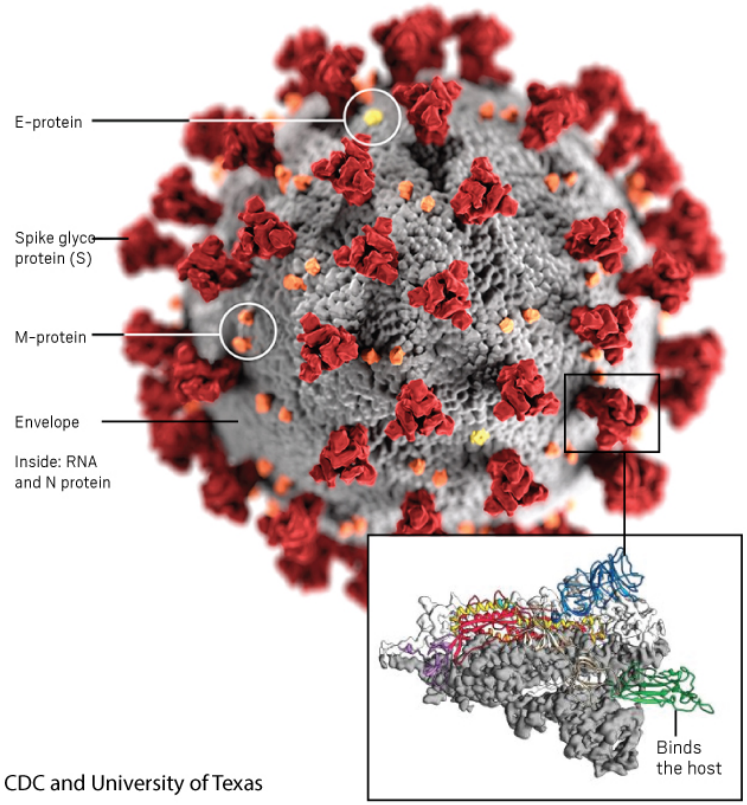
Rapid COVID-19 Diagnosis using Raman Spectroscopy and Machine Learning

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement



Author: Isaac Ghebregziabher

Capstone Project One

July 4, 2021

Springboard

World under COVID-19 pandemic crisis

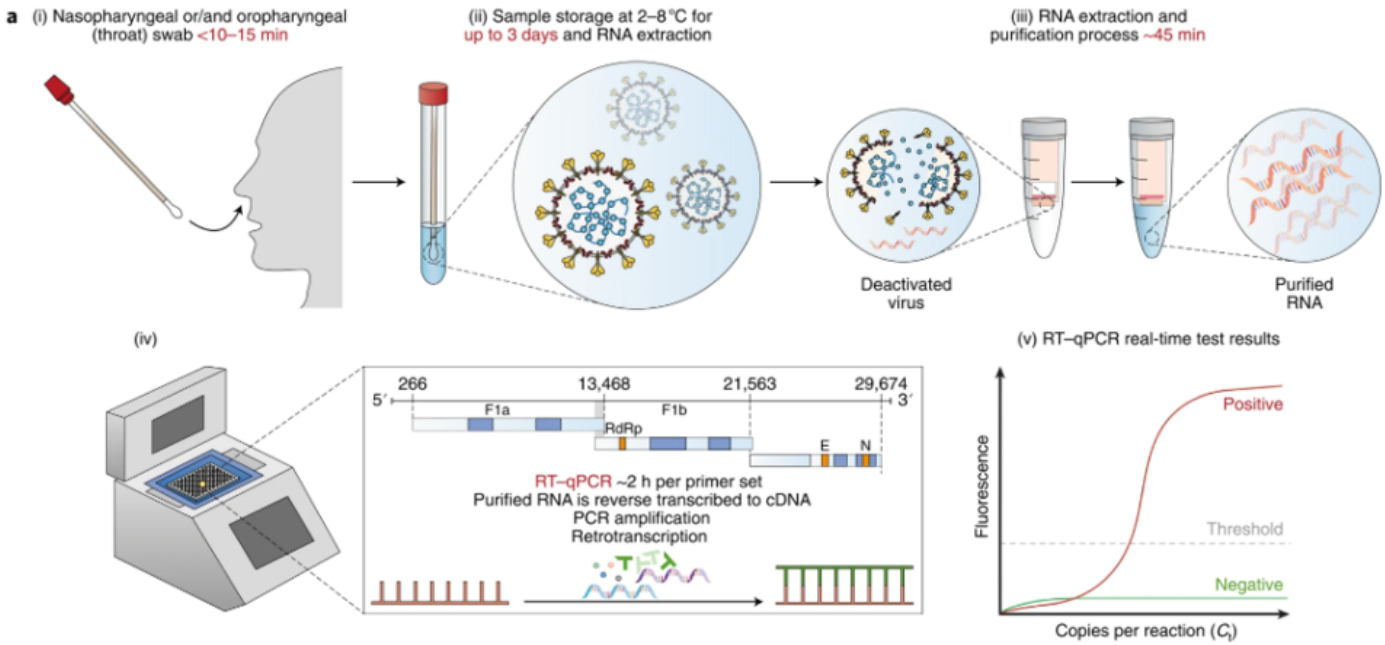
Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

- > 171 million active cases
- 3.5+ million deaths
- Fast and reliable diagnostic is needed

RT-PCR – Current COVID-19 detection method is time consuming and expensive

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

- 3 days for sample preparation and RNA extraction
- Expensive PCR



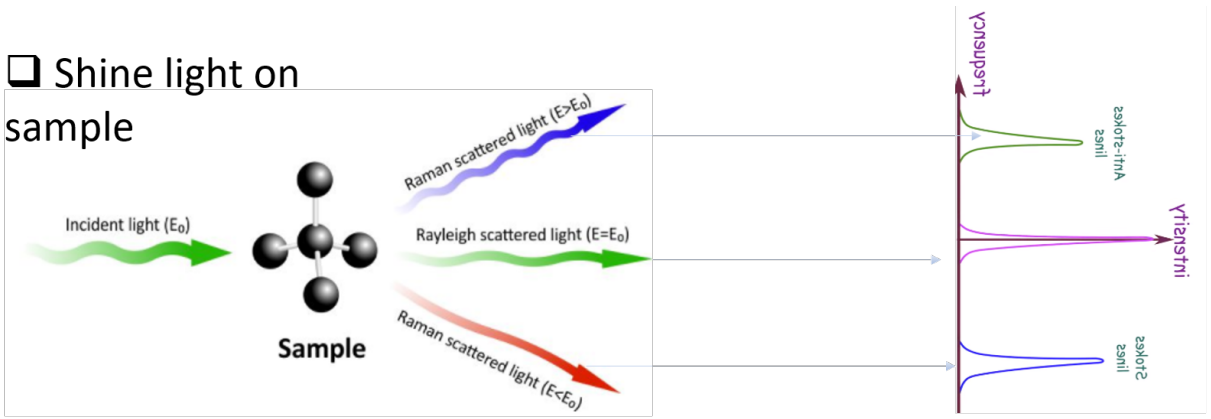
Springboard

Principle of Raman effect

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

- Most light scatters unaffected (Rayleigh scattering)
- A few percent gets Raman scattered
- Raman Scattered light is signature of molecular composition

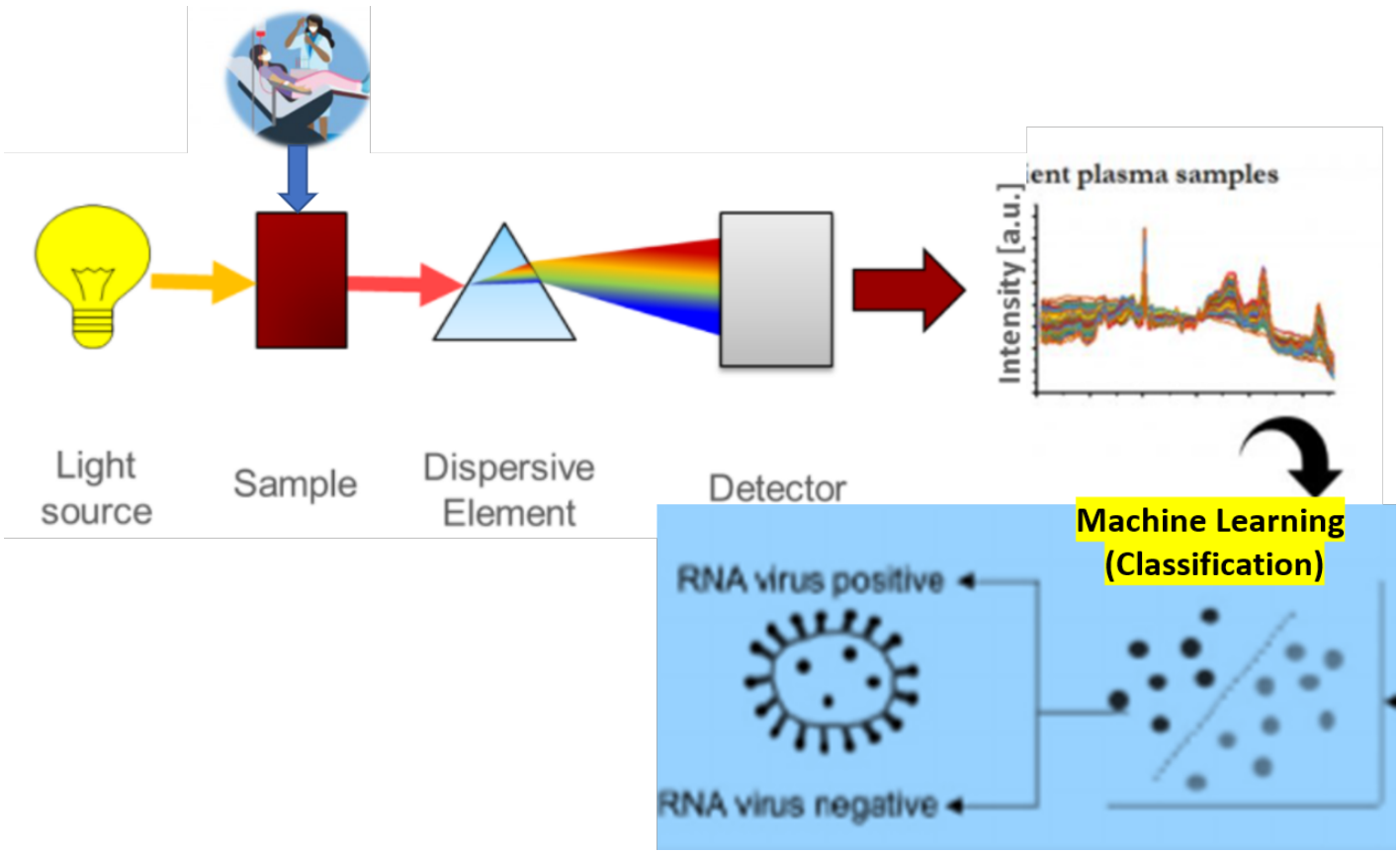
❑ Shine light on sample



❑ Most of the light is unaffected
❑ Small percentage of light undergoes frequency shift

Rapid detection of Covid-19 using Raman spectroscopy and Machine Learning

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement



Springboard

Overview of Dataset obtained from Kaggle

- 309 rows X 901 columns
- Each column → Raman Wavenumber
- Rows → Intensity for wavenumbers
- Each row corresponds to one observation
- Last column '***diagnostic***' is target variable

Feature names = Wavenumbers in units of cm-1

	400	402	405	407	410	412	415	417	420	422	...	diagnostic
0	0.0	-0.015237	-0.030607	-0.038309	-0.039078	-0.035809	-0.031176	-0.030395	-0.033311	-0.031603	...	Healthy
1	0.0	-0.012098	-0.028164	-0.035189	-0.036138	-0.031050	-0.026015	-0.027539	-0.028084	-0.027075	...	Healthy
2	0.0	-0.013000	-0.029058	-0.035021	-0.034994	-0.033025	-0.028413	-0.028470	-0.029737	-0.029198	...	Healthy
3	0.0	-0.015728	-0.034346	-0.045140	-0.047671	-0.044334	-0.040807	-0.040474	-0.041417	-0.040699	...	Healthy
4	0.0	-0.020355	-0.045839	-0.060556	-0.065805	-0.064988	-0.062097	-0.061955	-0.064759	-0.066886	...	Healthy

Raman scattered intensities for each Wavenumber

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

Raman detector might have dead pixels corresponding to all zero value intensities

- No missing values in dataset
- 9 features wave-numbers with 0 intensity value
- Drop null features (treated as missing)

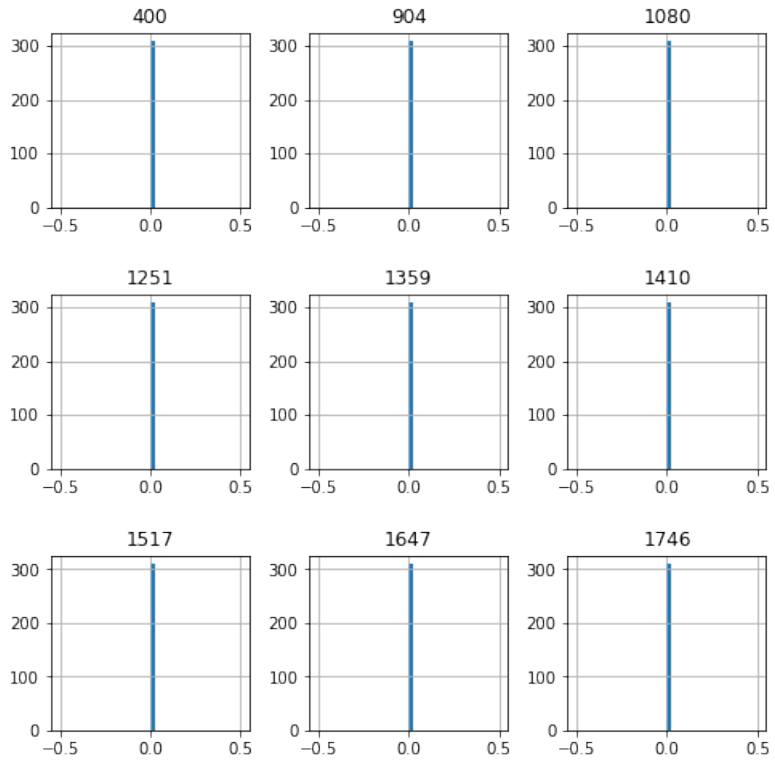


Figure: Single valued features (dropped).

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

Dataset is balanced

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

- No class imbalance issues
- Dataset is balanced with ≈ 50:50 class ratio.

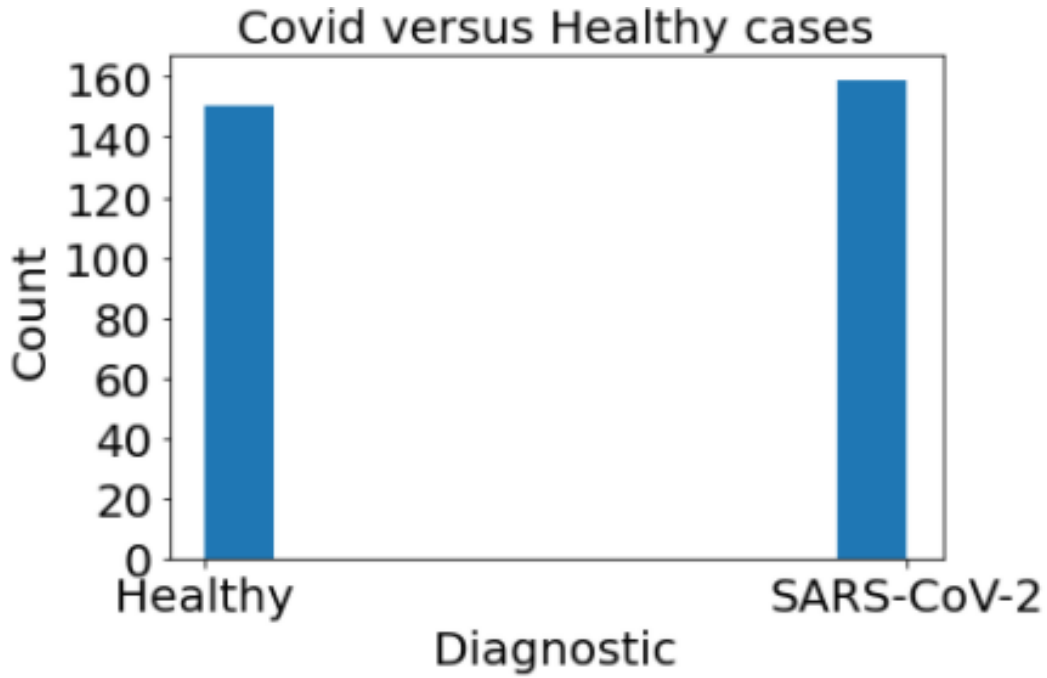
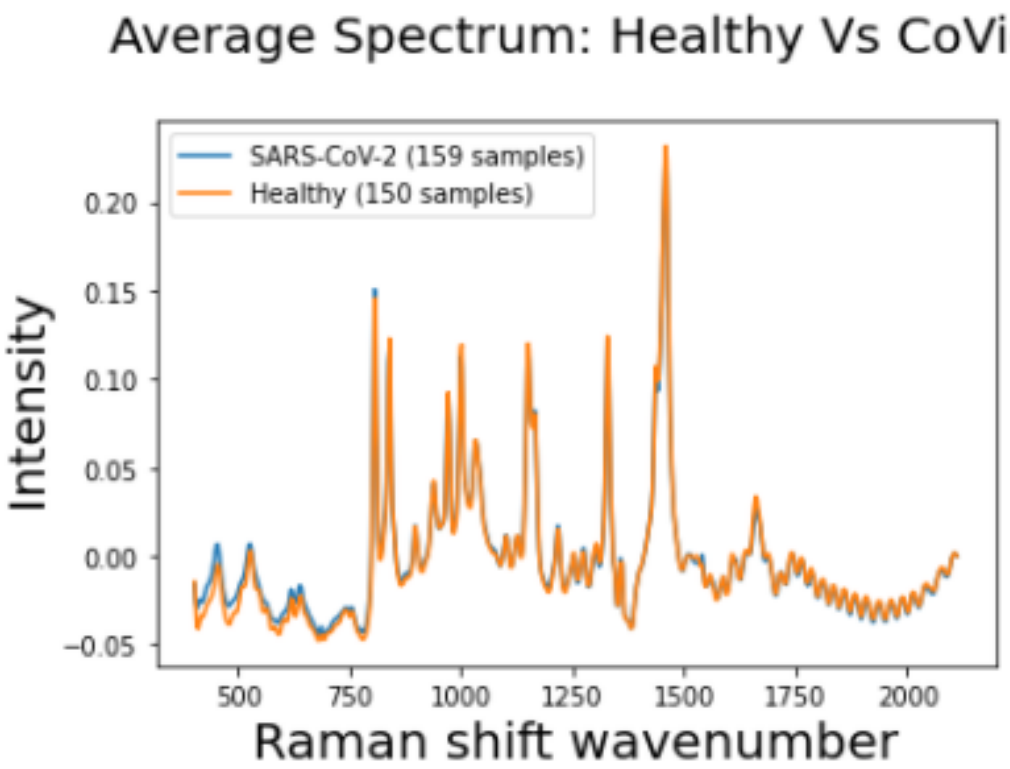


Figure: ≈ 50 : 50 COVID to Healthy class ratio.

Visually indiscernible Raman spectrum

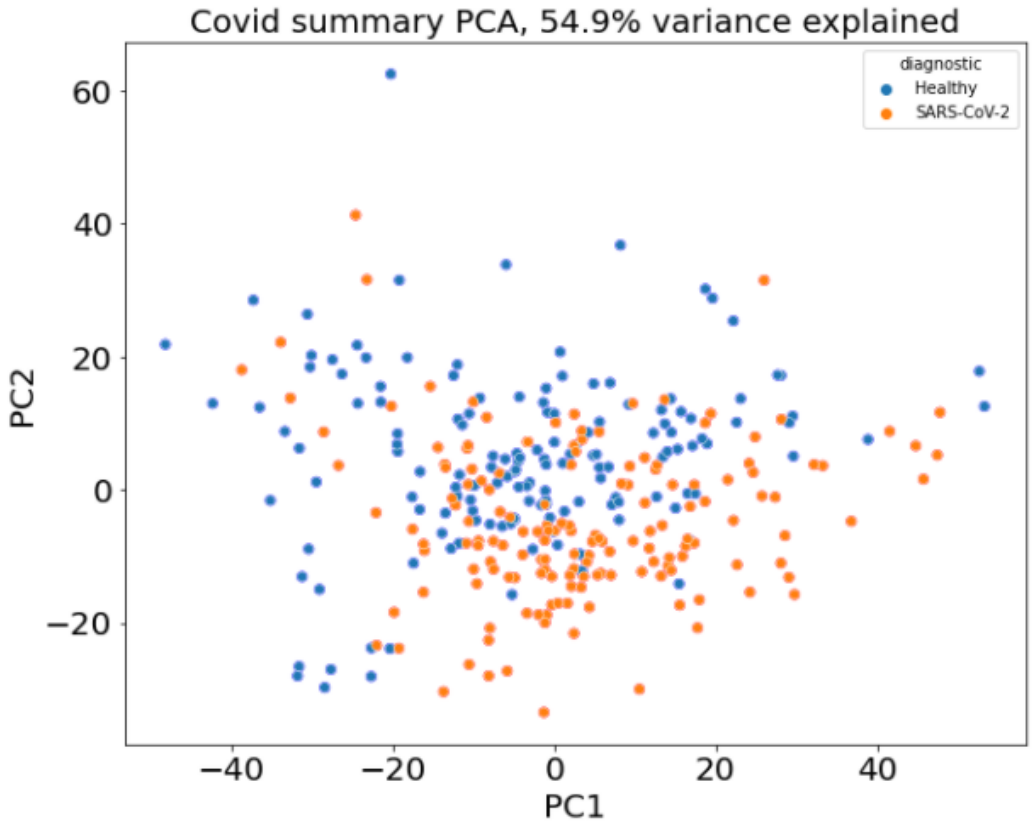
- Visually difficult to easily identify COVID from Healthy.
- Machine Learning model is needed for fast and reliable COVID diagnosis.



Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

High dimensional data visualization - PCA

- Over 50 percent variance explained with two principal components.
- No visual class separation

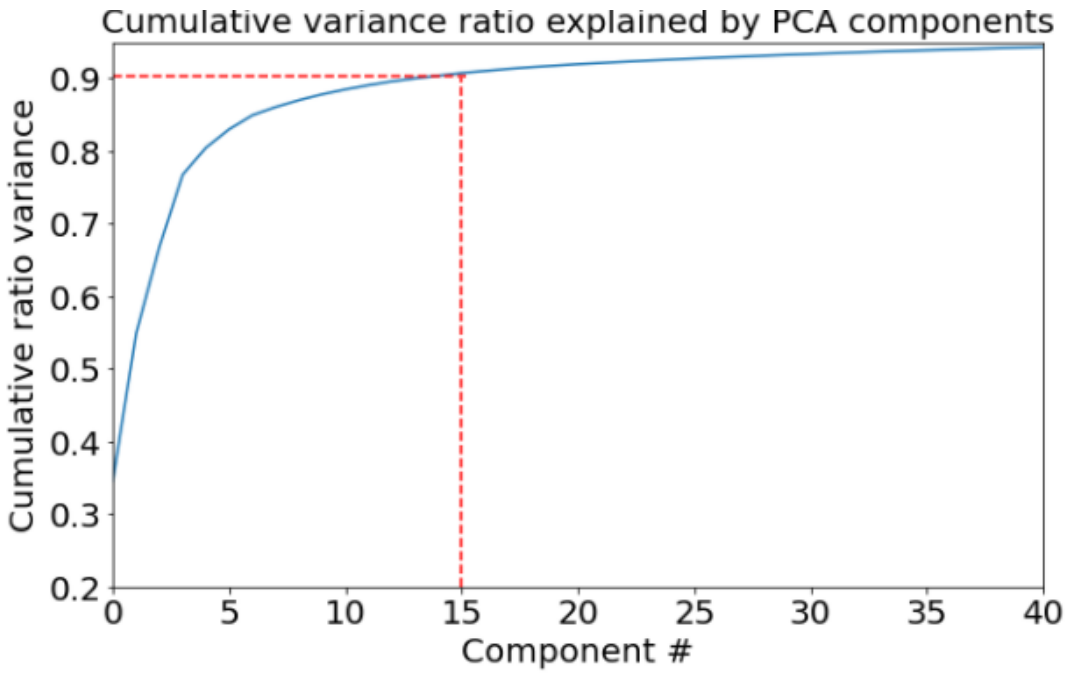


Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

Principal component analysis – feature reduction

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

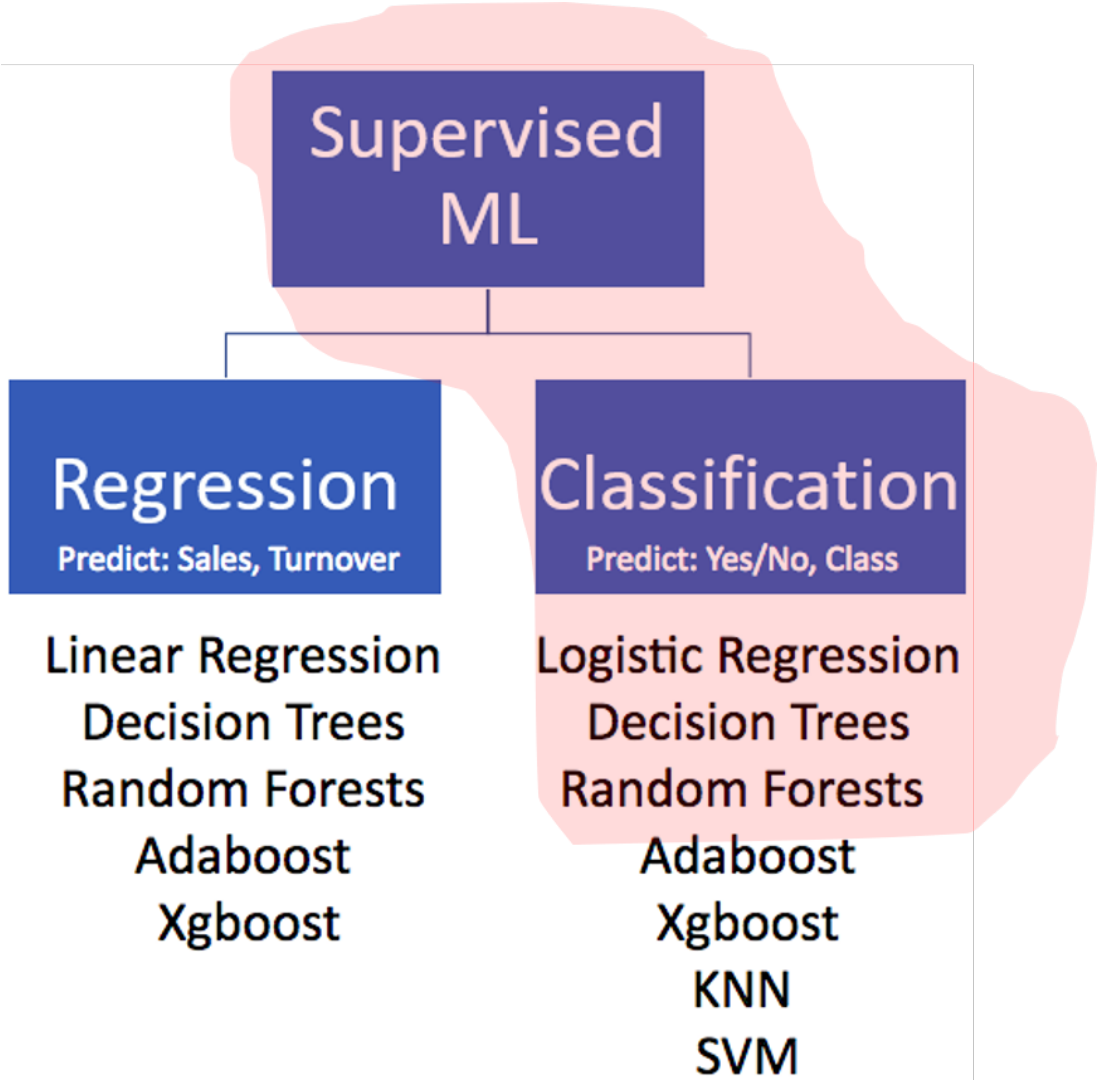
- Over 90% data variance explained with 15 components.
- Feature reduction to 15 from 900!



Modeling: Supervised Machine Learning

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

Three models considered:



Classification Report Training dataset: Logistic Regression

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

```
Misclassified samples: 6
Logistic regression: PCA 15 components
 Classification report for train-dataset:
              precision    recall  f1-score   support

     Healthy       0.98      0.96      0.97       112
   SARS-CoV-2      0.97      0.98      0.97       119

    accuracy                           0.97       231
   macro avg       0.97      0.97      0.97       231
weighted avg       0.97      0.97      0.97       231
```

97 percent accuracy!

Need to be tested with the test split.

Classification Report Training split: Decision Tree

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

```
Misclassified samples: 0
Decision tree: PCA 15 components
 Classification report for train-dataset:
              precision    recall  f1-score   support

     Healthy       1.00      1.00      1.00       110
   SARS-CoV-2      1.00      1.00      1.00       121

    accuracy                           1.00       231
   macro avg       1.00      1.00      1.00       231
weighted avg       1.00      1.00      1.00       231
```

100 percent accuracy! Model might be over-fitting.

Need to be tested with the test split.

Classification Report Training split: Random Forest

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

```
Misclassified samples: 0
Random Forest: PCA 15 components
 Classification report for train-dataset:
              precision    recall  f1-score   support

     Healthy       1.00      1.00      1.00       110
   SARS-CoV-2      1.00      1.00      1.00       121

    accuracy                           1.00       231
   macro avg       1.00      1.00      1.00       231
weighted avg       1.00      1.00      1.00       231
```

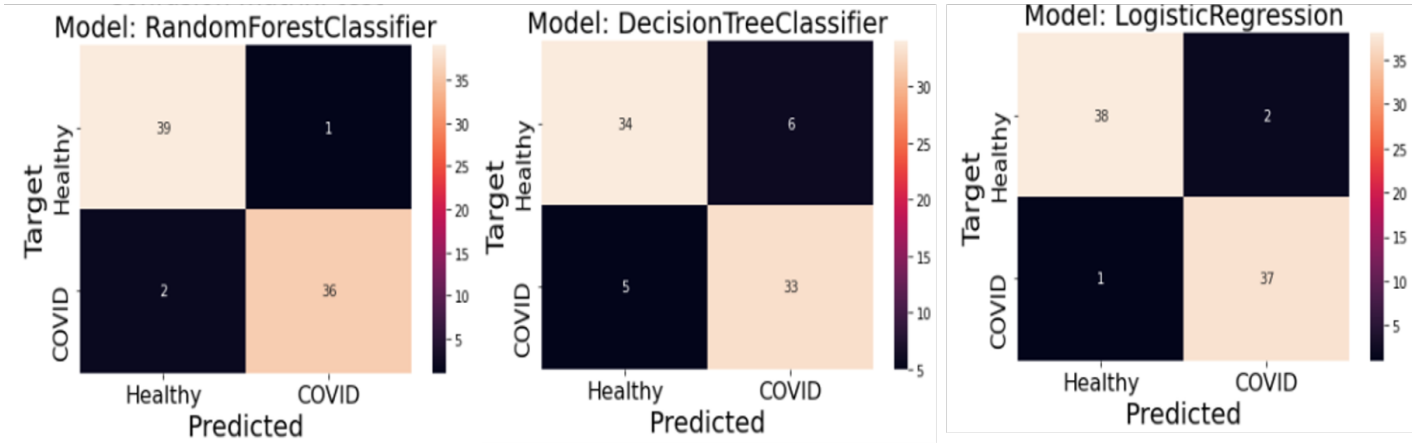
100 percent accuracy! Model might be over-fitting.

Need to be tested with the test split.

Models Testing With Test Split

- All models performed well
- RF and Logistic regression have highest accuracy

Confusion matrix: Test dataset



Accuracy = 96%

Accuracy = 86%

Accuracy = 96%

With PCA physical meaning of features is lost.

Need to model using all features without PCA to gain physical insight on features.

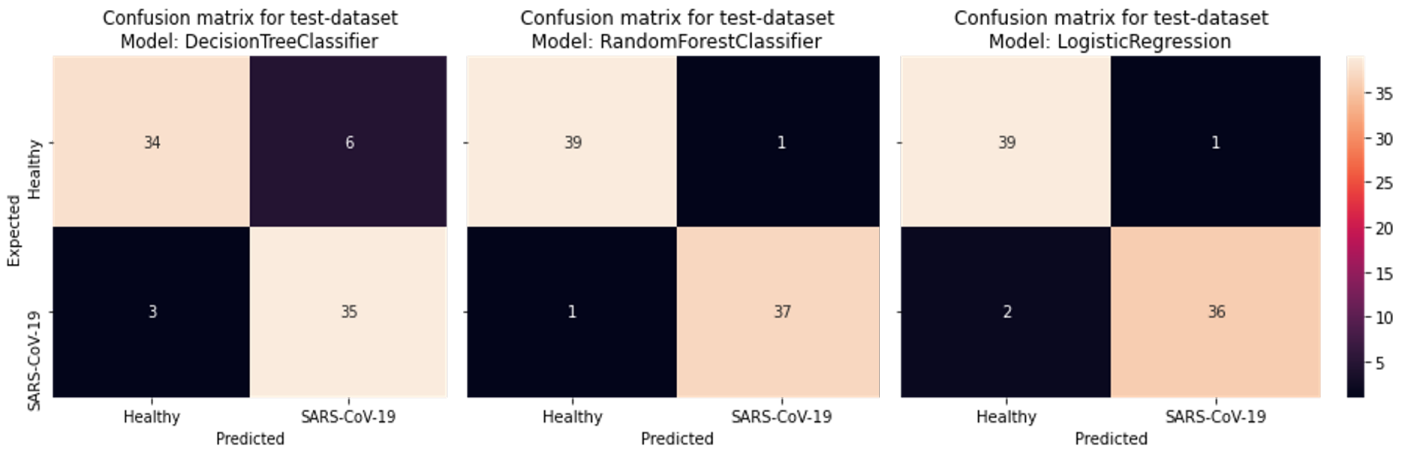
Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

Random forest classifier is the best performing model

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

Though all models persisted with good accuracy:

- Random fores performs the best
- 97% classification accuracy
- Random forest chosen for deployment



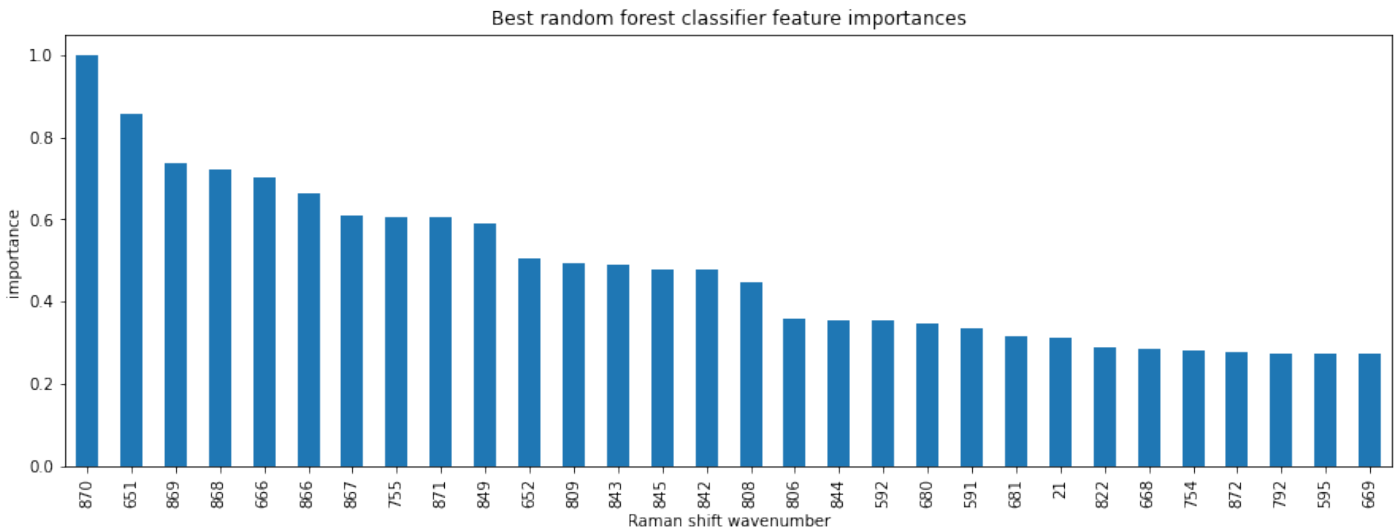
Accuracy = 0.88 Accuracy = 0.97 Accuracy = 0.96

Important features with high predictive power

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

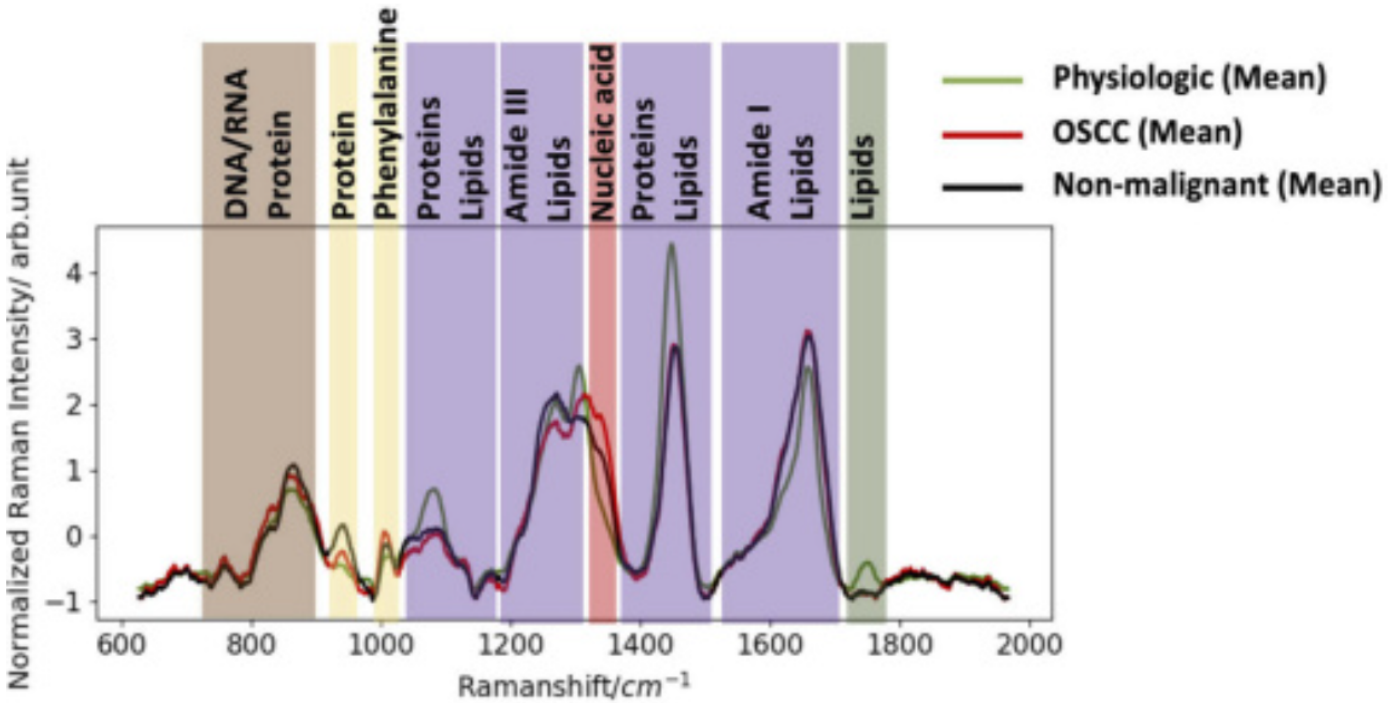
15 features out of 901 are the most important.

- Wavenumber in range [650, 870] has high predictive power
- feature 870 has the highest predictive power



Important features Raman band corresponds to RNA/DNA band

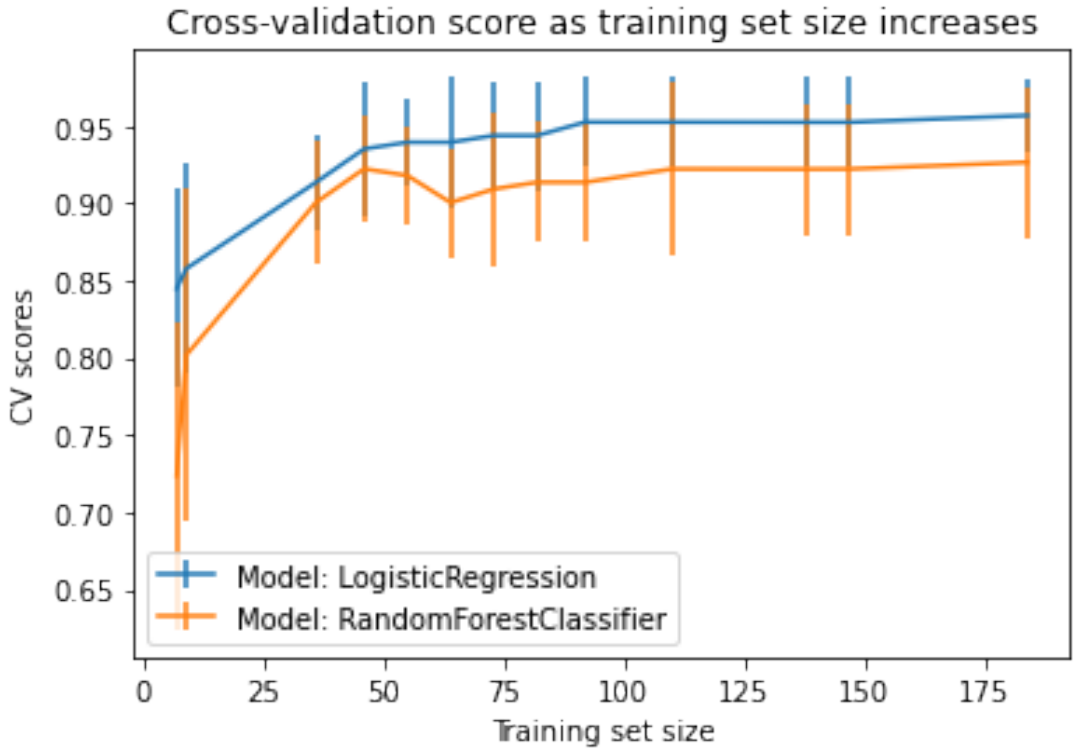
- Virus is an RNA/DNA protein
- Band $[700, 900]$ is prominent for RNA/DNA
- Band corresponds to predicted important features



Do we need more data to enhance model performance?

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

Model accuracy saturates well before the end of available data.



No need for more data.

Summary and Future Work

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

- We developed supervised machine learning models COVID detection using Raman spectroscopy data.
- Logistic regression, decision tree, and random forest supervised machine learning algorithms were considered.
- We find COVID detection using Random forest results in highest detection accuracy of 97 percent.
- Future works needs to be done with covid suspect and covid survived data for more comprehensive and reliable conclusion.

Acknowledgement

Introduction
Current diagnostic method
Raman spectroscopy
Raman and ML
Data Wrangling
Exploratory Data Analysis
Modeling
Summary
Acknowledgement

Springboard mentor: Yuxuan Xin

for time generous and insightful discussions