# X-RAY IMAGE CLASSIFICATION:
## A Machine Learning Approach

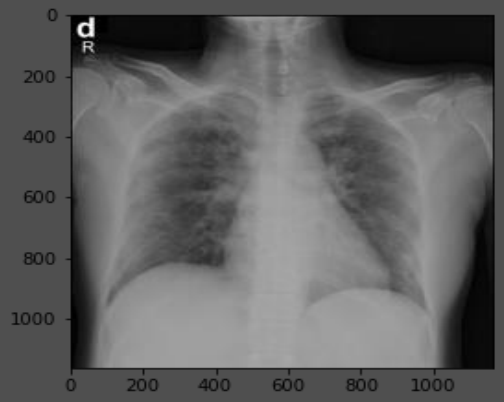
Gina McFarland

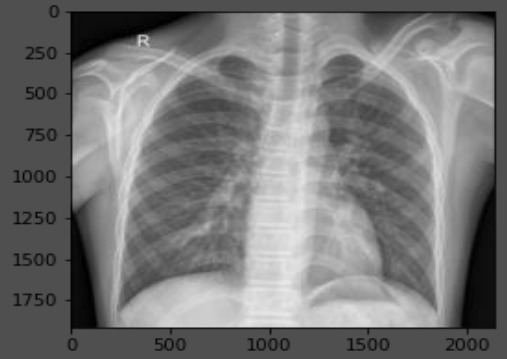
COMP 4449

# Research Question

Which machine learning model classifies lung x-ray images most accurately?

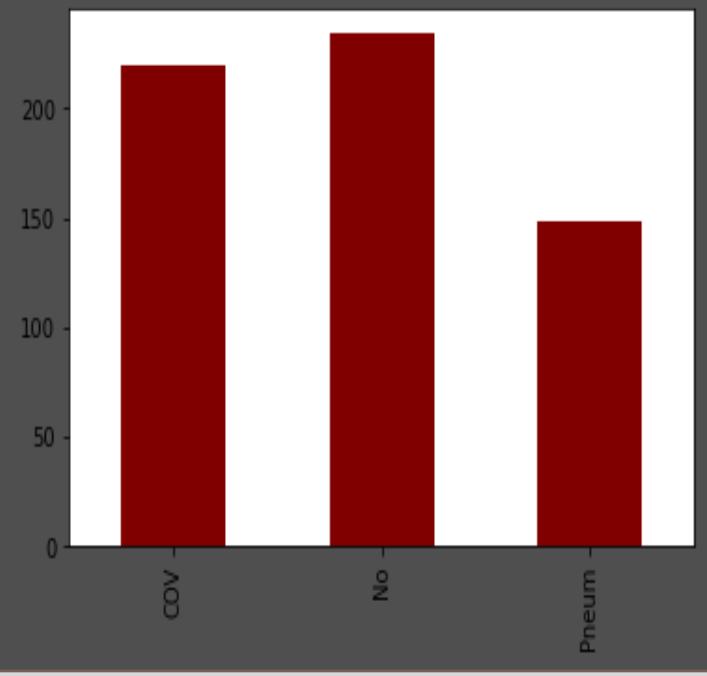
- Models examined: K Nearest Neighbor, Random Forest, Support Vector Machine, and Multilayer Perceptron Neural Networks




# Data

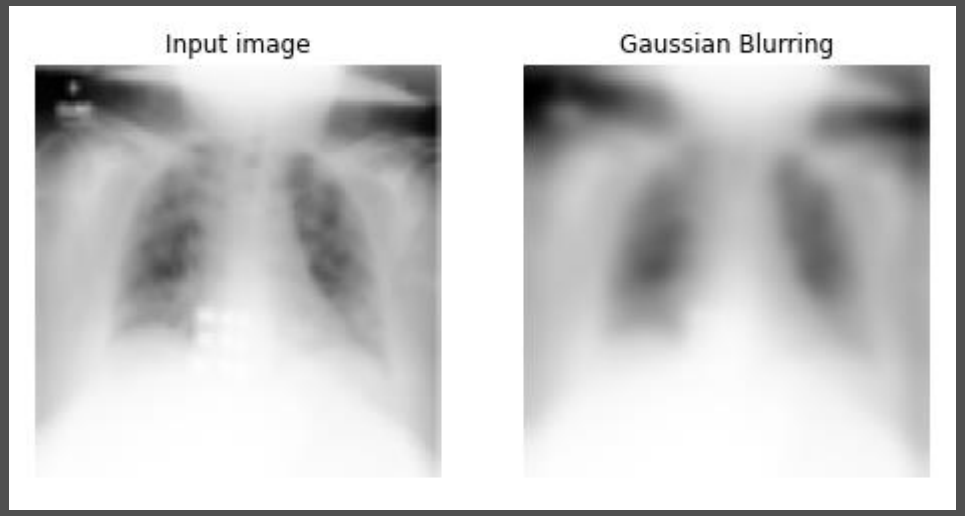
- 602 lung x-rays images
- Labelled Normal, Pneumonia, and COVID-19
- https://data.mendeley.com/datasets/fvk7h5dg2p/1



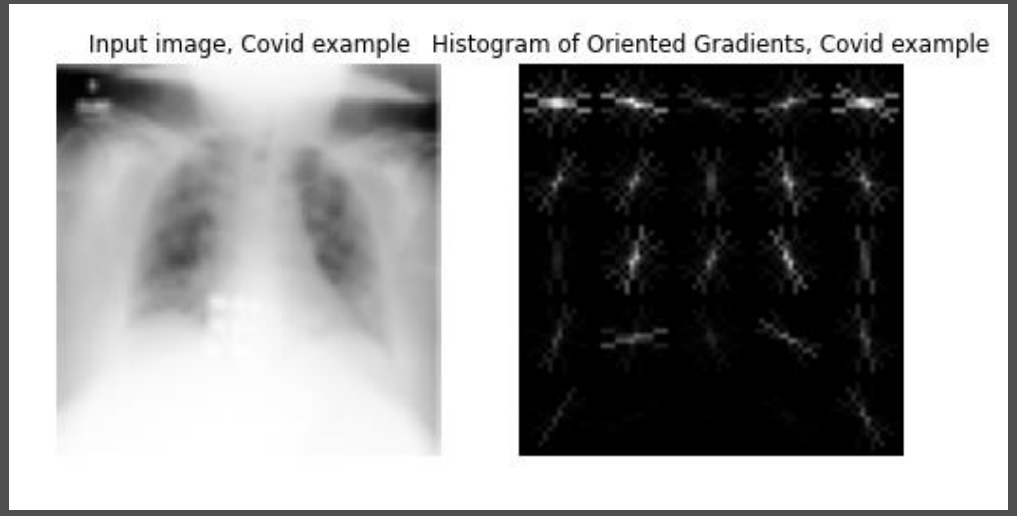
# Data Preparation

1. Resize Images
2. Gaussian Blurring
3. HOG Transform

## Gaussian Blurring



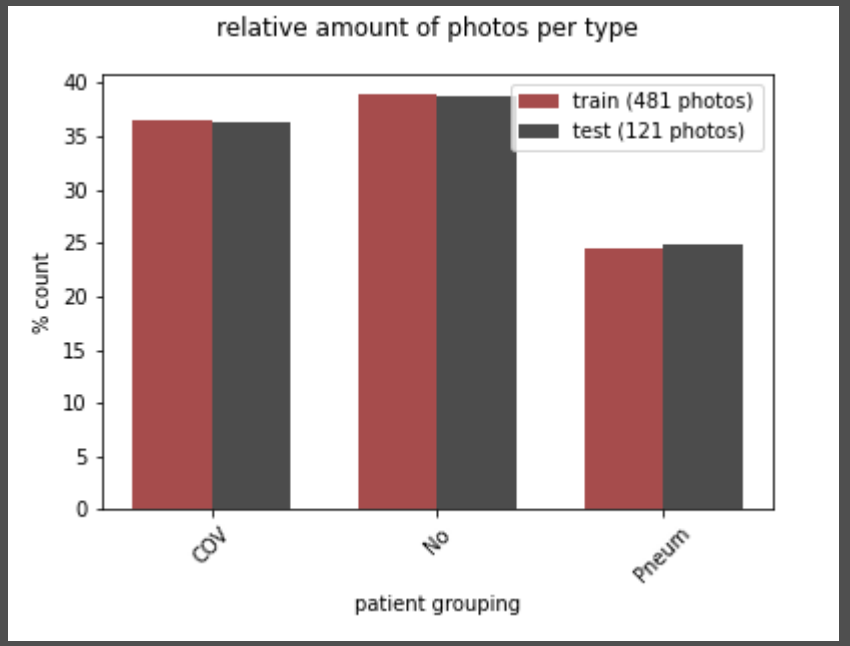

## Histogram of Oriented Gradients (HOG)



# Data Preparation

sklearn train_test_split package

- 80% train / 20 % test
- shuffle = True
- stratify = ‘y’



# Metrics

- Baseline Average Radiologist
  - Sample of radiologists (7) in China and US diagnosed x-ray images of COVID-19 and viral pneumonia
  - Average accuracy 74%
- Accuracy main metric
  - True positive and true negative most important
  - Classes reasonably balanced; No extreme skewing
- Precision, Recall, F1-score
  - These were also considered

# K Nearest Neighbor Model

- Concept is similar points are grouped together
- Calculates the distance of incoming points to all existing data points
- Classifies incoming point by taking a vote based on the label of the nearest K points
- The best number of neighbors to consider, (K), will differ based on the data

```
Classification Report:
 Labels: {'Normal', 'COVID19', 'Pneumonia'}
 Classifier: KNeighborsClassifier(n_neighbors=1, p=1):
              precision    recall  f1-score   support

         COV       1.00      0.77      0.87        44
          No       0.72      0.89      0.80        47
       Pneum       0.79      0.77      0.78        30

    accuracy                           0.82       121
   macro avg       0.84      0.81      0.82       121
weighted avg       0.84      0.82      0.82       121
```

# Multilayer Perceptron Neural Network

- Multilayer Perceptron (MLP) is a neural network supervised machine learning algorithm
- Composed of one or more neural layers
  - Input layer, hidden layers, predictions are made on the output layer
- Activation functions propagate data through network
  - Data trained backwards through network
  - Input data only propagates forward

```
Classification Report:
 Labels: {'Normal', 'COVID19', 'Pneumonia'}
 Classifier: MLPClassifier(max_iter=1000, random_state=0,
              precision    recall  f1-score   support

         COV       0.95      0.95      0.95        44
          No       0.84      0.89      0.87        47
       Pneum       0.89      0.80      0.84        30

    accuracy                           0.89       121
   macro avg       0.89      0.88      0.89       121
weighted avg       0.89      0.89      0.89       121
```

# Random Forest

- Random Forest Classifier is ensemble of decision trees
- Decision trees sensitive to the data they are trained on
- Random Forest avoids this with randomization
- Classification made from highest vote counts from created trees

Classifier: RandomForestClassifier(max_depth=8, n_estimators=57,

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| COV | 0.91 | 0.98 | 0.95 | 44 |
| No | 0.87 | 0.96 | 0.91 | 47 |
| Pneum | 0.95 | 0.70 | 0.81 | 30 |
| accuracy | | | 0.90 | 121 |
| macro avg | 0.91 | 0.88 | 0.89 | 121 |
| weighted avg | 0.91 | 0.90 | 0.90 | 121 |

# Support Vector Machine Classifier

- SVC finds best boundary to separates data points into different classes
- Creates boundary that is maximally distant from data points in each group to generalize for new data without breaking the model

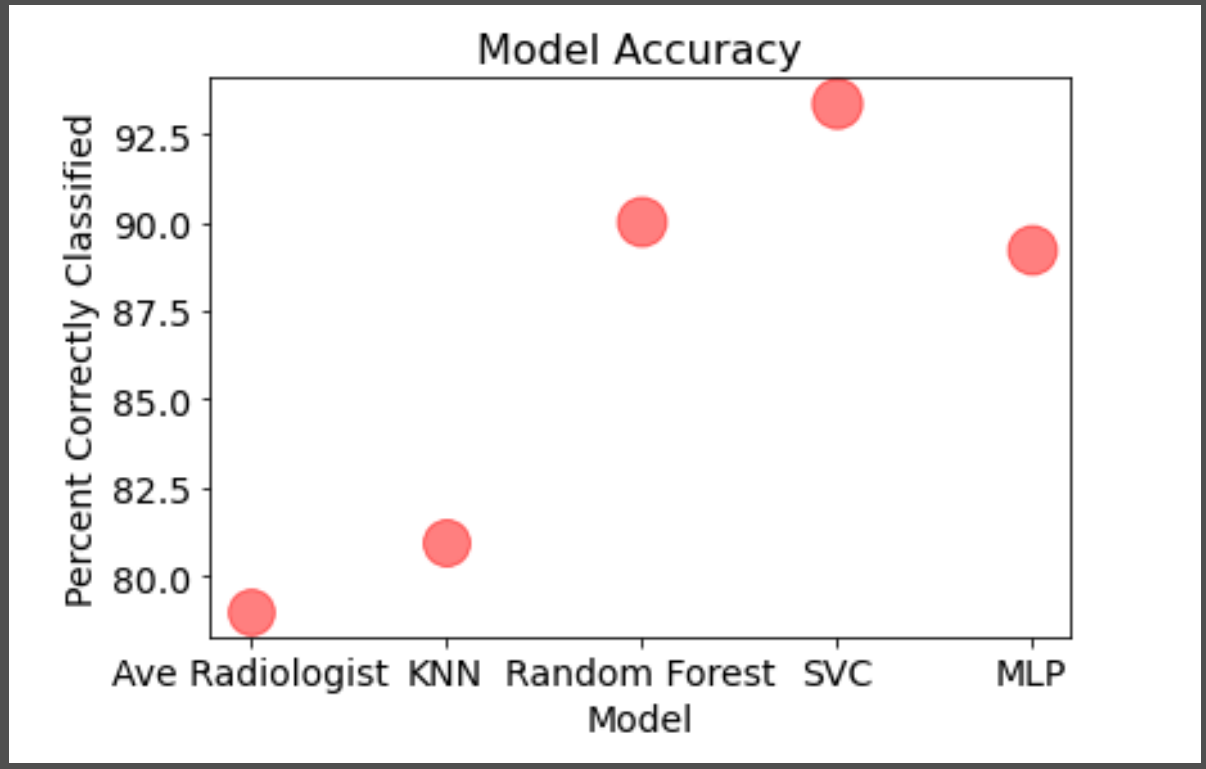
Labels: {'Normal', 'COVID19', 'Pneumonia'}
Classifier: SVC(C=1):

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| COV | 0.98 | 0.98 | 0.98 | 44 |
| No | 0.90 | 0.94 | 0.92 | 47 |
| Pneum | 0.93 | 0.87 | 0.90 | 30 |
| accuracy | | | 0.93 | 121 |
| macro avg | 0.93 | 0.93 | 0.93 | 121 |
| weighted avg | 0.93 | 0.93 | 0.93 | 121 |

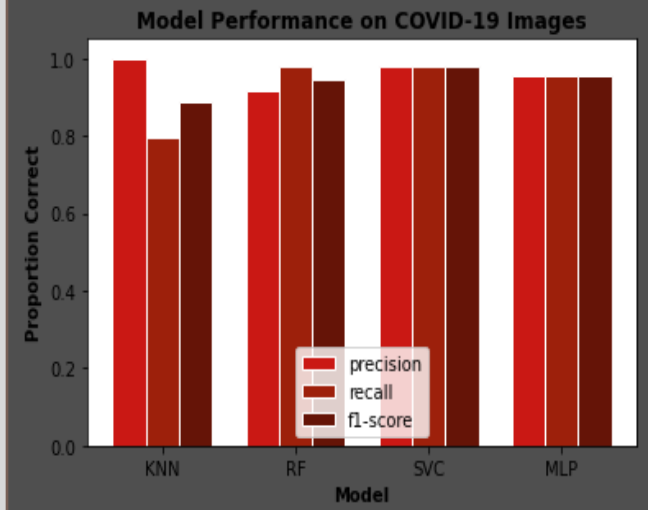
# Results

- SVC most accurate – 93%
- Random Forest – 90%
- MLP – 89%
- KNN – 82%
- Ave Radiologist - 74%

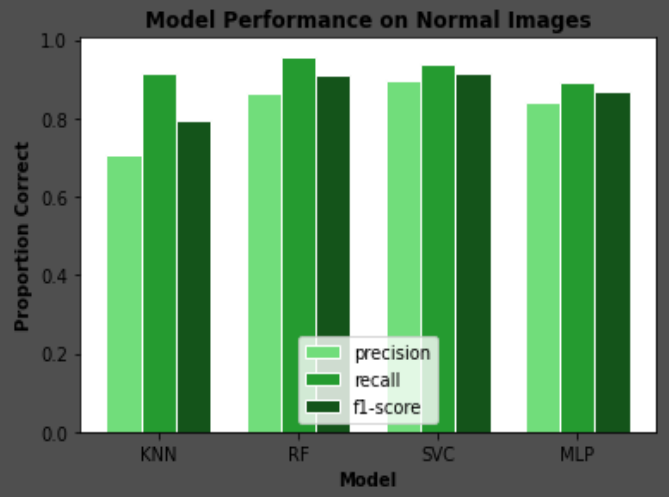
- Data prep biased toward SVC?



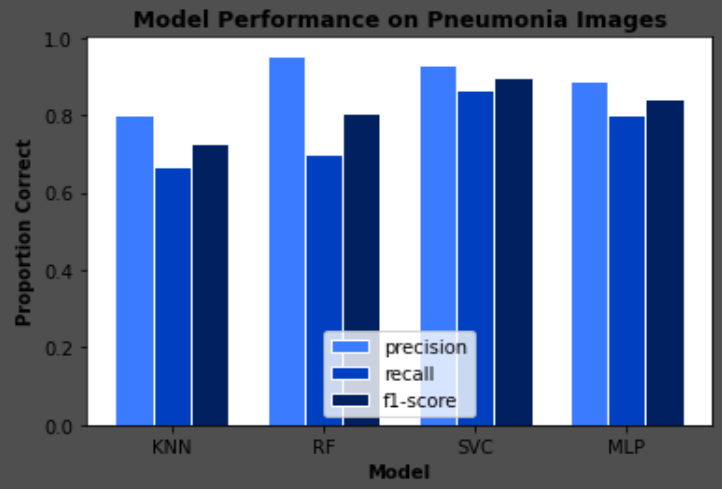
# Results




COVID-19




Normal




Pneumonia

Precision (leftmost bar), Recall (middle bar), F1-Score (rightmost bar)

- Models performed best on COVID-19 images
- RF and MLP performed well on COVID-19 and Normal images
  - Pneumonia set reduced performance
- SVC had consistent performance across all data sets

# Further Study

- Convolutional Neural Network (CNN)
  - Study used CNN for pneumonia x-ray detection
  - Achieved 98.4% accuracy(!)
  - CNNs were specifically designed to map image data
    - standard model for image predictions
- Differ image prep techniques
  - Without HOG transform, how would other models perform?
- Perform same analysis with balanced classes
  - MLP, SVC, and KNN sensitive to imbalanced classes
  - Random Forest more tolerant of data imbalances

# Questions?