A Classification Tool for COVID-19 and Bacterial Pneumonia Diagnosis using X-ray Imaging Data

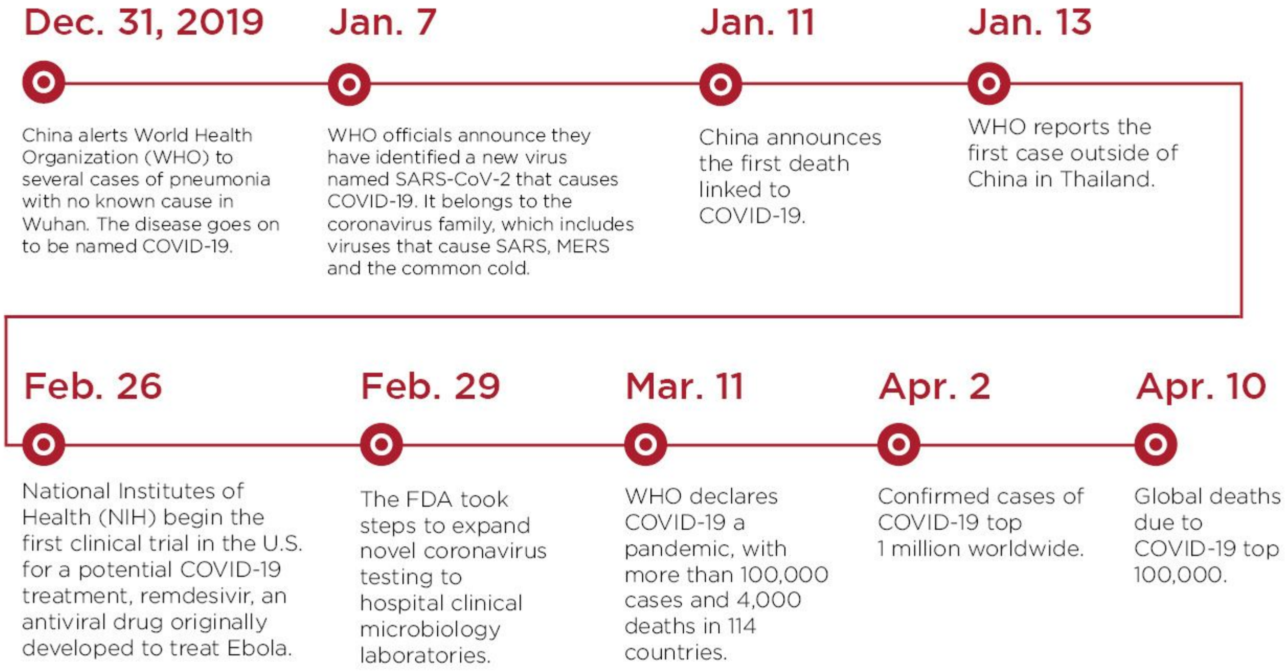
CHL7001

Vivian Ngo
Department of Statistical Sciences
University of Toronto

Andrew Hoan Tran
Department of Biostatistics
Dalla Lana School of Public Health

Introduction

Motivation



Source: https://asm.org/Press-Releases/2020/COVID-19-Resources

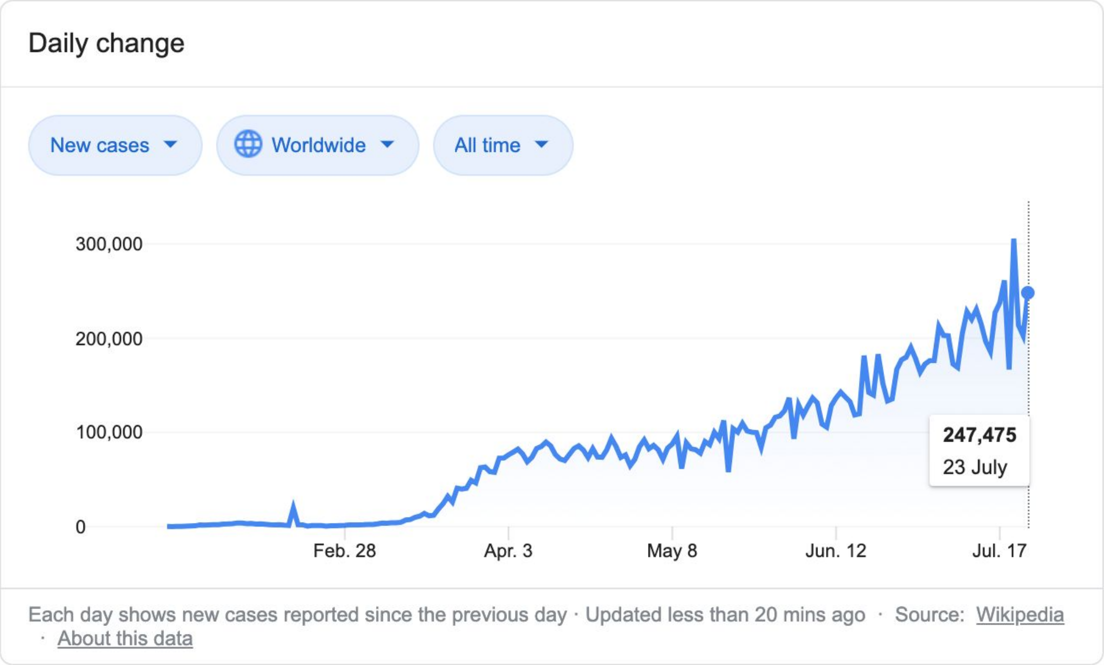
Motivation

Challenges:

- Diagnosis/Testing
- Limited resources
- COVID-19 vs bacterial pneumonia

Ideas:

- Imaging data,
- Computer-based methods



This Project

- **Objective:** Differentiate between healthy individuals, individuals with bacterial pneumonia, and individuals with COVID-19.
- **Idea:** Use deep learning techniques to build a multi-class classifier for these three categories

Data and Methods

Data

COVID:

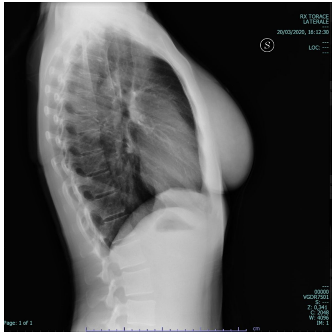
- X-rays and CT scans of patients with COVID-19
- Roughly 200 images
- https://github.com/ieee8023/covid-chestxray-dataset

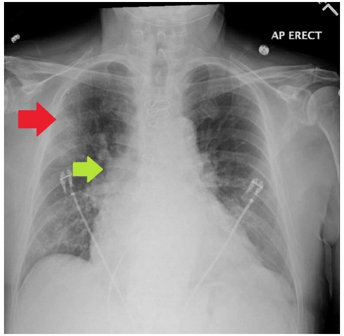
Bacterial Pneumonia + Healthy:

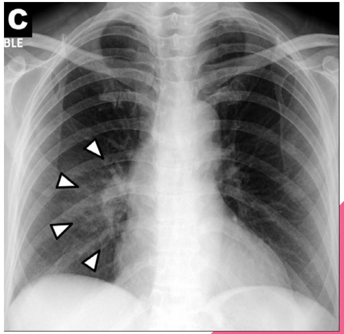
- Chest x-rays for individuals with and without pneumonia
- Over 5,000 images all together
- https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

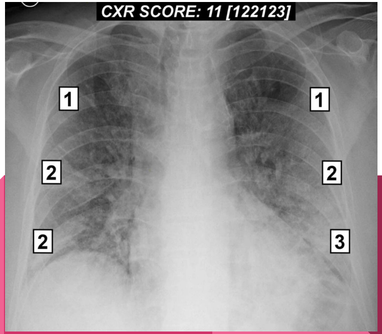
Data Processing

- Filter only x-rays from COVID dataset
- Remove MERS and SARS from COVID dataset
- Remove images that are wrong view or annotated
- Universal crop of all images







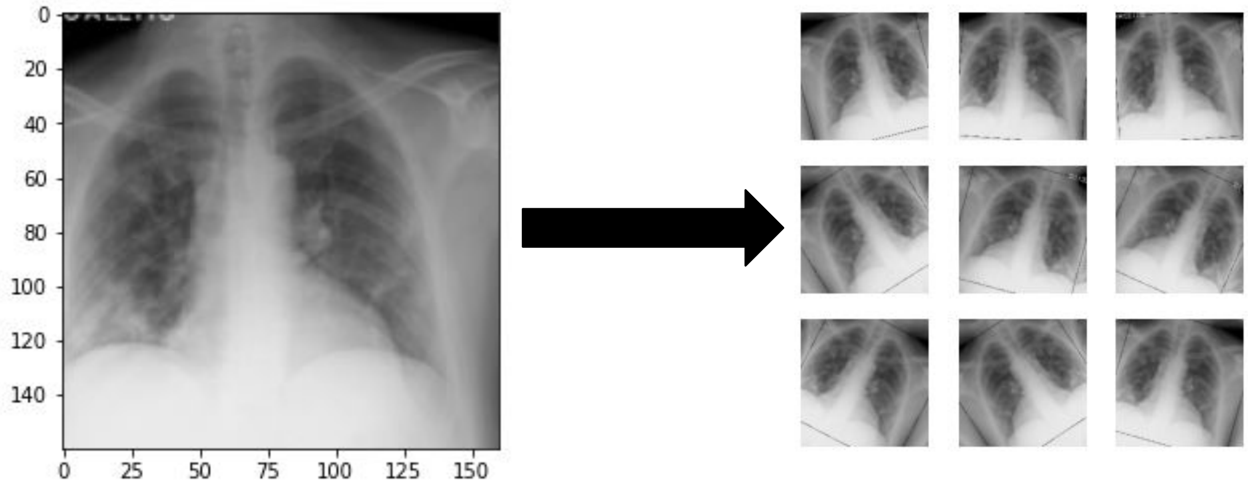


Methods

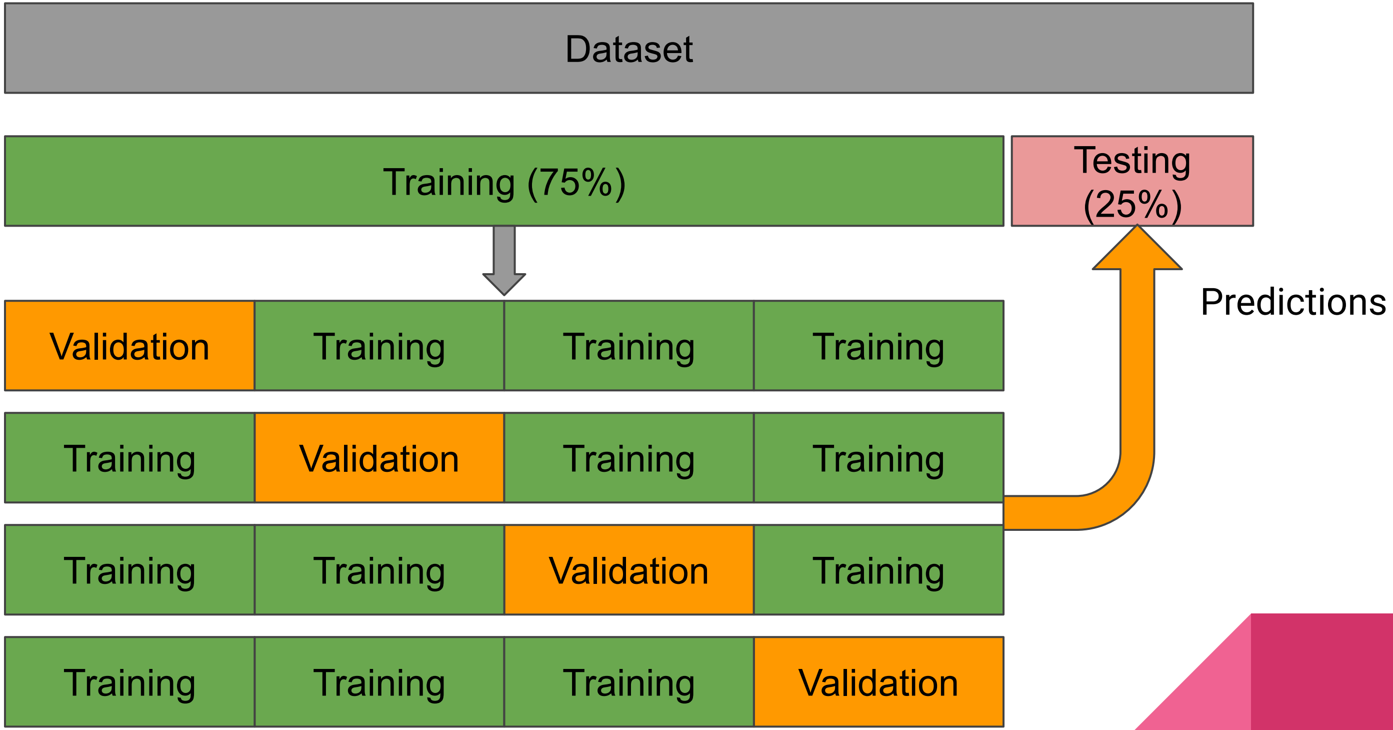
- Multi-class classification with class imbalance
- Limited COVID images:
  - Transfer learning
  - Data augmentation
  - Synthetic Minority Oversampling Technique (SMOTE)
  - Class Weights
- K-fold cross validation (stratified)

Data Augmentation

- Random horizontal flips
- Random rotations up to 36 degrees clockwise or counter-clockwise



Stratified K-fold cross validation



Evaluation

Macro average:

- Recall
- Precision
- F-score

$$Recall_M = \left(\Sigma_{i=1}^{l=3} \frac{TP_i}{TP_i + FN_i}\right)/3 \qquad (1)$$

$$Precision_M = \left(\Sigma_{i=1}^{l=3} \frac{TP_i}{TP_i + FP_i}\right)/3 \qquad (2)$$

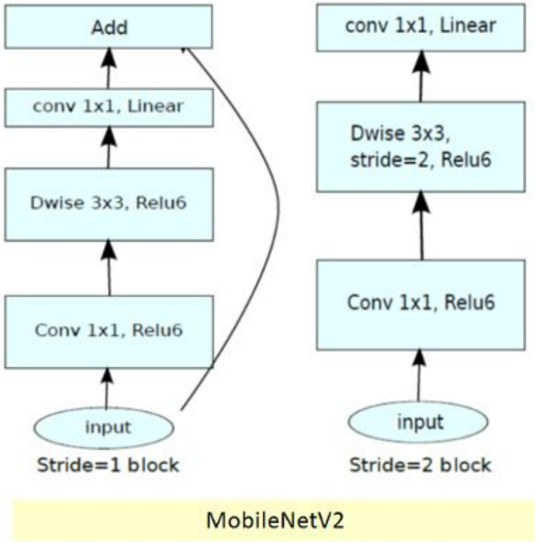
$$Fscore_M = \frac{(\beta^2 + 1) Precision_M Recall_M}{\beta^2 Precision_M + Recall_M} \qquad (3)$$

Transfer Learning Model Architecture

- Two data augmentation layers
- Pre-trained base model initialized with ImageNet weights (MobileNetV2)
- Global average pooling 2D layer
- Prediction layer with softmax activation

```
base_model = tf.keras.applications.MobileNetV2(input_shape=image_shape,
                                               include_top=False,
                                               weights='imagenet')
base_model.trainable = trainable

model = tf.keras.Sequential([layers.experimental.preprocessing.RandomFlip('horizontal'),
                             layers.experimental.preprocessing.RandomRotation(0.1),
                             base_model,
                             tf.keras.layers.GlobalAveragePooling2D(),
                             tf.keras.layers.Dense(3, activation='softmax')])
```



Model Experimentations

Model 1: NP3SMOTE

- Not pretrained
- SMOTE in the training set

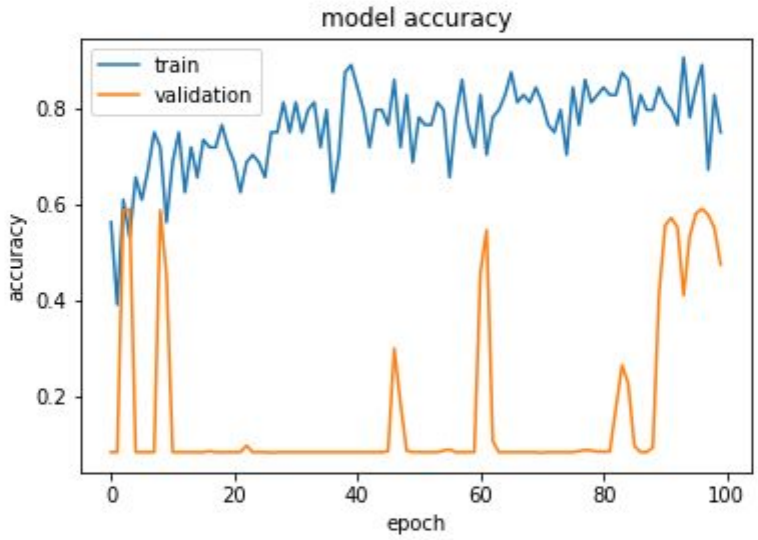
Model Summary:

- Model: not pretrained
- Epochs: 100
- Learning rate: 0.005
- Class weights: No
- SMOTE: Yes

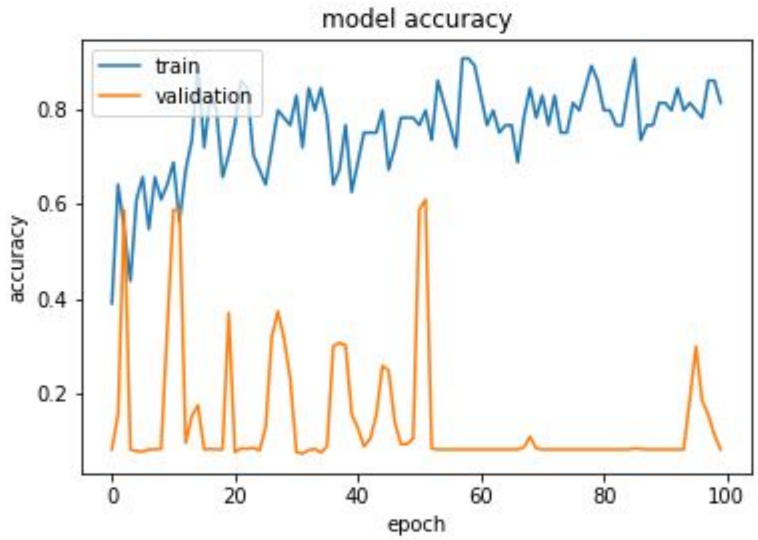
Fold	Accuracy	Loss
1	47.42	2.76
2	8.41	4.75
3	19.75	2.16
4	8.31	4.23
mean	20.97	3.48

Model 1: NP3SMOTE

Best fold



Worst fold



Model 1: NP3SMOTE

Predicted Classes

True Classes

Best fold

	precision	recall	f1-score	support	pred_0	pred_1	pred_2
class 0	0.58	0.72	0.64	524	377	147	0
class 1	0.19	0.62	0.29	74	28	46	0
class 2	0.00	0.00	0.00	294	244	50	0
accuracy	0.47	0.47	0.47	0.47	-	-	-
macro avg	**0.26**	**0.45**	**0.31**	**892**	**-**	**-**	**-**
weighted avg	**0.36**	**0.47**	**0.40**	**892**	**-**	**-**	**-**

Worst fold

	precision	recall	f1-score	support	pred_0	pred_1	pred_2
class 0	0.00	0.00	0.00	523	0	523	0
class 1	0.08	1.00	0.15	74	0	74	0
class 2	0.00	0.00	0.00	294	0	294	0
accuracy	0.08	0.08	0.08	0.08	-	-	-
macro avg	**0.03**	**0.33**	**0.05**	**891**	**-**	**-**	**-**
weighted avg	**0.01**	**0.08**	**0.01**	**891**	**-**	**-**	**-**

Model 2: P0SMOTE

- +pretrained
- +SMOTE
- +decreased lr

Model Summary:

- Model: pretrained MobileNet
- Epochs: 100
- Learning rate: 0.0001
- Class weights: No
- SMOTE: Yes

Previous Model:

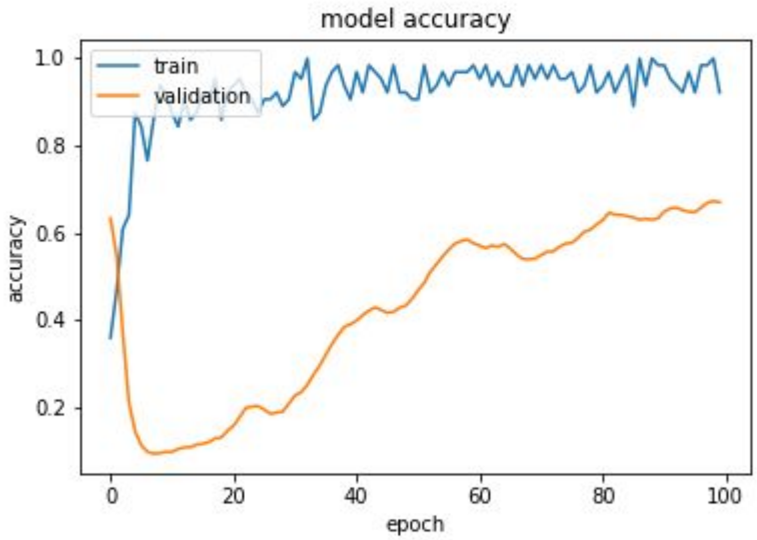
Fold	Accuracy	Loss
1	47.42	2.76
2	8.41	4.75
3	19.75	2.16
4	8.31	4.23
mean	20.97	3.48

Now:

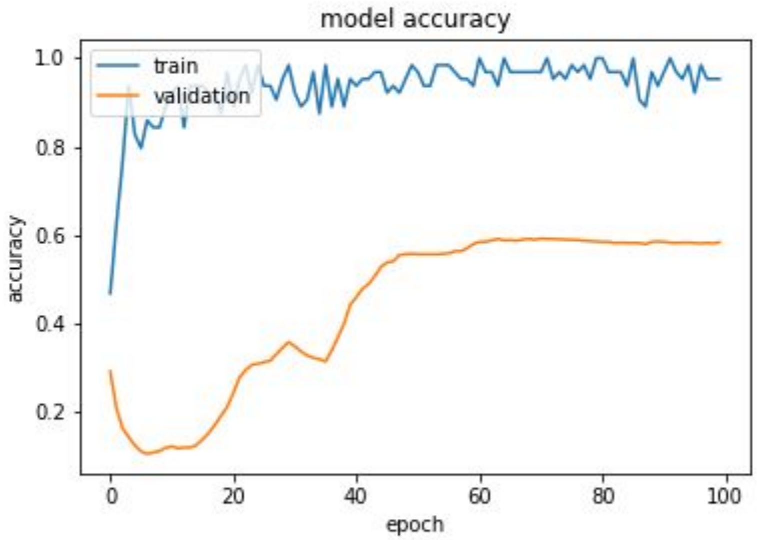
Fold	Accuracy	Loss
1	67.03	0.88
2	60.05	2.86
3	58.42	3.09
mean	61.83	2.28

Model 2: P0SMOTE

Best fold



Worst fold



Model 2: P0SMOTE

Best fold

	precision	recall	f1-score	support	pred_0	pred_1	pred_2
class 0	0.66	0.92	0.77	698	642	41	15
class 1	0.47	0.36	0.41	99	52	36	11
class 2	0.82	0.30	0.44	392	273	0	119
accuracy	0.67	0.67	0.67	0.67	-	-	-
macro avg	**0.65**	**0.53**	**0.54**	**1189**	**-**	**-**	**-**
weighted avg	**0.70**	**0.67**	**0.63**	**1189**	**-**	**-**	**-**

Worst fold

	precision	recall	f1-score	support	pred_0	pred_1	pred_2
class 0	0.59	0.99	0.74	697	690	7	0
class 1	0.36	0.04	0.07	98	94	4	0
class 2	0.00	0.00	0.00	393	393	0	0
accuracy	0.58	0.58	0.58	0.58	-	-	-
macro avg	**0.32**	**0.34**	**0.27**	**1188**	-	-	-
weighted avg	**0.37**	**0.58**	**0.44**	**1188**	-	-	-

Model 3: P1SMOTE

- +Class weights
  - {0: 4., 1: 26., 2: 7.}

Model Summary:

- Model: pretrained MobileNet
- Epochs: 100
- Learning rate: 0001
- Class weights: {0: 4., 1: 26., 2: 7.}
- SMOTE: Yes

Previous Model:

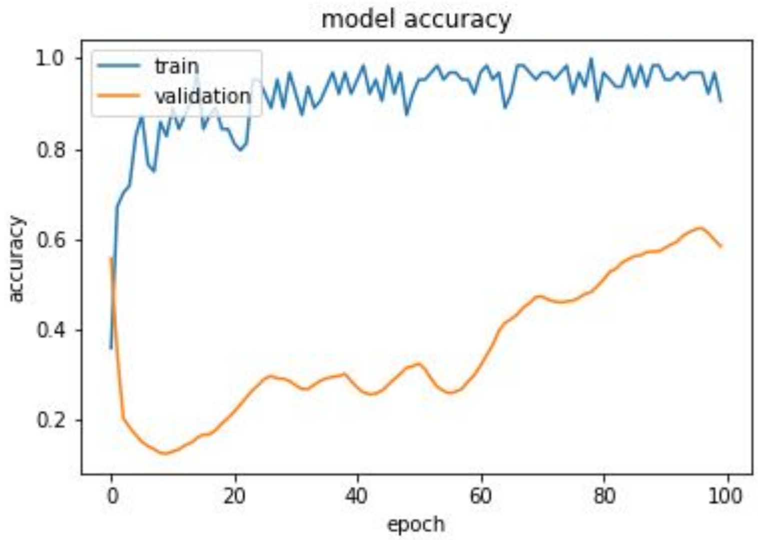
Fold	Accuracy	Loss
1	67.03	0.88
2	60.05	2.86
3	58.42	3.09
mean	61.83	2.28

Now:

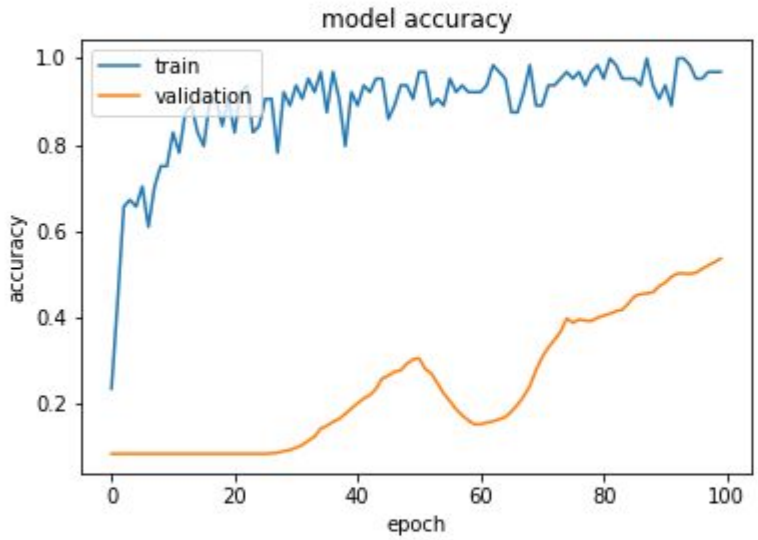
Fold	Accuracy	Loss
1	58.45	1.15
2	53.57	1.31
3	57.58	1.04
mean	56.53	1.17

Model 3: P1SMOTE

Best fold



Worst fold



Model 3: P1SMOTE

Best fold

	precision	recall	f1-score	support	pred_0	pred_1	pred_2
class 0	0.61	0.81	0.70	698	562	136	0
class 1	0.30	0.60	0.40	99	40	59	0
class 2	1.00	0.19	0.32	392	314	4	74
accuracy	0.58	0.58	0.58	0.58	-	-	-
macro avg	**0.64**	**0.53**	**0.47**	**1189**	**-**	**-**	**-**
weighted avg	**0.71**	**0.58**	**0.55**	**1189**	**-**	**-**	**-**

Worst fold

	precision	recall	f1-score	support	pred_0	pred_1	pred_2
class 0	0.60	0.78	0.68	698	547	151	0
class 1	0.30	0.80	0.43	99	20	79	0
class 2	1.00	0.03	0.05	392	345	36	11
accuracy	0.54	0.54	0.54	0.54	-	-	-
macro avg	**0.63**	**0.54**	**0.39**	**1189**	-	-	-
weighted avg	**0.71**	**0.54**	**0.45**	**1189**	-	-	-

Model 4: P2SMOTE

- +adjusted class weights
    - {0: 4., 1: 26., 2: 12.}

Model Summary:

- Model: pretrained MobileNet
- Epochs: 100
- Learning rate: 0.0001
- Class weights: {0: 4., 1: 26., 2: 12.}
- SMOTE: Yes

Previous Model:

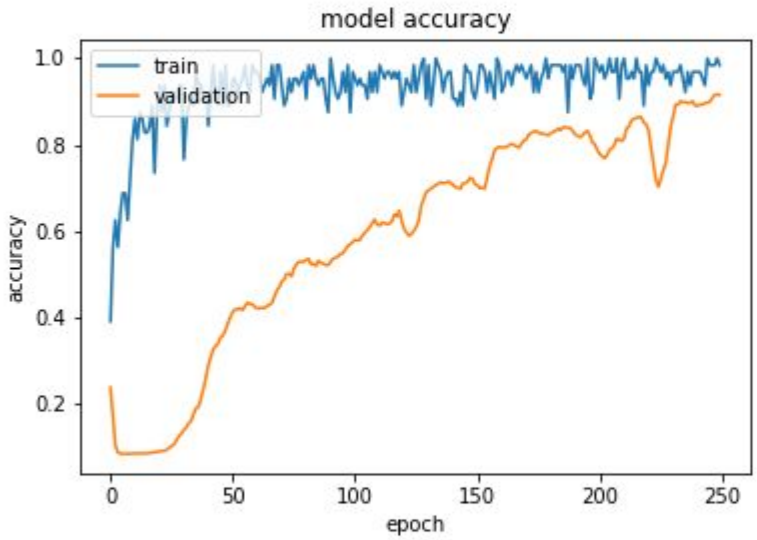
Fold	Accuracy	Loss
1	58.45	1.15
2	53.57	1.31
3	57.58	1.04
mean	56.53	1.17

Now:

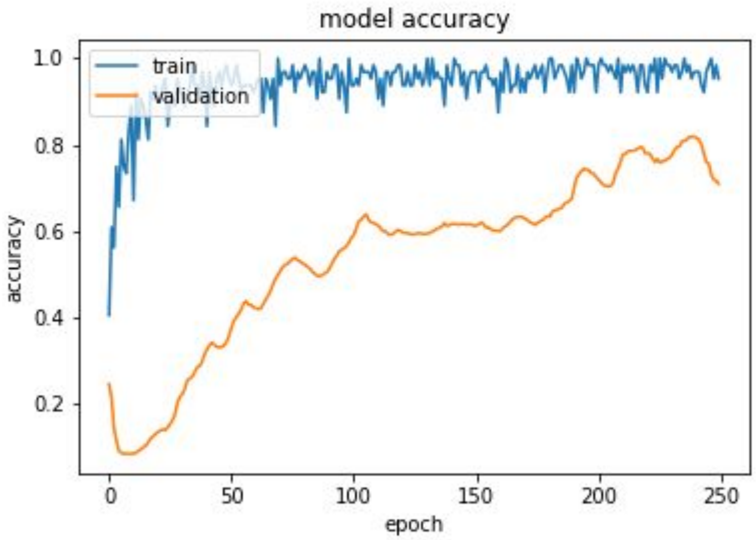
Fold	Accuracy	Loss
1	91.50	0.23
2	75.69	0.68
3	71.04	0.64
mean	79.41	0.52

Model 4: P2SMOTE

Best fold



Worst fold



Model 4: P2SMOTE

Best fold

	precision	recall	f1-score	support	pred_0	pred_1	pred_2
class 0	0.95	0.92	0.93	698	643	14	41
class 1	0.84	0.82	0.83	99	8	81	10
class 2	0.88	0.93	0.9	392	27	1	364
accuracy	0.92	0.92	0.92	0.92	-	-	-
macro avg	**0.89**	**0.89**	**0.89**	**1189**	-	-	-
weighted avg	**0.92**	**0.92**	**0.92**	**1189**	-	-	-

Worst fold

	precision	recall	f1-score	support	pred_0	pred_1	pred_2
class 0	0.69	0.96	0.8	697	669	6	22
class 1	0.89	0.52	0.66	98	31	51	16
class 2	0.77	0.32	0.45	393	269	0	124
accuracy	0.71	0.71	0.71	0.71	-	-	-
macro avg	**0.78**	**0.6**	**0.64**	**1188**	-	-	-
weighted avg	**0.73**	**0.71**	**0.67**	**1188**	-	-	-

Model 5: P3SMOTE

- +increased steps per epoch

Model Summary:

- Model: pretrained MobileNet
- Epochs: 100
- Learning rate: 0.0001
- Class weights: No
- SMOTE: Yes

Previous Model:

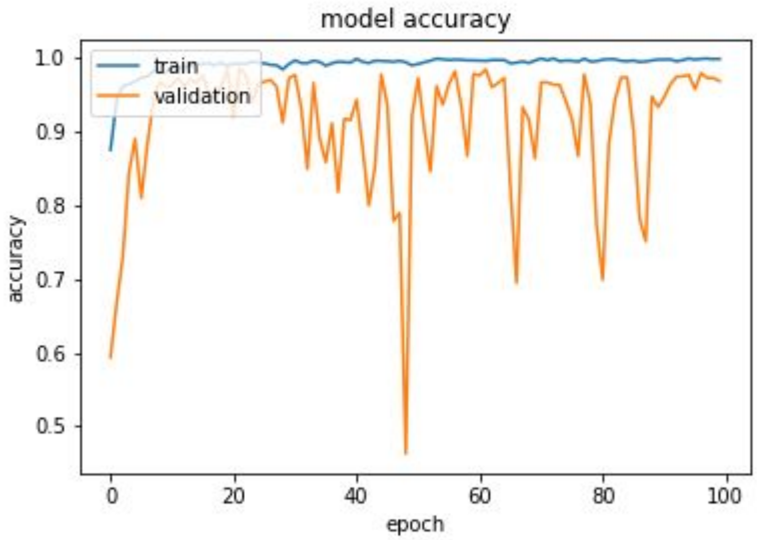
Fold	Accuracy	Loss
1	91.50	0.23
2	75.69	0.68
3	71.04	0.64
mean	79.41	0.52

Now:

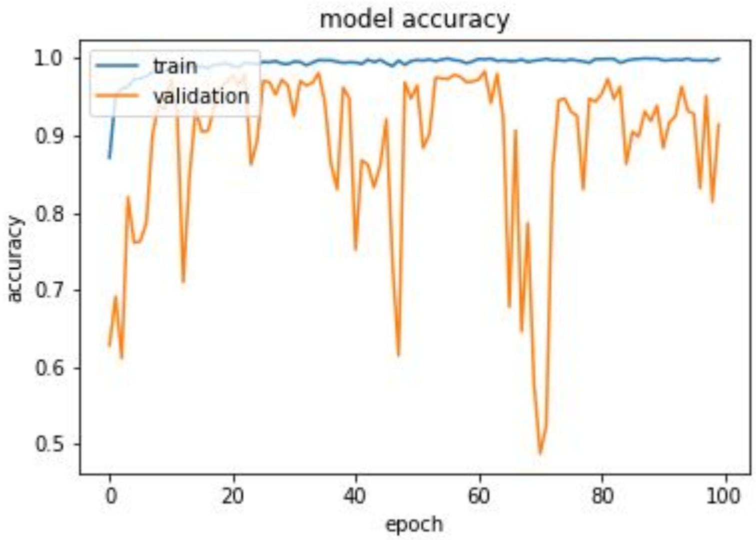
Fold	Accuracy	Loss
1	92.60	0.65
2	96.89	0.21
3	91.41	0.52
mean	93.63	0.46

Model 5: P3SMOTE

Best fold



Worst fold



Model 5: P3SMOTE

Best fold

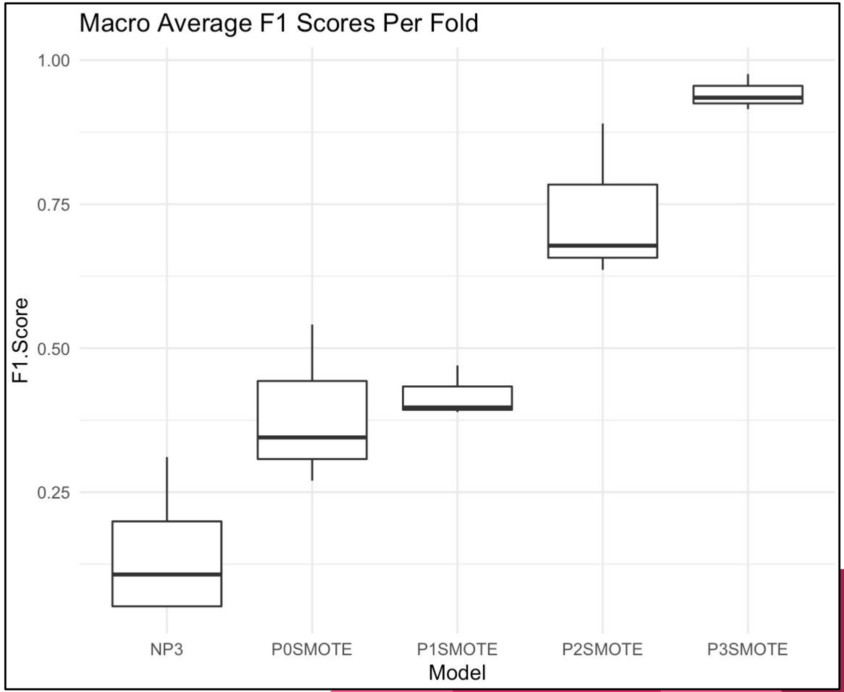
	precision	recall	f1-score	support	pred_0	pred_1	pred_2
class 0	0.99	0.95	0.97	698	665	0	33
class 1	1	1	1	99	0	99	0
class 2	0.92	0.99	0.95	392	4	0	388
accuracy	0.97	0.97	0.97	0.97	-	-	-
macro avg	**0.97**	**0.98**	**0.98**	**1189**	-	-	-
weighted avg	**0.97**	**0.97**	**0.97**	**1189**	-	-	-

Worst fold

	precision	recall	f1-score	support	pred_0	pred_1	pred_2
class 0	0.99	0.88	0.93	697	611	0	86
class 1	1.00	0.87	0.93	98	1	85	12
class 2	0.80	0.99	0.89	393	3	0	390
accuracy	0.91	0.91	0.91	0.91	-	-	-
macro avg	**0.93**	**0.91**	**0.92**	**1188**	-	-	-
weighted avg	**0.93**	**0.91**	**0.92**	**1188**	-	-	-

Model experimentation

Model	Change	Macro Average F1-Score for Fold				Average F1-Score
		1	2	3	4	
NP3		0.311	0.052	0.162	0.051	0.144
P0SMOTE	+pretrained +SMOTE +decreased lr	0.541	0.345	0.27	-	0.385
P1SMOTE	+class weights	0.47	0.389	0.397	-	0.419
P2SMOTE	+adjusted class weights	0.89	0.678	0.636	-	0.735
P3SMOTE	+increased steps per epoch	0.935	0.976	0.915	-	0.942
P6SMOTE	-decreased steps per epoch -decreased epochs	**0.924**	**0.893**	**0.974**	**-**	**0.930**



Challenges

- All images getting classified into the same category
- Balancing class weight
- Real world interpretation of the results

Next Steps

- Tune parameters
    - Class weights
    - Learning rates
    - Epochs
- Evaluate model performance on test set
- Clean up our code

Q & A

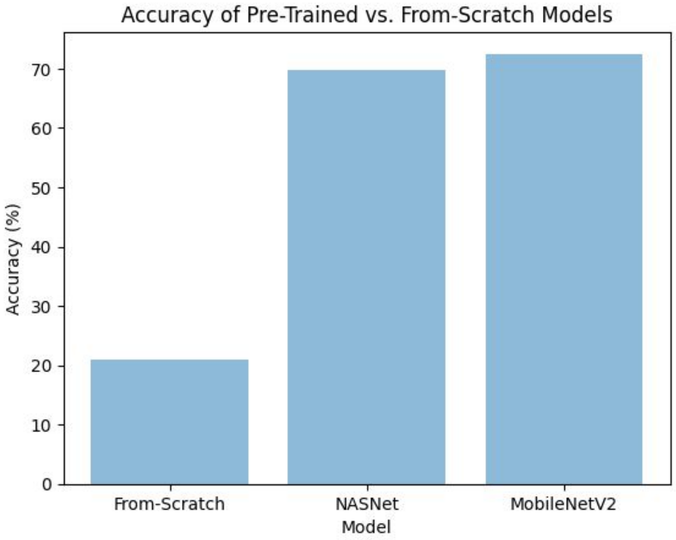
Thanks for listening! Any questions?

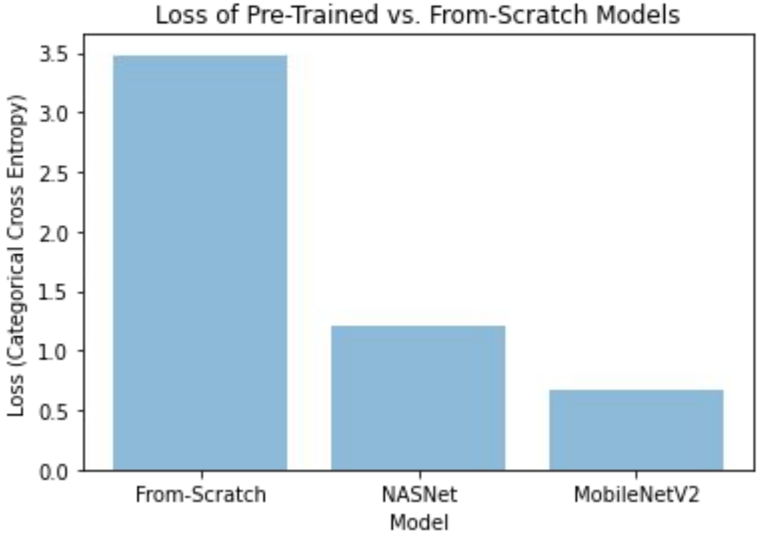
References

1. Molinaro AM, Simon R, and Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. bioinformatics. 2005;21(15):3301-3307. doi:10.1093/bioinformatics/bti499.
2. Wannian (PRC) Aylward, Bruce (WHO); Liang. 2020. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). The WHOChina Joint Mission on Coronavirus Disease 2019,2019(February):16–24.
3. Gayani Chandrarathne, Kokul Thanikasalam, and Amalka Pinidiyaarachchi. 2020. A Comprehensive Study on Deep Image Classification with Small Datasets.
4. Francois Chollet et al. Metrics.
5. Joseph Paul Cohen, Paul Morrison, and Lan Dao. 2020. Covid-19 image data collection. arXiv 2003.11597.
6. Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalena Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. 2018. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell, 172(5):1122– 1131.e9.
7. Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection.In Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI' 95, page 1137–1143, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
8. V Mayr, Dobrescu Ai, A Chapman, E Persad, I Klerings, G Wagner, U Siebert, C Christof, C Zachariah, G Gartlehner, V Mayr, Dobrescu Ai, A Chapman, E Persad, I Klerings, G Wagner, U Siebert, C Christof, and C Zachariah. 2020. Measures to control COVID-19 : a rapid review.
9. Paul Mooney. 2018. Chest x-ray images (pneumonia).
10. Public Health Ontario. 2020. Coronavirus disease 2019 (covid-19) testing.
11. Danish Rafiq, Asiya Batool, and M A Bazaz. 2020. Three months of COVID-19: A systematic review and meta-analysis. Reviews in medical virology, (April):e2113.
12. Mahendra Sahare and Hitesh Gupta. 2012. A review of multi-class classification for imbalanced data. International Journal of Advanced Computer Research, 2:160–164.
13. Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 6.
14. Tanu Singhal. 2020. A Review of Coronavirus Disease-2019 (COVID-19). Indian Journal of Pediatrics, 87:281–286.
15. Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4):427 – 437.
16. Pengfei Sun, Xiaosheng Lu, Chao Xu, Wenjuan Sun, and Bo Pan. 2020. Understanding of COVID-19 based on current evidence. Journal of Medical Virology, 92(6):548–551.
17. Samir S. Yadav and Shivajirao M. Jadhav. 2019. Deep convolutional neural network based medical

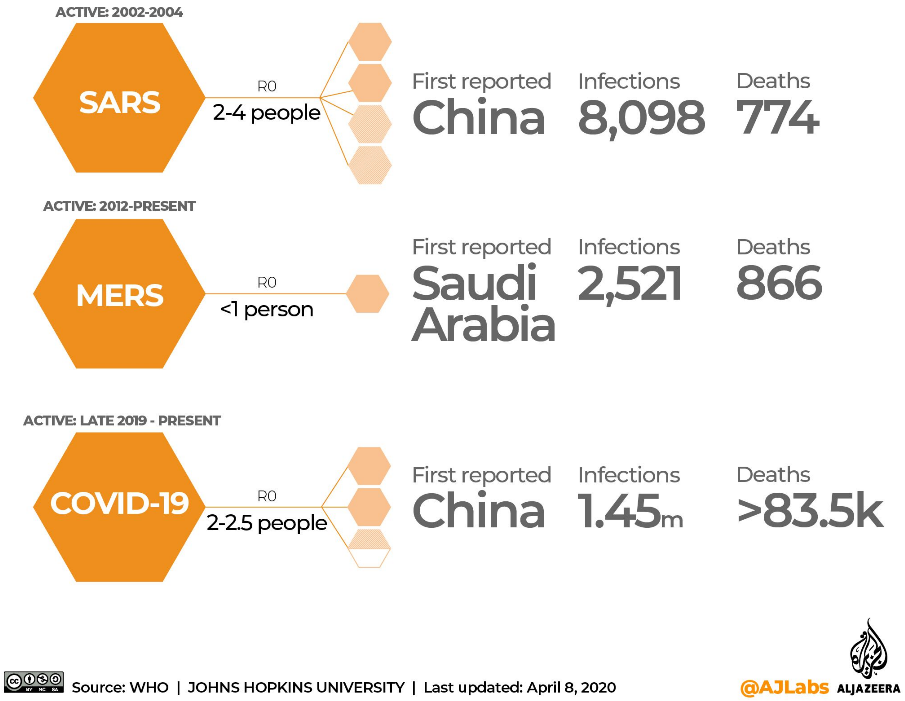
Appendix

Pre-trained models comparison





Motivation


ACTIVE: 2002-2004
SARS
R0
2-4 people
First reported
China
Infections
8,098
Deaths
774
ACTIVE: 2012-PRESENT
MERS
R0
<1 person
First reported
Saudi Arabia
Infections
2,521
Deaths
866
ACTIVE: LATE 2019 - PRESENT
COVID-19
R0
2-2.5 people
First reported
China
Infections
1.45m
Deaths
>83.5k
Source: WHO | JOHNS HOPKINS UNIVERSITY | Last updated: April 8, 2020
@AJLabs
ALJAZEERA