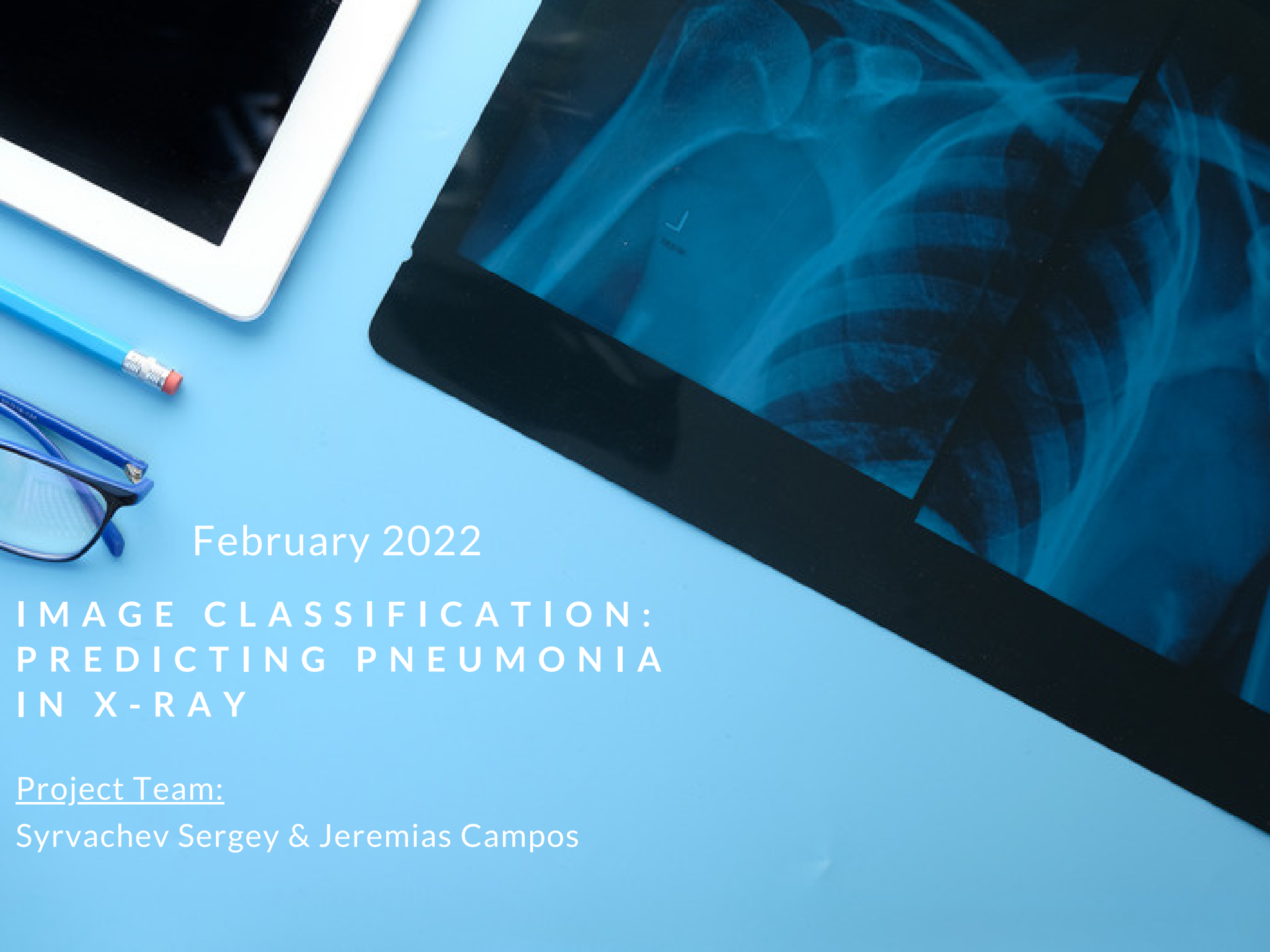


February 2022

# IMAGE CLASSIFICATION: PREDICTING PNEUMONIA IN X-RAY


Project Team:

Syrvachev Sergey & Jeremias Campos

**AS TO DISEASES, MAKE A HABIT OF TWO THINGS — TO HELP, OR AT LEAST, TO DO NO HARM.**

– Hippocrates

# PNEUMONIA

- <u>What is it?:</u>
  - Infection that inflames air sacs in one or both lungs, which may fill with fluid.
- 1.5 million+ US cases per year (VERY COMMON)
- 40,000 Death each year
- Lab tests or imaging are always required

Source: CDC
https://www.cdc.gov/

| 1 | 2 | 3 |
| --- | --- | --- |
| INTRODUCTION | RESEARCH | FINALIZE |

- Business Problem
- Data Understanding
- Metrics

- Raw Data
- Modeling
- Modeling Results

- Business recommendations
- Next Steps

# INTRODUCTION

- Business problem
- Key Idea
- Data sources & Methods

# BUSINESS PROBLEM

- Our team was hired by an unnamed local hospital **to create a model which automatically classifies if a patient has pneumonia by looking a x-ray.**

- 10-15% of all diagnoses are incorrect.[1]
- 1 in 3 misdiagnoses result in serious injury or death.[2]

- Our model should provide over 90% accuracy and 95% recall in **predicting pneumonia.**

**Source:**
[1] BMJ journal
https://qualitysafety.bmj.com/content/22/Suppl_2/ii21.info
[2] Johns Hopkins University School of Medicine
https://www.degruyter.com/document/doi/10.1515/dx-2019-0019/html

# KEY IDEA BEHIND

- Fundamental approach:

“Good health and effective medical care are essential for a society's ability to function.”

- Philosophy:

“The philosophy of healthcare is the study of the ethics, processes, and people which constitute the maintenance of health for human beings.”

# DATA UNDERSTANDING

## METHODS:

- Image recognition using Deep Learning techniques:
  - Fully Connected Neural Networks
  - Convolutional Neural Networks
  - Pretrained Convolutional Neural Networks

## DATA SOURCES:


- Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), "Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images", Mendeley Data https://data.mendeley.com/datasets/rscbjbr9sj/3

# METRICS

We are focused on finding people with pneumonia.

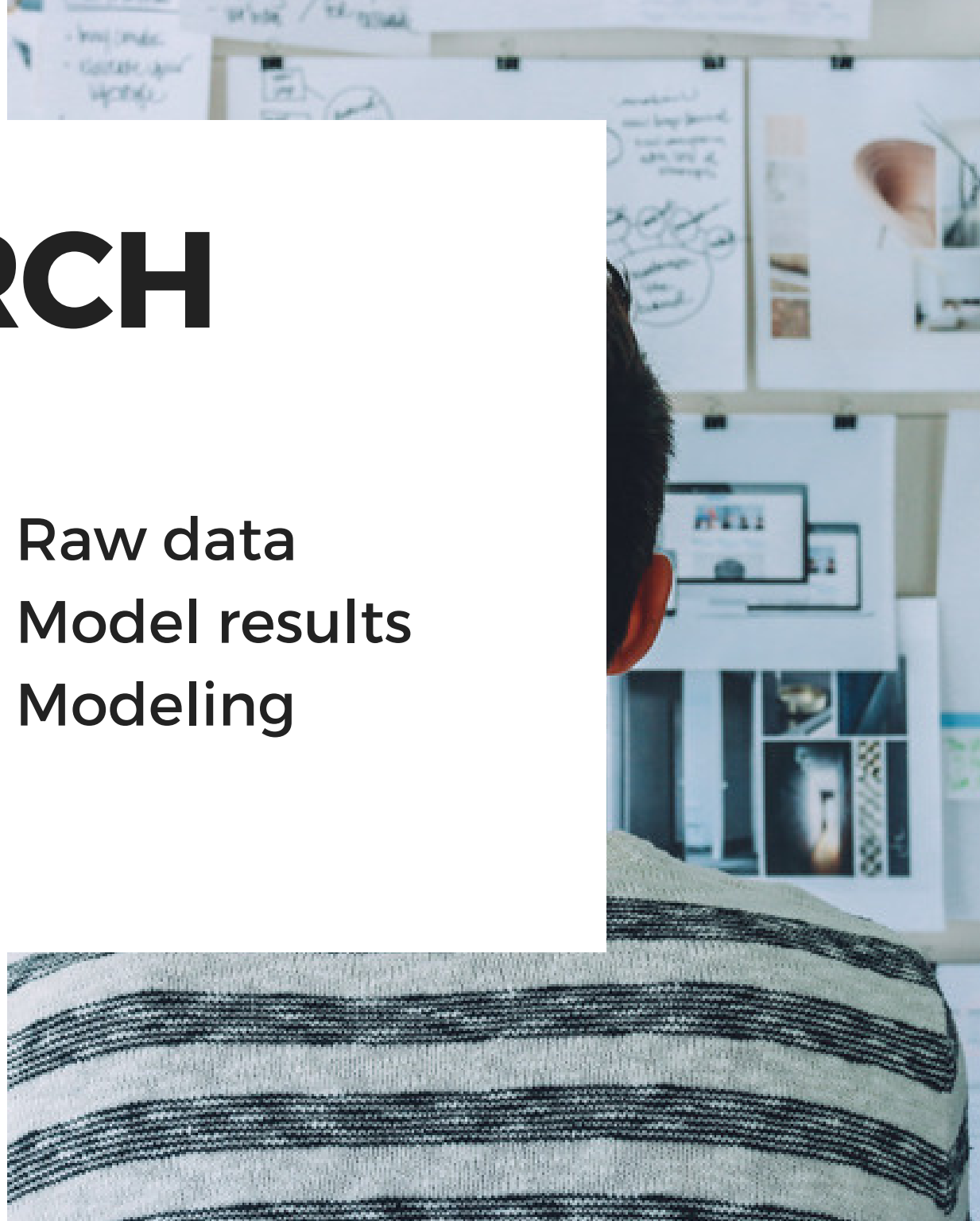
- **Hypothesis:**
  - H0 - Person has pneumonia.
  - HA - There is statistically significant proof that the person doesn't' have pneumonia.

# USED METRICS:

- **Recall** - Health of people is our priority, we will be focused on finding pneumonia cases.
  - **Business requirement**: our system should have at least 95% recall.
- **Accuracy** - How accurate our results are considering false identified cases.

# RESEARCH

- Raw data
- Model results
- Modeling

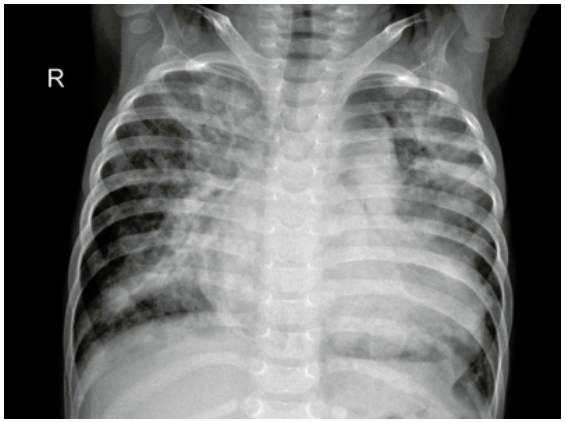


# RAW DATA

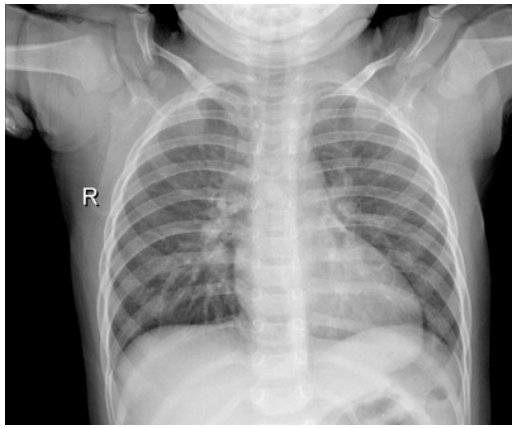


- TOTAL 6462 X-RAY IMAGES
  - 4879 HAS PNEUMONIA
  - 1583 ARE NORMAL

PNEUMONIA X-RAY




NORMAL X-RAY

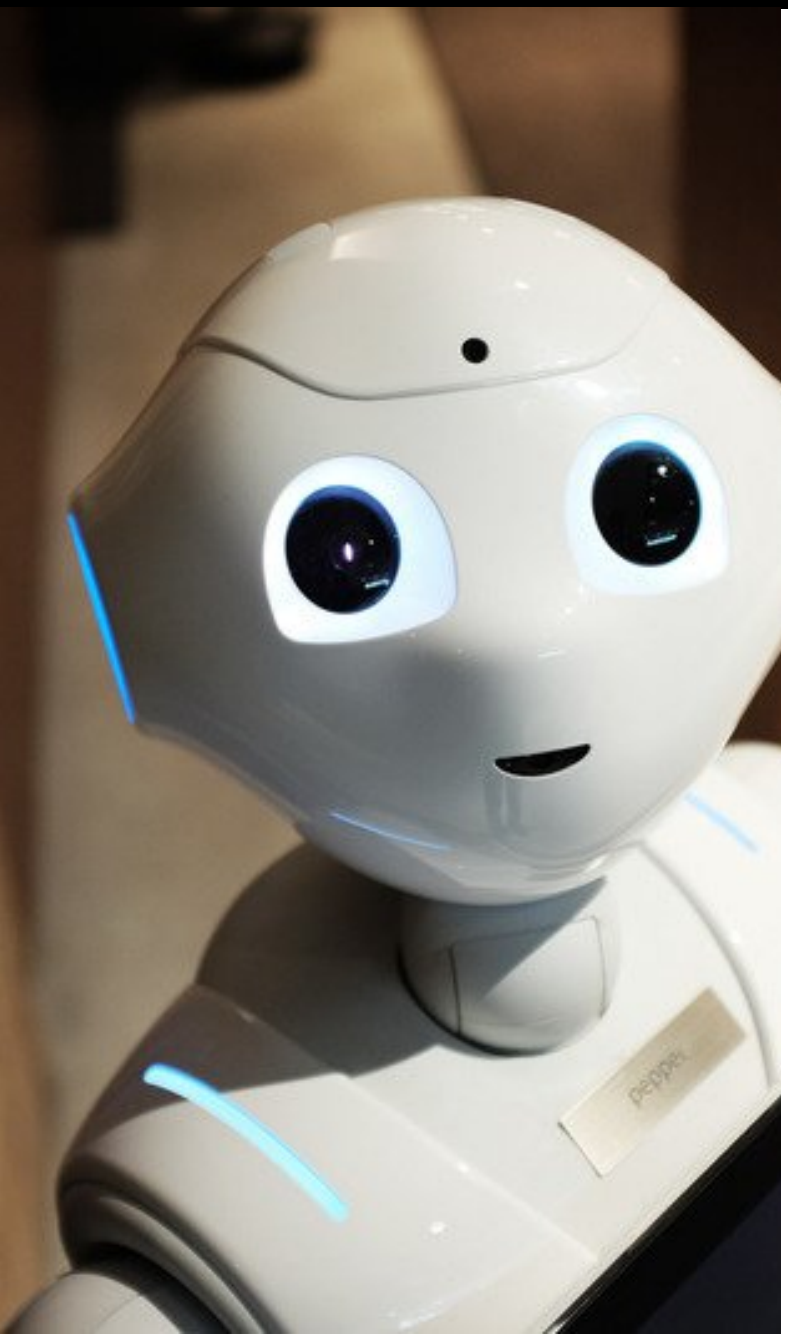


# MODEL RESULTS



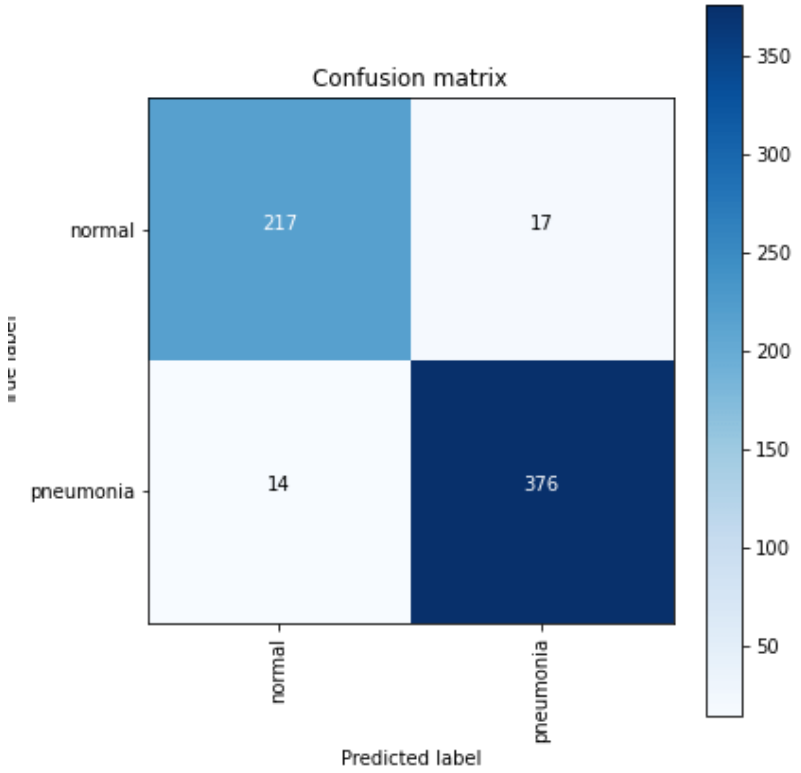
| Model name | Image size | Parameters | Train time | Train accuracy | Validation accuracy | Test accuracy | Test Recall |
|---|---|---|---|---|---|---|---|
| Baseline_model | (64, 64) | 614971 | 12.0 | 0.9514 | 0.4834 | 0.6266 | 0.9915 |
| Baseline with reg | (64, 64) | 615996 | 31.0 | 0.9061 | 0.7749 | 0.625 | 0.0 |
| Baseline with reg, dropout | (64, 64) | 615996 | 31.0 | 0.7689 | 0.7691 | 0.625 | 0.0 |
| Baseline with reg, dropout, optimizer | (64, 64) | 615996 | 40.0 | 0.7689 | 0.7691 | 0.625 | 0.0 |
| Baseline with reg, dropout, optimizer + extra ... | (64, 64) | 615996 | 83.0 | 0.7689 | 0.7691 | 0.625 | 0.0 |
| Basic CNN | (64, 64) | 107553 | 456.0 | 0.9609 | 0.9486 | 0.7837 | 0.4444 |
| Basic CNN with reg, dropout | (64, 64) | 72705 | 293.0 | 0.9688 | 0.8869 | 0.6763 | 0.1368 |
| Basic CNN 100x100 | (100, 100) | 138241 | 617.0 | 0.7689 | 0.7691 | 0.625 | 0.0 |
| Basic CNN 100x100 with reg and dropout | (100, 100) | 173089 | 756.0 | 0.7689 | 0.7691 | 0.625 | 0.0 |
| Augmented CNN 100x100 with reg | (100, 100) | 376421 | 1359.0 | 0.8404 | 0.8562 | 0.5016 | 0.6282 |
| Augmented CNN 100x100 with reg and RMSprop opt... | (100, 100) | 390757 | 920.0 | 0.7626 | 0.785 | 0.383 | 0.9658 |
| Augmented CNN 100x100 with additional reg, inc... | (100, 100) | 390757 | 2736.0 | 0.8854 | 0.9038 | 0.4615 | 0.6923 |
| Pre-trained Augmented CNN 100x100 frozen layer... | (100, 100) | 15009729 | 1476.0 | 0.8939 | 0.9375 | 0.4631 | 0.7308 |
| Pre-trained Augmented CNN 100x100 frozen layer... | (100, 100) | 15009729 | 1132.0 | 0.9262 | 0.9217 | 0.5385 | 0.6692 |
| Pre-trained Augmented CNN 100x100 frozen layer... | (100, 100) | 15009729 | 997.0 | 0.9231 | 0.936 | 0.8333 | 0.9846 |
| Pre-trained Augmented CNN 224x224 frozen layer... | (200, 200) | 6272193 | 1318.0 | 0.9438 | 0.932 | 0.9503 | 0.9641 |
| Pre-trained Augmented CNN 224x224 frozen layer... | (200, 200) | 16320449 | 4090.0 | 0.94 | 0.93 | 0.9263 | 0.9821 |
| Pre-trained Augmented CNN 200x200 frozen layer... | (200, 200) | 15894465 | 3486.0 | 0.9394 | 0.94 | 0.9279 | 0.9744 |

BEST MODEL:
PRETRAINED CONVOLUTIONAL NEURAL NETWORK MOBILENETV2



## BEST MODEL BEFORE TUNING:

- PRETRAINED CONVOLUTIONAL NEURAL NETWORK MOBILENETV2
  - 96.41% RECALL
  - 95.03% ACCURACY



## 3 TYPES OF TUNING:

DEFAULT MODEL:

- 96.41% RECALL
- 95.03% ACCURACY

MAX RECALL

- 99.49% RECALL
- 91.35% ACCURACY

MAX ACCURACY:

- 97.69% RECALL
- 95.19% ACCURACY

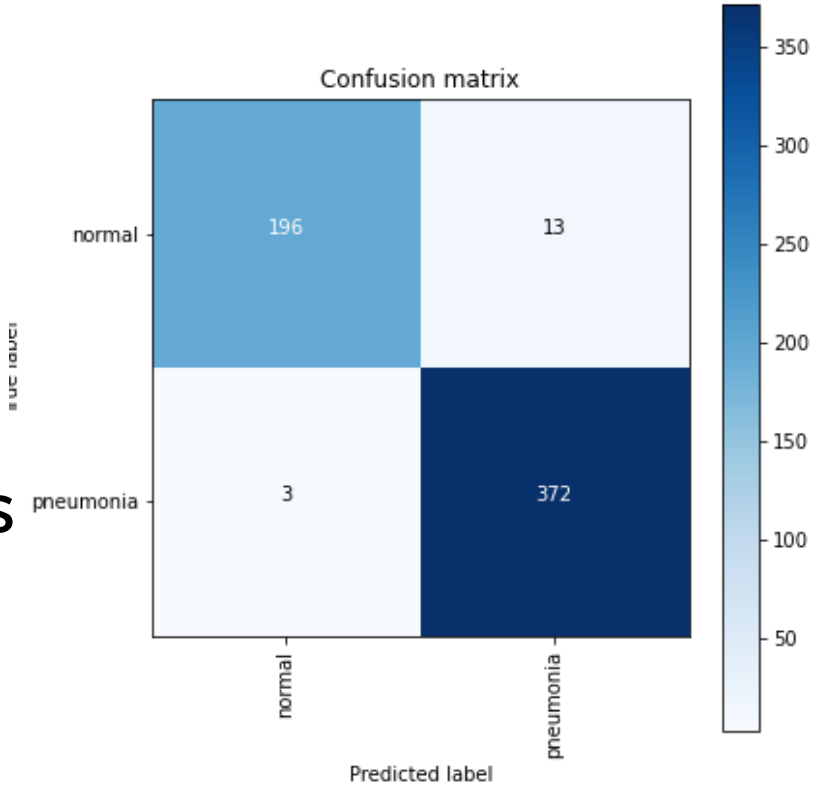
## SUSPICIOUS CASE REMOVAL APPROACH:

MAX ACCURACY OPTIMIZED MODEL:

- 97.69% RECALL
- 95.19% ACCURACY

SUSPICIOUS CASES REMOVED

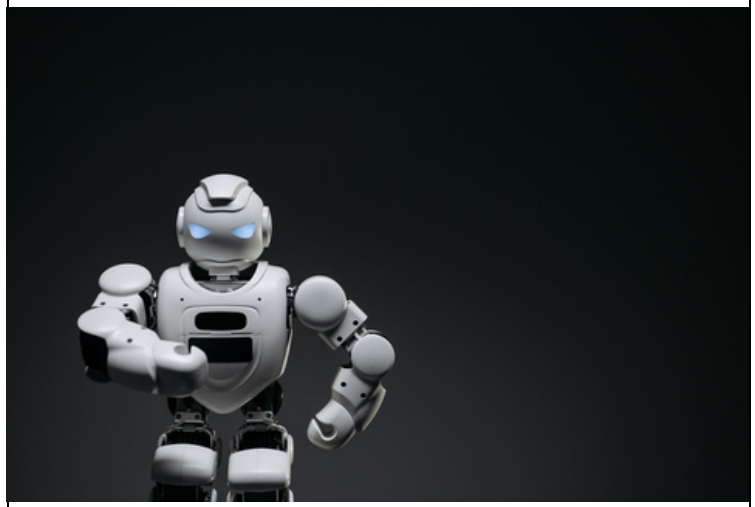
- 99.92% RECALL
- 97.26% ACCURACY
- 6.41% SUSPICIOUS CASES



# FINALIZE

- Business recommendations
- Next Steps
- QA

# Recommendations



01

**MODEL OPTIMIZATION**

Based on hospital treatment policy model should be tuned

02

**USE OF MODEL**

The model gives less error than doctors

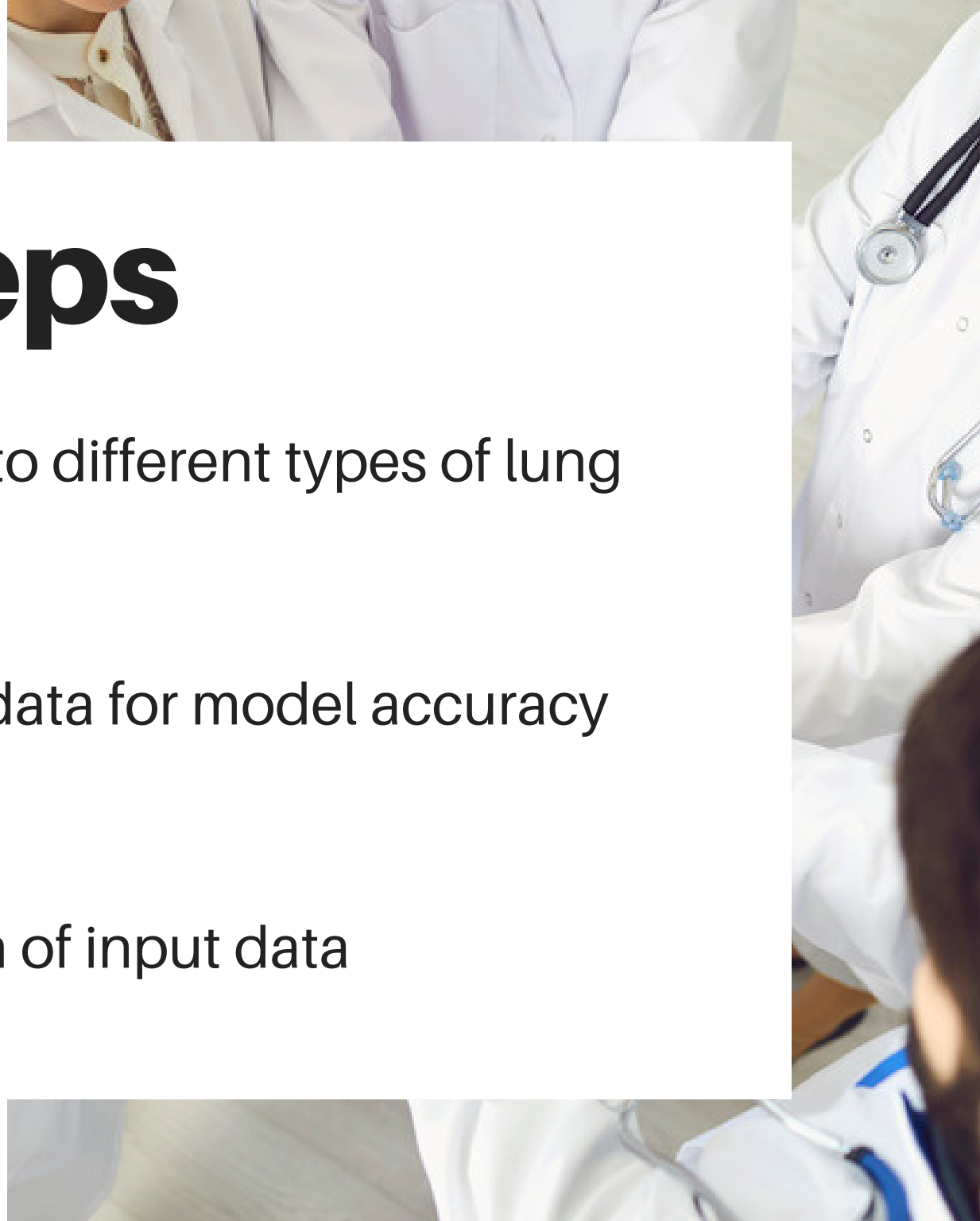
03

**INFORMATION INPUTS**

Model can be extended to diagnose other lungs illnesses



MEDICAL IMAGE CLASSIFICATION



# Next steps

- Extend the model to different types of lung desease.
- Gather additional data for model accuracy improvement
- Double verification of input data

# Q & A:

## Thank you for joining today's presentation.


SYRVACHEV SERGEY

- DATA SCIENTIST
- DEST.STUDIO@GMAIL.COM
- LINKEDIN: /SSYRVACHEV
- GITHUB: 314KA4Y
- MEDIUM: @SERGEYSYRVACHEV

- JEREMIAS CAMPOS
- DATA SCIENTIST
- JEREMIASCAMPOS3@GMAIL.COM
- LINKEDIN: /JEREMIASCAMPOS
- GITHUB: DATAJCAMPOS
- MEDIUM: @JEREMIASCAMPOS3