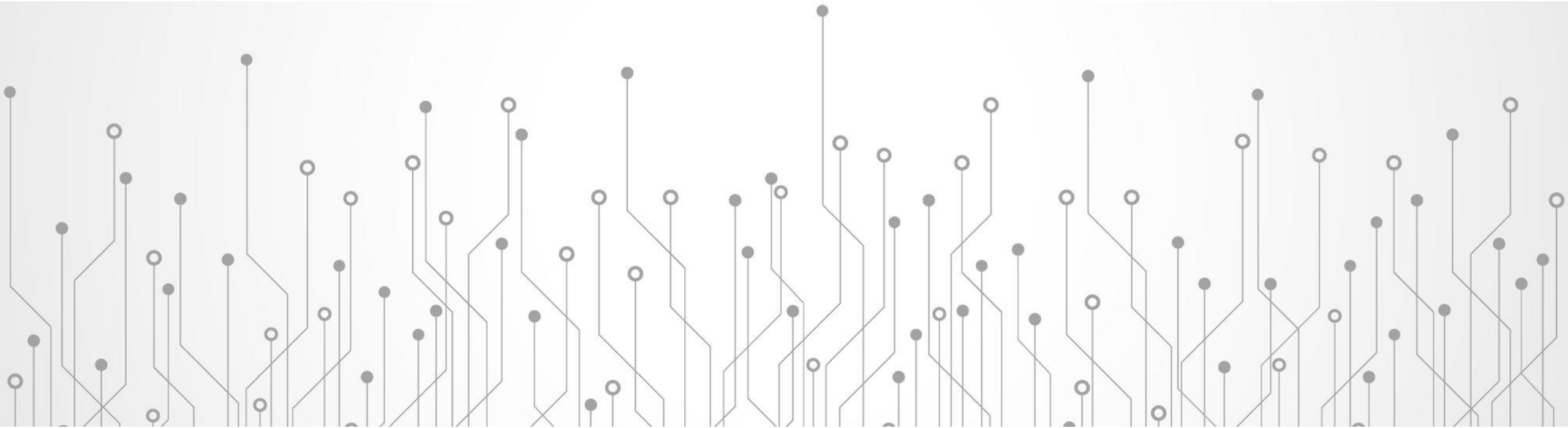


# Data Science

## SHOWCASES



By Kyung Myung Lee

Master of Science in informatics and Analytics

Master of Science in Computer Science and Engineering

Post-Baccalaureate Certificate in Business Analytics

# Contents

## Machine Learning Model

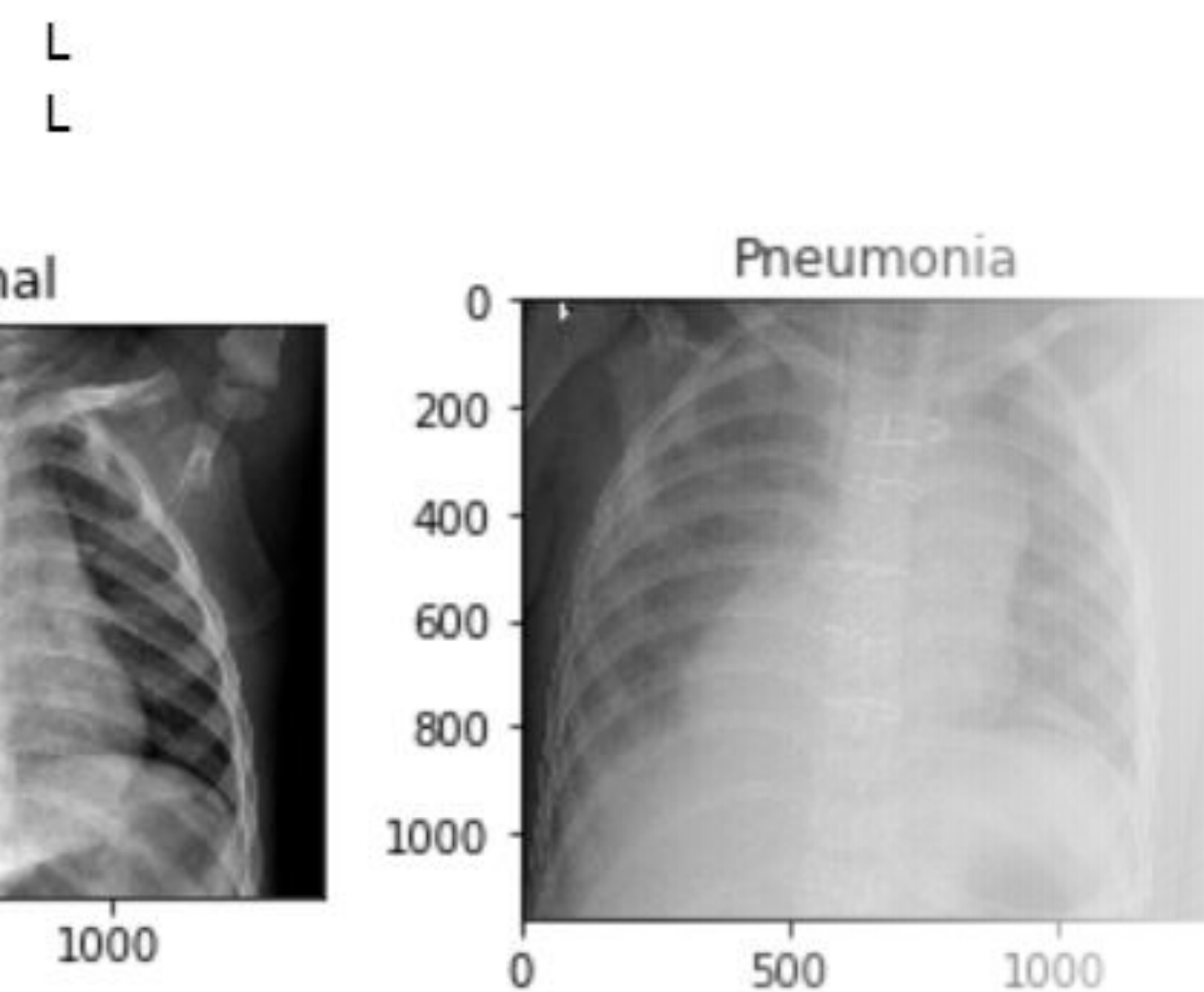
- Pneumonia Detection using Convolutional Neural Network
- Tools & Environment: Python 3, Keras(Tensorflow) on local and Cloud platform(IBM Watson Studio & AWS)

## Statistical Model

- Comparison between cancer centers and genetic counselor utilization using Fisher's exact test and Logistic Regression
- Tool: R Language

❖ Article published (Aug. 2021, co-author (Kyung Lee)) :

Journal of Genetic Counseling,  
<https://onlinelibrary.wiley.com/doi/10.1002/jgc4.1495>



## Pneumonia Detection using Convolutional Neural Network

---

\*X-ray images from Dataset: Kermayn, Daniel; Zhang, Kang; Goldbaum, Michael (2018),  
“Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification”,  
Mendeley Data, v2 <http://dx.doi.org/10.17632/rscbjbr9sj.2> (License: CC BY 4.0)

# Pneumonia Detection

Problem

- Children( age 1 ~ 5) are too young to express symptoms, especially before they get seriously ill.

Model Solution

- Binary image classification
- Earlier pneumonia detection matters.

Dataset

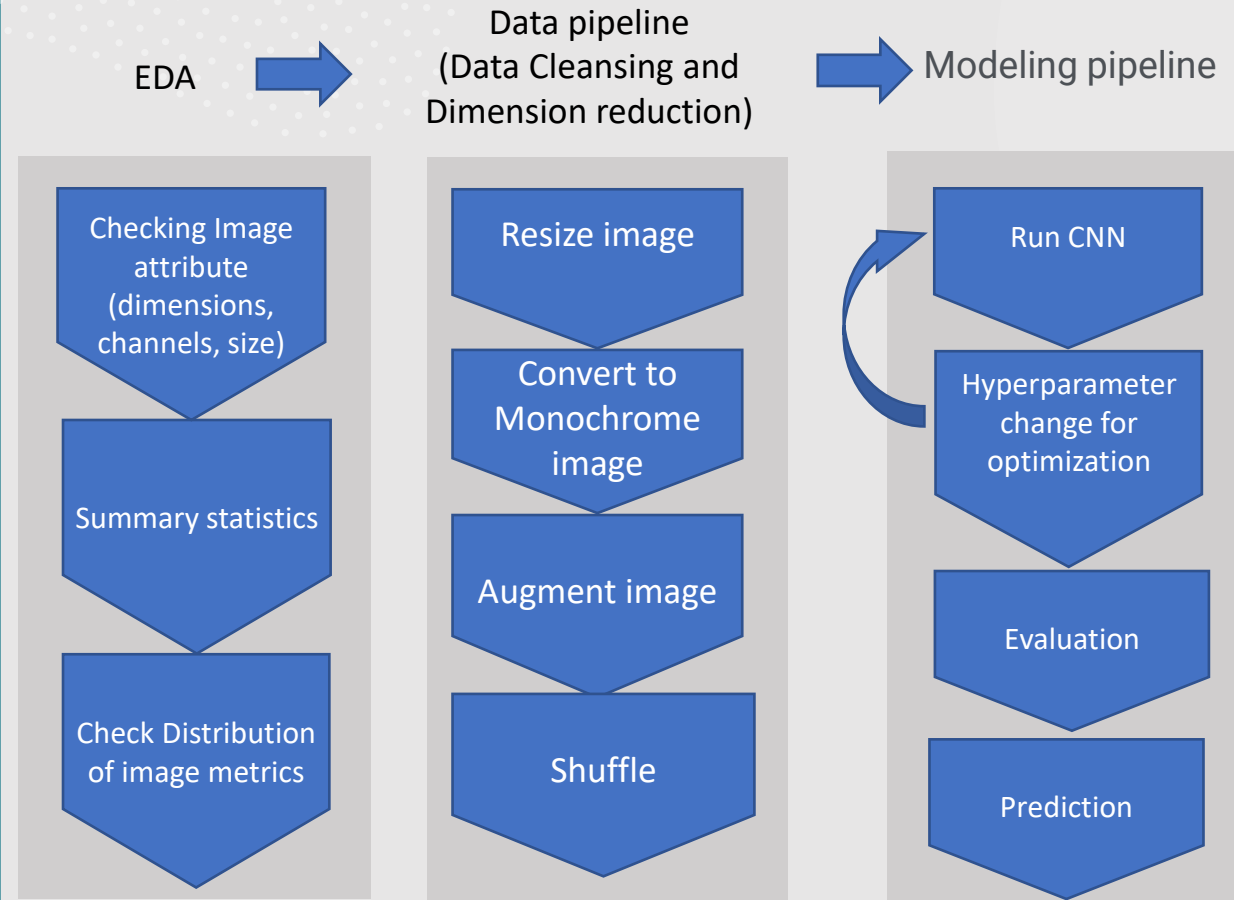
- 5,863 X-ray images (JPEG) and 2 classes (Pneumonia/Normal)

Technique

- CNN(Convolutional Neural Network)

Tools and Env.

- Pandas, numpy, keras(tensorflow) libraries on IBM Watson Studio



# Pneumonia Detection

## Exploratory Data Analysis

**Imbalance in terms of class labels  
(1,341 vs 3,875)**

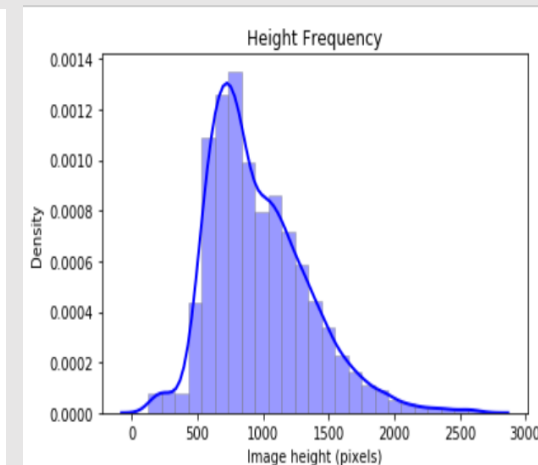
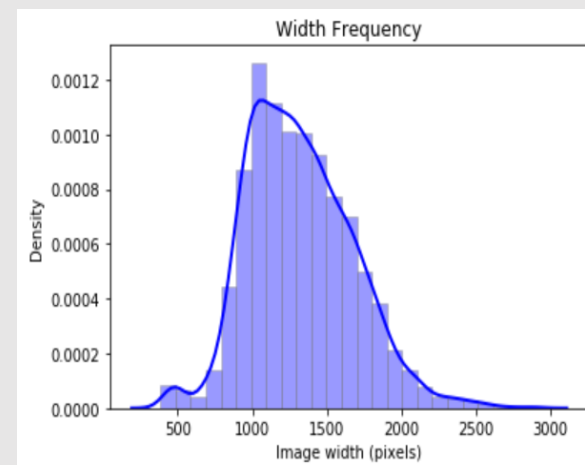
Train Normal dataset:

	width	height
count	1341.000000	1341.000000
mean	1667.734526	1381.431022
std	289.210512	326.320734
min	912.000000	672.000000
25%	1466.000000	1152.000000
50%	1640.000000	1328.000000
75%	1824.000000	1542.000000
max	2916.000000	2663.000000

Train Pneumonia dataset:

	width	height
count	3875.000000	3875.000000
mean	1200.483613	825.026839
std	291.305676	277.073758
min	384.000000	127.000000
25%	1000.000000	640.000000
50%	1168.000000	776.000000
75%	1368.000000	968.000000
max	2772.000000	2304.000000

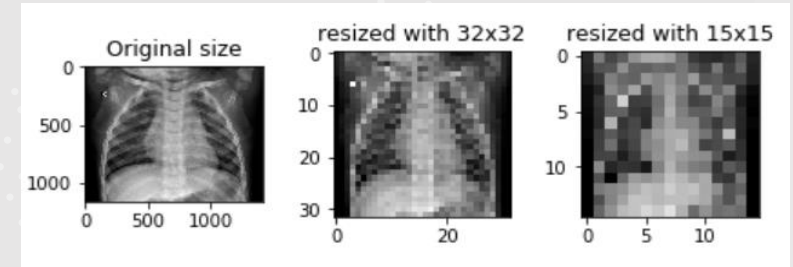
**Distribution of the frequency of width and  
height size (unit:pixel)**



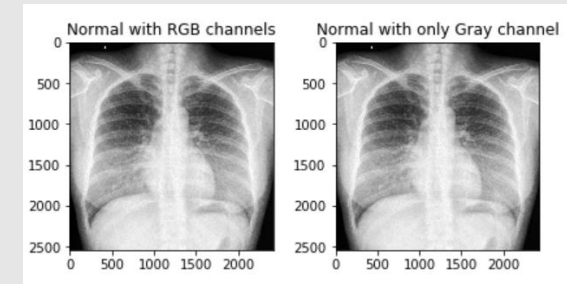
# Pneumonia Detection

## Data Cleansing

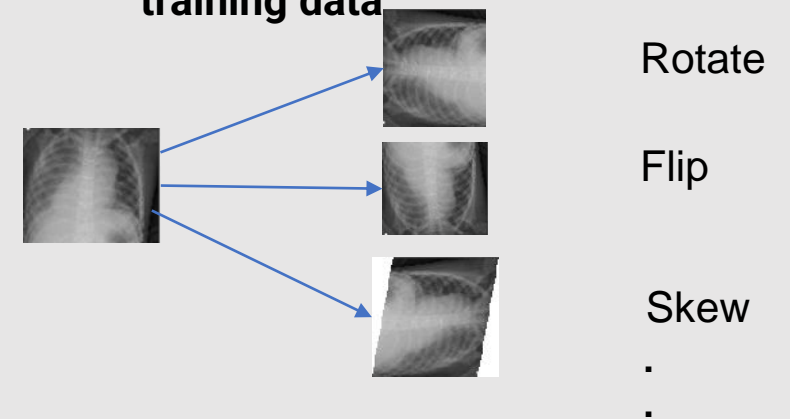
### Resizing Images



### Converting to Gray



**Image augmentation for effect of random sampling only to training data**



# Pneumonia Detection

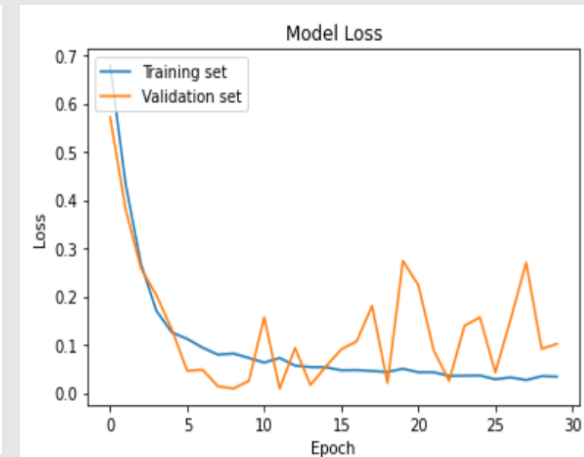
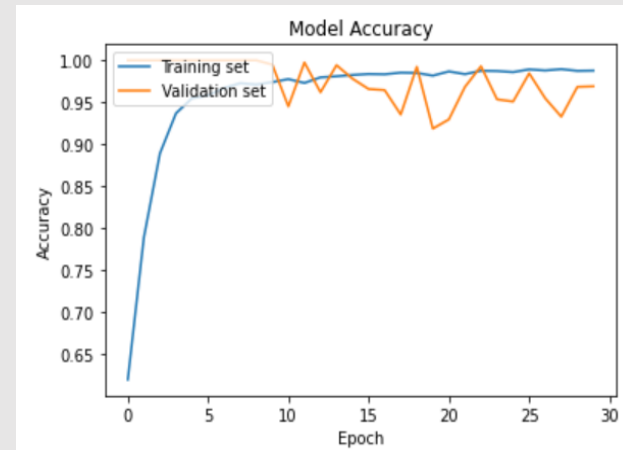
## Modeling

## Convolutional Neural Network

Model Implementation:

[https://github.com/kmleeDS/portfolio/blob/main/Showcase\\_attached\\_to\\_JobKorea\\_CNN\\_32x32\\_Images.ipynb](https://github.com/kmleeDS/portfolio/blob/main/Showcase_attached_to_JobKorea_CNN_32x32_Images.ipynb)

## Accuracy and Loss



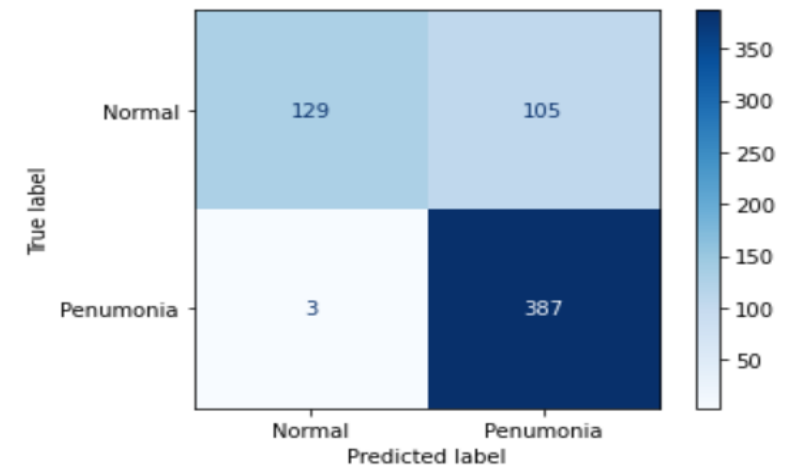
# Pneumonia Detection

## Model Evaluation

Model	accuracy with test dataset	F1 Score	Training vs Test data ratio (80 : 20)
Model 1 (32 x 32 )	83%	88%	7,750 vs 155
Model 2 (60 x 60 )	76%	84%	7,750 vs 155

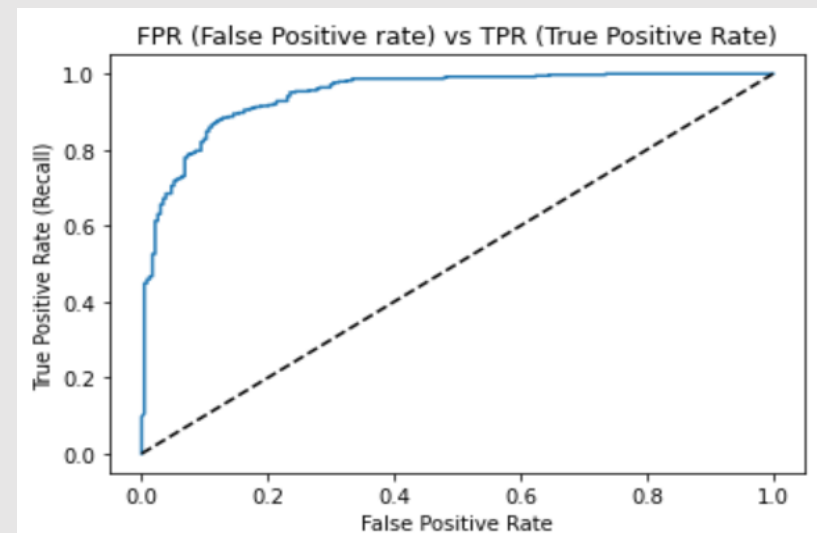
\*Model 1 is a winner with f1, 88% and type II error is very low with only 3 obs. according to the Confusion Matrix

Confusion Matrix (Model 1)



accuracy of chest X-ray for pneumonia:0.83  
Precison of chest X-ray for pneumonia:0.79  
Recall of chest X-ray for pneumonia:0.99  
f1 of chest X-ray for pneumonia:0.88

ROC with ACU curve (Model 1)





# Pneumonia Detection

## Model application and future work


### Model application

- It can be utilized as a pre-diagnosis tool before doctors' confirmed diagnoses
- Proactive treatment is possible because the model is able to quickly identify a patient with pneumonia once the x-ray image is ready to use

### Future work

Apply better high dimensionality reduction techniques

Fine-tune hyperparameter to get optimal parameters of the model



## **Comparison between cancer centers and genetic counselor utilization**

**(Research with Cone Health  
Cancer Center)**

---



# Following NCCN guidelines within one hospital system in the United States: Comparison between cancer centers and genetic counselor utilization

Journal of Genetic Counseling, co-author),  
<https://onlinelibrary.wiley.com/doi/10.1002/jgc4.1495>


Received: 12 February 2021 | Revised: 26 July 2021 | Accepted: 28 July 2021

DOI: 10.1002/jgc4.1495

National Society of  
Genetic Counselors  
WILEY

## ORIGINAL ARTICLE

# Following NCCN guidelines within one hospital system in the United States: Comparison between cancer centers and genetic counselor utilization

Karen Powell<sup>1</sup>  | Jonathan Rakestraw<sup>2</sup> | Sat Gupta<sup>3</sup> | Wenhao Shou<sup>3</sup> | Kyung Lee<sup>4</sup> | Ofri Leitner<sup>5</sup>

<sup>1</sup>Genetic Counseling Program, Cone Health Cancer Center, Greensboro, NC, USA

<sup>2</sup>Oncology Informatics System, Cone Health Cancer Center, Greensboro, NC, USA

<sup>3</sup>Department of Mathematics and Statistics, The University of North Carolina, Greensboro, NC, USA

<sup>4</sup>Informatics and Analytics Program, The University of North Carolina, Greensboro, NC, USA

<sup>5</sup>Skypax, Chapel Hill, NC, USA

### Correspondence

Karen Powell, Genetic Counseling Program, Cone Health Cancer Center, Greensboro, NC, USA.

Email: Karen.powell@conehealth.com

### Abstract

Genetic testing is an instrumental tool used to determine whether an individual has a predisposition to certain cancers. Knowing of a hereditary cancer predisposition may allow a patient and their family to consider high-risk screening or risk-reducing options. Genetic counselors work with physicians to identify patients at increased risk for genetic testing using available guidelines such as those provided by the National Comprehensive Cancer Network (NCCN). Information within one hospital system's cancer registry was used to identify individuals who qualify for genetic testing. This includes patients with a history of cancer of the breast (diagnosis  $\leq 45$ , triple negative (TN)  $\leq 60$ , and male), ovaries, colon (diagnosis  $\leq 50$ ), or uterus (diagnosis  $\leq 50$ ). Within this hospital system's registry, there are six cancer centers. Data were collected from cancer centers that utilized genetic counselors (GCs), and cancer centers that did not (non-GC) to determine whether there was a difference in genetic testing rates between GC and non-GC cancer centers. An analysis of 695 patients demonstrated a significantly higher proportion of eligible patients undergoing genetic testing at the GC cancer centers than at the non-GC cancer centers (91.6% versus 68.7%,  $p < .001$ ). Further analysis of specific cancers showed a significantly higher uptake of genetic testing for eligible patients with colon cancer (90.8% versus 50%,  $p < .001$ ), breast cancer  $\leq 45$  (99.5% versus 86%,  $p < .001$ ), and ovarian cancer (91.3% versus 62.8%,  $p < .001$ ) at the GC cancer centers than at the non-GC cancer centers. There was no significant difference in the proportion of testing of TN breast cancer  $\leq 60$  or uterine cancer  $\leq 50$  between cancer centers. These data suggest that having a GC working within a cancer center increases the ability to identify and offer testing to patients who meet NCCN genetic testing criteria based on their cancer type.

### KEYWORDS

cancer, genetic counselor, genetic services, genetic testing, NCCN guidelines