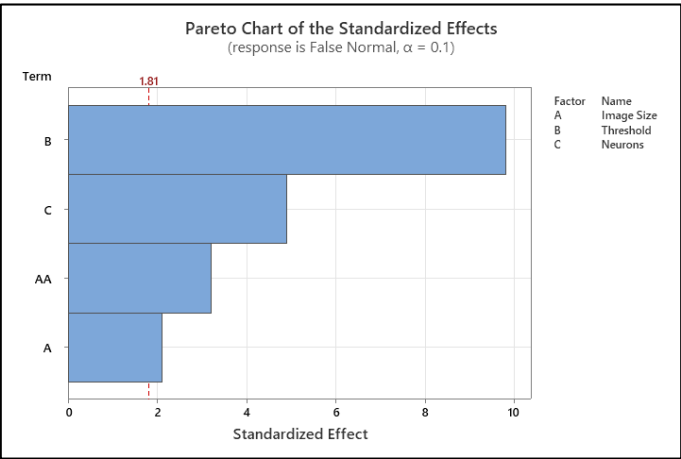
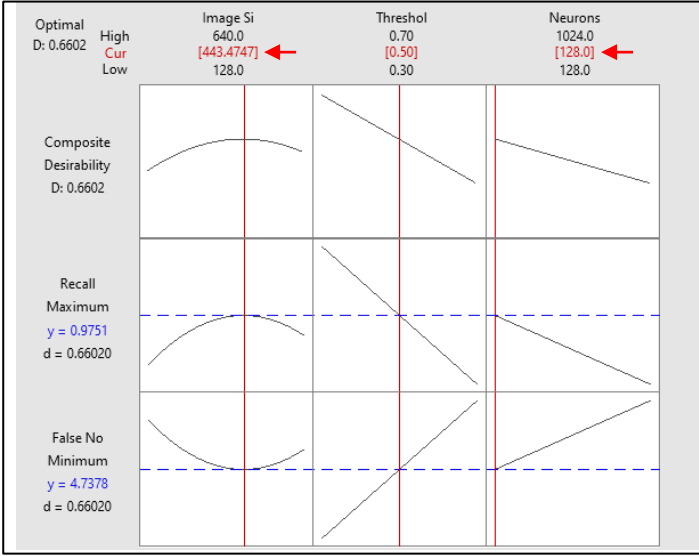


DoE Runs from Box-Behnken Design

	Predictors		
Leg	Image Size	Threshold	Neurons
1	640	0.7	576
2	640	0.3	576
3	384	0.7	1024
4	128	0.5	1024
5	640	0.5	128
6	384	0.3	1024
7	128	0.5	128
8	384	0.5	576
9	128	0.7	576
10	128	0.3	576
11	384	0.5	576
12	384	0.5	576
13	384	0.3	128
14	640	0.5	1024
15	384	0.7	128

Solution						
Solution	Image Size	Threshold	Neurons	Fit	Recall	False Normal
1	443.475	0.5	128	0.975064	4.73779	0.660201

Multiple Response Prediction						
Variable	Setting	Model Summary				
Image Size	443.475	S	R-sq	R-sq(adj)	R-sq(pred)	
Threshold	0.5	0.0053192	93.09%	90.32%	82.73%	
Neurons	128					
Response	Fit	SE Fit	95% CI	95% PI		
Recall	0.97506	0.00271	(0.96902, 0.98111)	(0.96176, 0.98837)		
False Normal	4.738	0.515	(3.590, 5.886)	(2.210, 7.265)		



# Hyperparameters Correlation: Box-Behnken Design

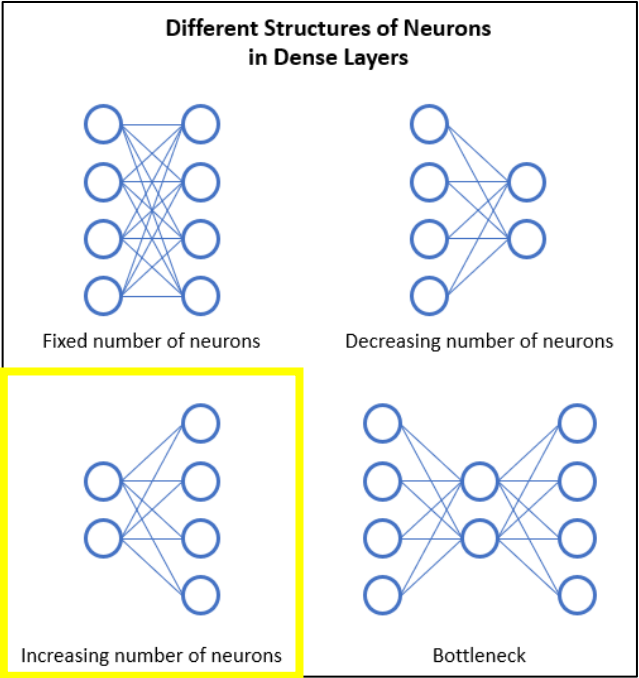
- The following parameters appear to give the highest possible Recall and the lowest possible False Normal.

Image Size	Neurons
443	128

- Recall to be maximized, False Normal to be minimized.
- The regression model fits the data quite well with R-squared value of 93%.
- From the contour plot, the optimal image size appears to fall somewhere **between 400x400 and 500x500**. This result aligns with the findings from other researchers. [8]
- The optimal number of neurons seems to be somewhere **below 300**.
  - The number of neurons for all dense layers is made equal in this case.
- Note: The Box-Behnken Design assumes a quadratic relationship, whereas in Deep Learning (DL) neural networks, the correlation could get very complex and often non-linear.

# Model Optimization: Keras Tuner

- Generally, there are a few types of structures.



- To understand the impact of different image sizes, the number of neurons in the first dense layer, and the number of neurons in the second dense layer, the following hyperparameters are selected.

	Image Size	Neurons 1	Neurons 2
LL	128 x 128	32	32
HH	640 x 640	512	512

## Results:

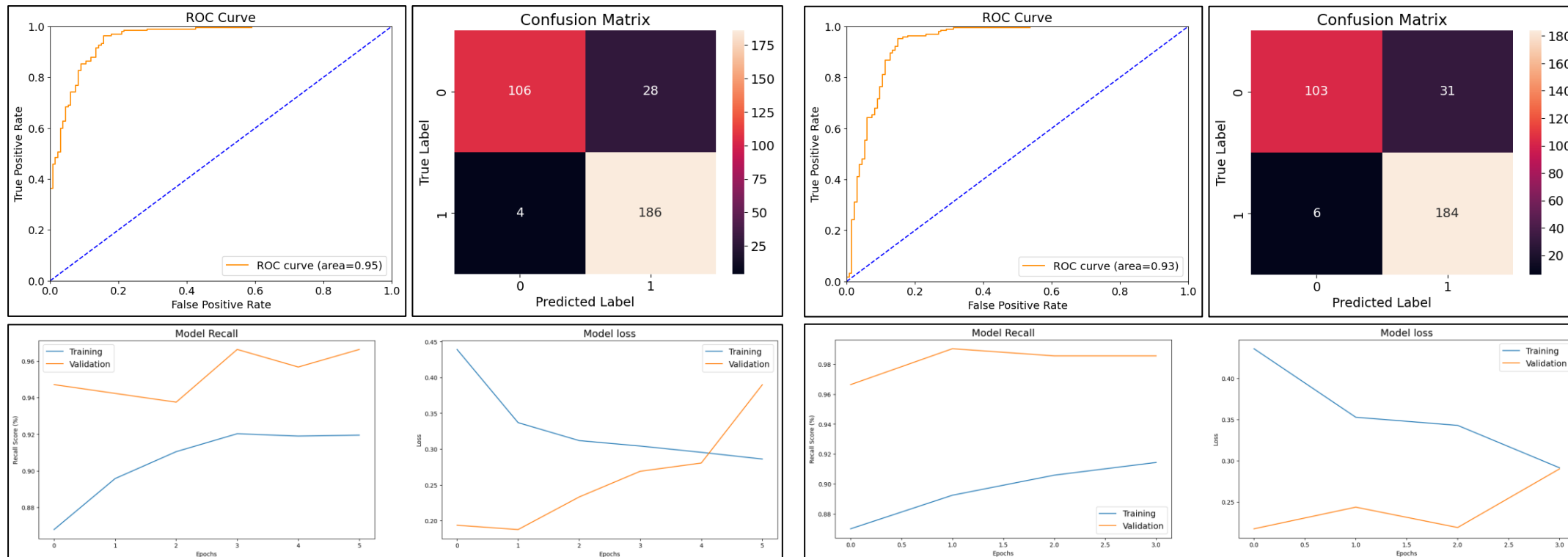
- RandomSearch returns the best hyperparameters from the three trials conducted.

```
Trial 0
Hyperparameters {'image_size': 448, 'neuron_one': 32, 'neuron_two': 224}
Score 0.9855769276618958
-----
Trial 1
Hyperparameters {'image_size': 256, 'neuron_one': 480, 'neuron_two': 160}
Score 0.9855769276618958
-----
Trial 2
Hyperparameters {'image_size': 256, 'neuron_one': 480, 'neuron_two': 288}
Score 0.9839743773142496
-----
```

	Image Size	Neurons 1	Neurons 2
Best	448 x 448	32	224

- Limitation: Only 3 different combinations of hyperparameters were evaluated due to computing limitations.

# Full Stage Data Augmentation



```
# Instantiate data generators
train_datagen = ImageDataGenerator(
    rescale=1./255.,
    rotation_range=5,
    width_shift_range=0.1,
    height_shift_range=0.1,
    horizontal_flip=True
)

val_datagen = ImageDataGenerator(
    rescale=1./255.,
    rotation_range=5,
    width_shift_range=0.1,
    height_shift_range=0.1,
    horizontal_flip=True
)

test_datagen = ImageDataGenerator(
    rescale=1./255.,
    rotation_range=5,
    width_shift_range=0.1,
    height_shift_range=0.1,
    horizontal_flip=True
)
```

Accuracy	Precision	Recall	F1
0.901235	0.869159	0.978947	0.920792

**Without** full stage data augmentation

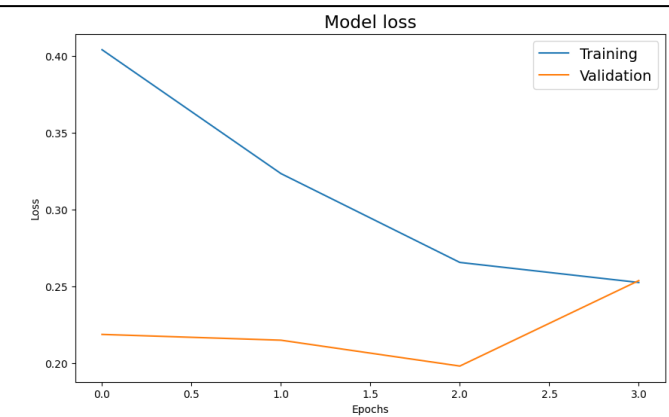
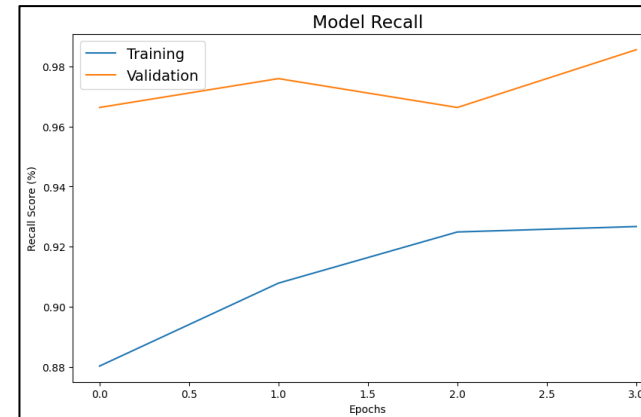
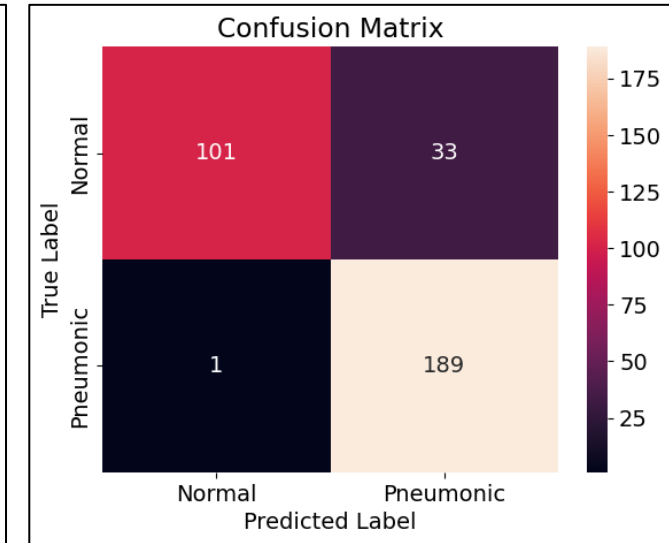
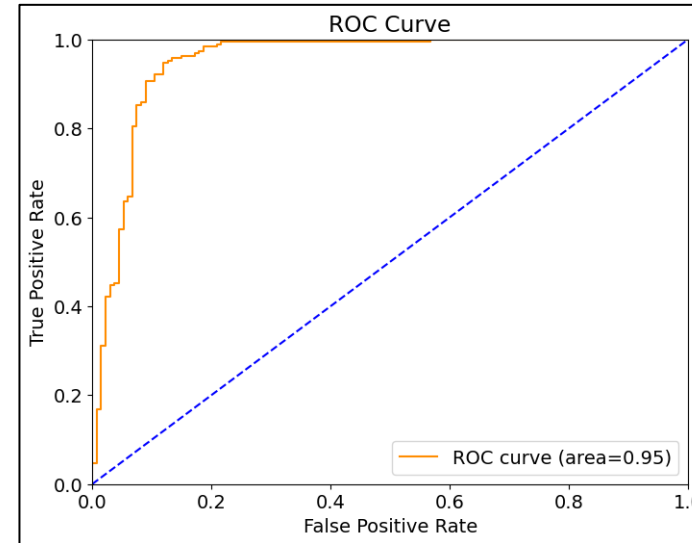
Accuracy	Precision	Recall	F1
0.885802	0.855814	0.968421	0.908642

**With** full stage data augmentation

- Full stage data augmentation: To apply augmentation to validation and test image data as well, rather than just training data.
- Contrary to other studies, the overall results with full stage data augmentation actually perform worse than without it.
  - In other studies, CIFAR-10 and CIFAR-100 datasets were used, which consist of color (RGB) images and much more diversified image categories.
  - The type and magnitude of augmentations used could also influence the outcome.
- However, the validation loss (val\_loss) along the epochs is generally lower (better) than without full stage data augmentation.

# Finalized Model & Discussions

- The model managed to achieve an AUC score of 95% and performs reliably on test images.
- The goal is to minimize False Negatives (Normal) to reduce the risk of misdiagnosing positive cases.
- The best recall score was already achieved at the first epoch with minimal loss.
  - Different hyperparameters were attempted, and all combinations achieved the best recall score in less than five epochs.
  - As a trade-off, False Positives would increase as False Negatives reduce. This is probably acceptable depending on the medications and treatments given to patients who are misclassified as 'positive'.
  - The recall and False Negative rates can be further reduced by adjusting the classification threshold (default = 0.5), if needed. ▶
- The recall score achieved by the model is **99.47%**.



Accuracy	Precision	Recall	F1
0.895062	0.851351	0.994737	0.917476



# Summary & Final Thoughts

- There is randomness when deep neural networks train and make predictions in different runs. It's computationally costly to do multiple runs to confirm results.
- The results align with previous studies that show DenseNet performs quite well on X-Ray images. [7][8]
- Both BBD and Keras Tuner show that an image size of 448x448 performs the best. This also aligns with the findings from other researchers. [8]
- In real-world applications, the desired metrics should be gathered and reviewed together with certified medical professionals, and the model can be further tuned according to the requirements. One of the quickest ways is by adjusting the classification threshold accordingly.
- Full stage data augmentation did not perform as well as in other studies. This could be due to the datasets and/or different ways of implementing augmentations.
- The built model makes reliable classifications which may help in the business problems.

## Future Works:

- Experiment with more hyperparameter tuning.
- Combine different pre-trained CNN models to make predictions.
- Obtain more chest X-Ray image data from different sources and combine them with the existing datasets.



---

# Appendix

---

# References

- [1] H. J. Koo, S. Lim, J. Choe, S. H. Choi, H. Sung, and K. H. Do, "Radiographic and CT Features of Viral Pneumonia," *RadioGraphics: A Review Publication of the Radiological Society of North America, Inc*, vol. 38, no. 3, pp. 719-739, Published Online, May 1, 2018. [Online]. Available: <https://doi.org/10.1148/rg.2018170048>
- [2] "How to Read a Chest Xray II: Pneumonia," *Medchrome*, June 7, 2015. [Online]. Available: <https://medchrome.com/mbbs-exams/how-to-read-a-chest-xray-pneumonia/>. [Accessed: 18- March- 2024].
- [3] "Artificial Intelligence in Radiology," *Siemens Healthineers*, [Online]. Available: <https://www.siemens-healthineers.com/medical-imaging/digital-transformation-of-radiology/ai-in-radiology>. [Accessed: 18- March- 2024].
- [4] L. Berlin, "Faster Reporting Speed and Interpretation Errors: Conjecture, Evidence, and Malpractice Implications," *Journal of the American College of Radiology*, vol. 12, no. 9, pp. 894-896, 2015.
- [5] L. L. Plesner, F. C. Müller, J. D. Nybing, L. C. Laustrop, F. Rasmussen, O. W. Nielsen, M. Boesen, and M. B. Andersen, "Autonomous Chest Radiograph Reporting Using AI: Estimation of Clinical Impact," *Radiology*, vol. 307, no. 3, e222268, Mar. 2023. [Online]. Available: <https://doi.org/10.1148/radiol.222268>.
- [6] F. Pesapane, M. Codari, and F. Sardanelli, "Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine," *European Radiology Experimental*, vol. 2, no. 35, 2018.
- [7] Z. Campos-Lopez, J. Diaz-Roman, B. Mederos-Madrado, N. Gordillo-Castillo, J. Cota-Ruiz, and J. Mejia-Muñoz, "Identification of Pneumonia with X-ray Images Using Deep Transfer Learning," in *XLVI Mexican Conference on Biomedical Engineering, IFMBE Proceedings*, vol. 96, pp. 32-40, Published Online, Oct. 26, 2023. [Online]. Available: [https://doi.org/10.1007/978-3-031-46933-6\\_4](https://doi.org/10.1007/978-3-031-46933-6_4)
- [8] C. F. Sabottke and B. M. Spieler, "The Effect of Image Resolution on Deep Learning in Radiography," *Radiology: Artificial Intelligence*, vol. 2, no. 1, Published Online, Jan. 22, 2020. [Online]. Available: <https://doi.org/10.1148/ryai.2019190015>

# References

- [9] Q. Zheng, M. Yang, X. Tian, N. Jiang, and D. Wang, "A Full Stage Data Augmentation Method in Deep Convolutional Neural Network for Natural Image Classification," Discrete Dynamics in Nature and Society, vol. 2020, Article ID: 4706576, Published Online, Jan. 11, 2020. [Online]. Available: <https://doi.org/10.1155/2020/4706576>
- [10] M. Nagaraju, P. Chawla, and N. Kumar, "Performance improvement of Deep Learning Models using image augmentation techniques," in 1197: Advances in Soft Computing Techniques for Visual Information-based Systems, Multimedia Tools and Applications, vol. 81, pp. 9177-9200, Published Online, Jan. 24, 2022. [Online]. Available: <https://doi.org/10.1007/s11042-021-11869-x>
- [11] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, Article number: 60, Published Online, July 6, 2019. [Online]. Available: <https://doi.org/10.1186/s40537-019-0197-0>

---

# App Demonstration!

- Aims to deploy the deep learning image classification model in **real-world applications** to address the **business problems**.
- A simple application has been created using CustomTkinter. This application could become a potential **business solution** to the challenges faced in healthcare.



---

# End of Presentation

---