X-ray Image Classification using Machine Learning Algorithm

Jannatul Ferdous Chhoa

Department of Mathematics
University of Houston

11 May, 2021

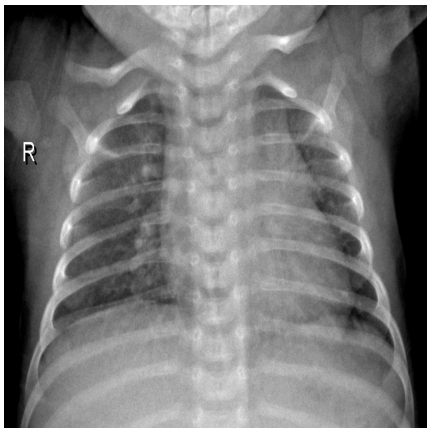
Introduction

- **Coronavirus Casualties:** Over 100 million infected, 3.3 million deaths
- **Main Tool:** Reverse transcription polymerase chain reaction (RT-PCR)
- **Goal:** Utility of classic machine learning algorithm in the rapid and accurate detection of COVID-19 from chest X-ray images

Introduction

Database: Consists of **3886** COVID-19 chest X-ray images collected from Kaggle.com

- 1200 images of patients with COVID-19
- 1345 images of patients with viral pneumonia
- 1341 images of patients with COVID-19 negative



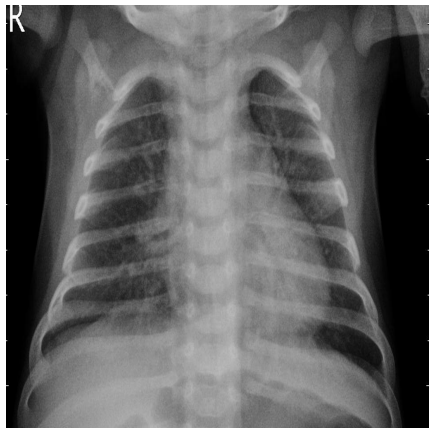


Figure: Examples of (i)Normal (ii)COVID (iii)Viral Pneumonia X-ray Images

Introduction

- **Data Pre-processing:** Resizing and extracting features from the data
- **Data Splitting:** K-fold Cross Validation
- **Image Classification Techniques:** 3 Supervised Algorithms:
  - Support Vector Machine (**SVM**)
  - Random Forest Classifier
  - Logistic Regression
- **Model Selection:** Grid Search CV
- **Results:** Best model with best estimators

Features in Wavelet

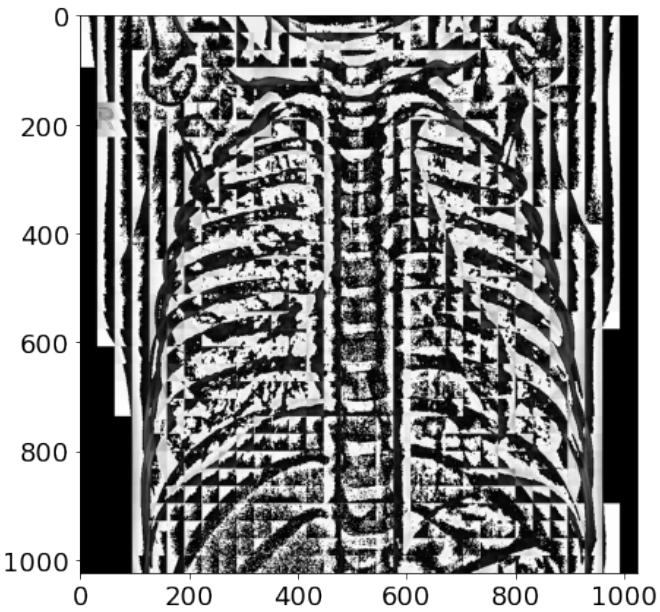


Figure: Normal Wavelet Image

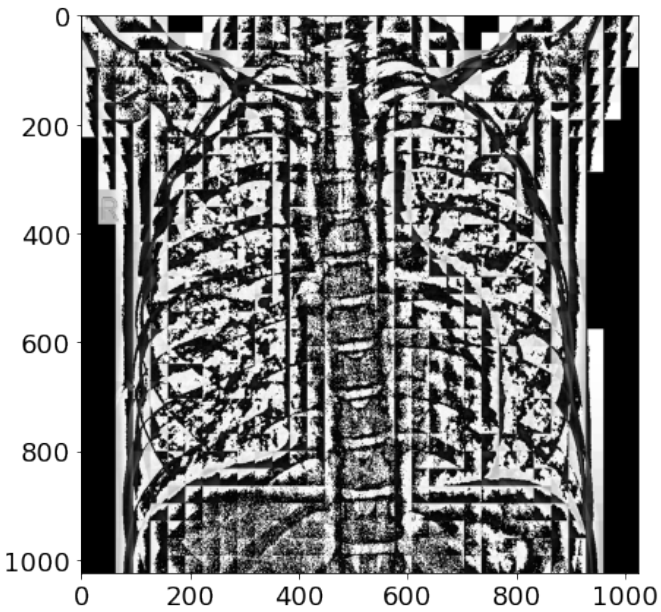


Figure: Covid Wavelet Image

Robust Principal Component Analysis

Decomposition of data matrix:

$$\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}$$

$\boldsymbol{X}$ is the data matrix

$\boldsymbol{L}$ is a structured low-rank matrix

$\boldsymbol{S}$ is a sparse matrix

In our case, $\boldsymbol{X}$ has rank 2500 and $\boldsymbol{L}$ has rank 1358

Robust Principal Component Analysis

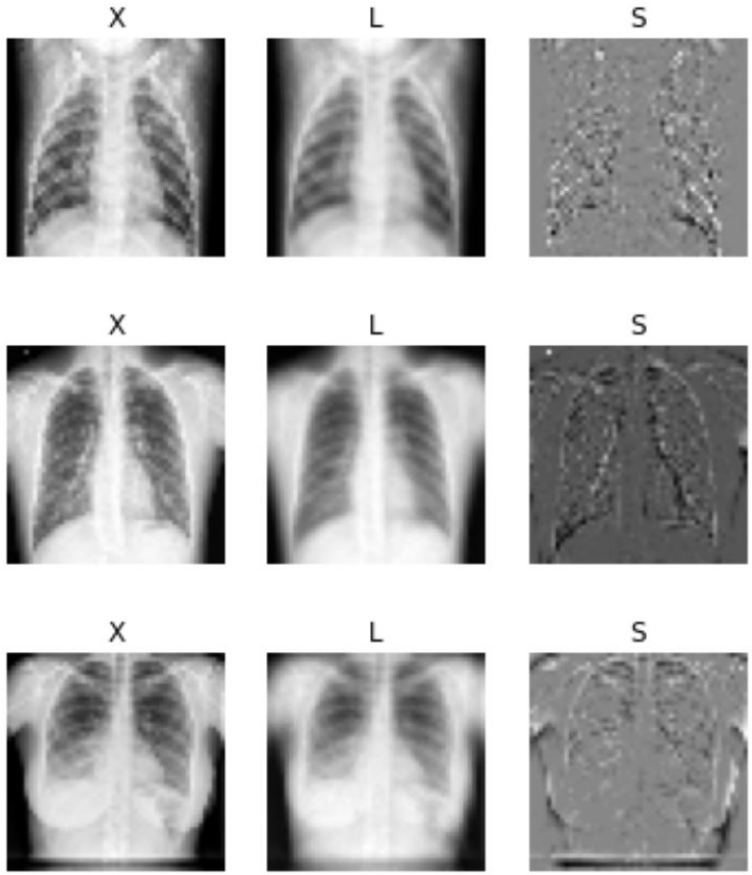
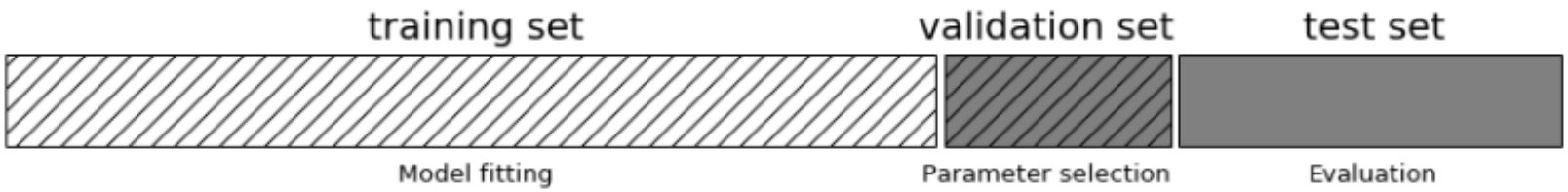


Figure: Extraction of the Low rank part and the Sparse part from the images.

Cross Validation



Types	Number of Images used:			
	Total	Training Set	Validation	Test Set
COVID-19	1200	720	180	300
Viral Pneumonia	1200	720	180	300
Normal	1200	720	180	300

Table: Number of images per class and per fold (5-fold Cross Validation will be used)

Model Parameters

- **Support Vector Machine:**
  - Regularization Parameter (**C**): 1, 10, 100, 1000
  - Kernels: rbf, linear
- **Random Forest:**
  - Number of trees in the Forest: 1, 5, 10
- **Logistic Regression:**
  - Regularization Parameter (**C**): 1, 5, 10

Grid Search CV

Figure: Grid Search CV

Outline Revisit

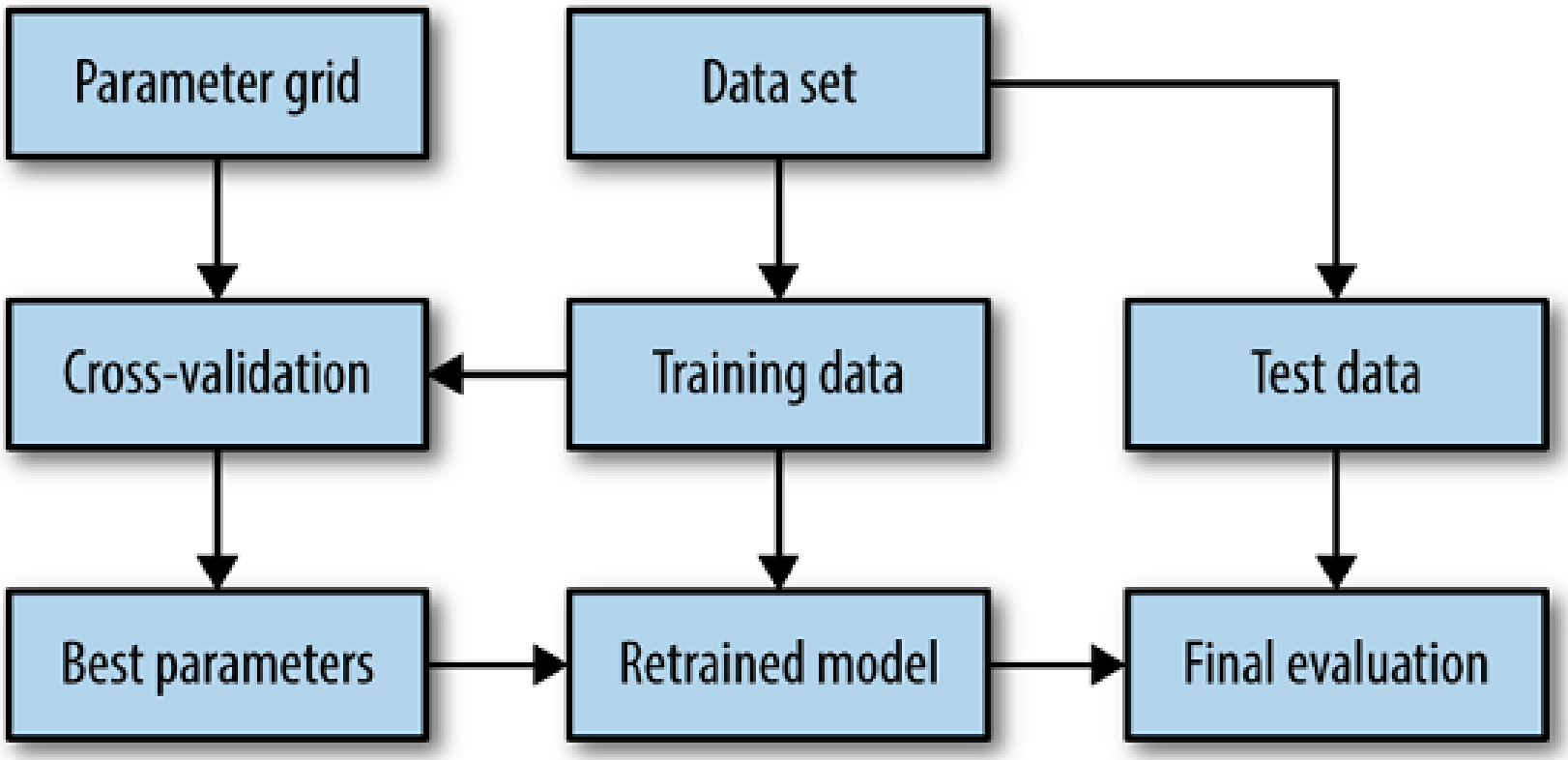


Figure: Outline

Comparison

	model	best_score	best_params
0	svm	0.889630	{'svc__C': 1, 'svc__kernel': 'rbf'}
1	random_forest	0.863704	{'randomforestclassifier__n_estimators': 10}
2	logistic_regression	0.856667	{'logisticregression__C': 1}

	model	best_score	best_params
0	svm	0.925185	{'svc__C': 1, 'svc__kernel': 'rbf'}
1	random_forest	0.878148	{'randomforestclassifier__n_estimators': 10}
2	logistic_regression	0.892963	{'logisticregression__C': 10}

	model	best_score	best_params
0	svm	0.940741	{'svc__C': 1, 'svc__kernel': 'rbf'}
1	random_forest	0.894074	{'randomforestclassifier__n_estimators': 10}
2	logistic_regression	0.908889	{'logisticregression__C': 1}

Figure: Best fine-tuned parameters using Grid Search CV using first (i)10 PCA (ii)15 PCA (iii)20 PCA (On Training Data)

Comparison

Table: Best Scores (On Test Data)

Model	Best Score
SVM	**0.944**
Random Forest	0.874
Logistic Regression	0.897

Prediction Result

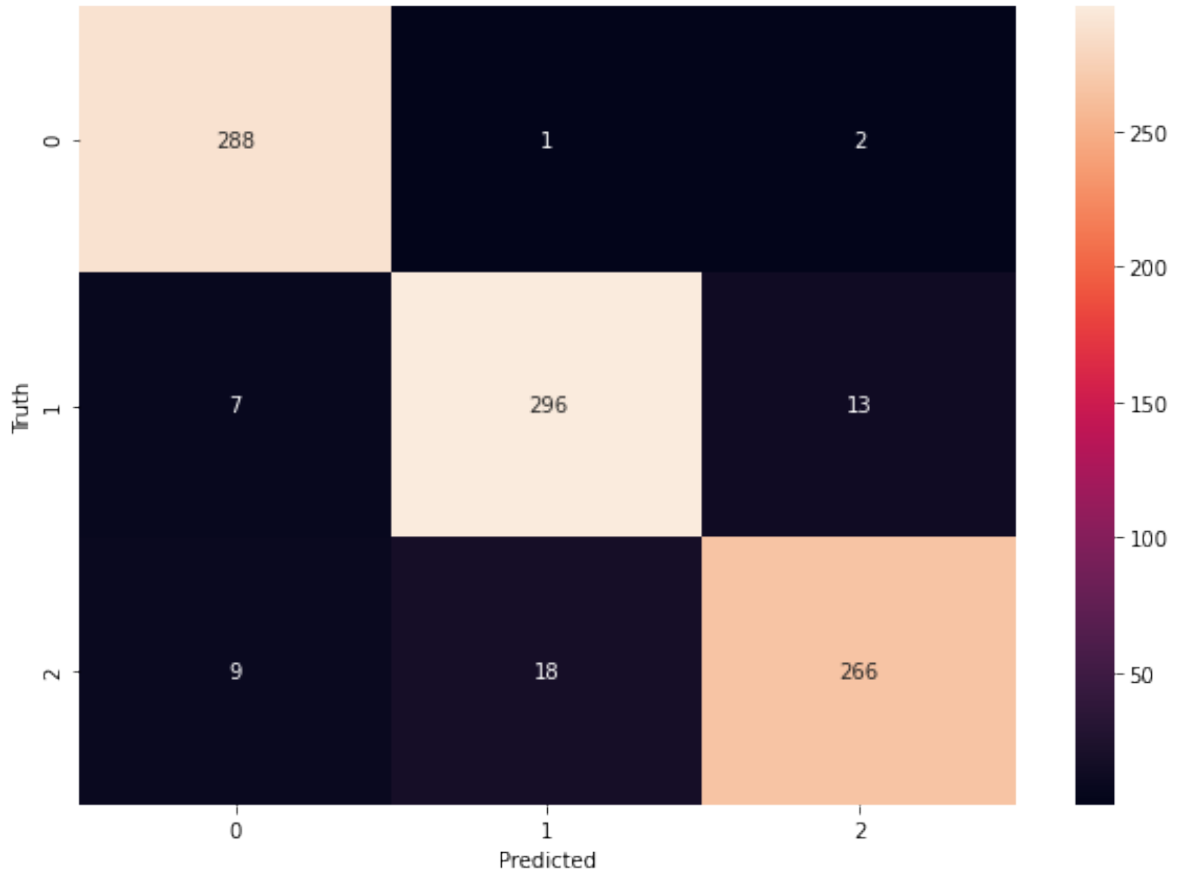


Figure: Confusion Matrix produced by the best model (SVM)

Conclusion

M. E. H. Chowdhury et al., "Can AI Help in Screening Viral and COVID-19 Pneumonia?," in IEEE Access,vol.8,pp.132665-132676, 2020,doi: 10.1109/ACCESS.2020.3010287.

Accuracy using CNN: 99%

Accuracy using RPCA and SVM: 94%

Thank You!
Q/A