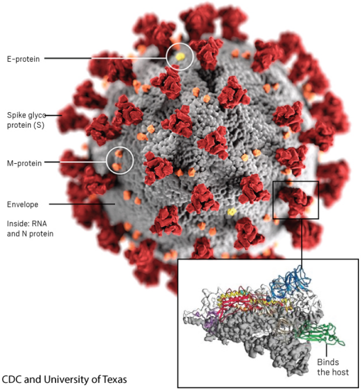
Rapid COVID-19 Diagnosis using Raman Spectroscopy and Machine Learning



CDC and University of Texas

Author: Isaac Ghebregziabher

Capstone Project One

June 3, 2021

Springboard

Side Deck - Capsone 1

Introduction

Current diagnostic method

Raman spectroscopy

Raman and ML

Data Wrangling

Exploratory Data Analysis

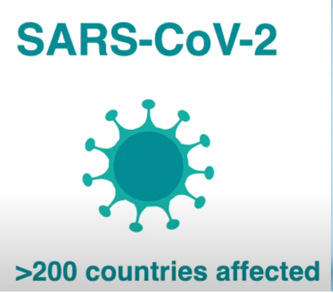
Modeling

Summary

Acknowledgement

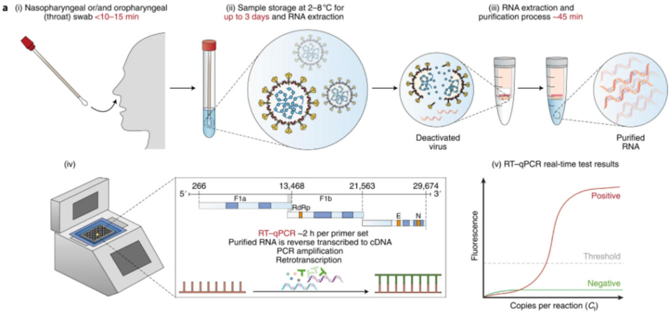
World under COVID-19 pandemic crisis

- > 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

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



a
(i) Nasopharyngeal or/and oropharyngeal (throat) swab <10–15 min
(ii) Sample storage at 2–8 °C for up to 3 days and RNA extraction
(iii) RNA extraction and purification process ~45 min
Deactivated virus
Purified RNA
(iv)
266
13,468
21,563
29,674
5′
3′
F1a
F1b
RdRp
E
N
RT–qPCR ~2 h per primer set
Purified RNA is reverse transcribed to cDNA
PCR amplification
Retrotranscription
(v) RT–qPCR real-time test results
Fluorescence
Positive
Threshold
Negative
Copies per reaction (C_t)

Side Deck - Capsone 1

Introduction

Current diagnostic method

Raman spectroscopy

Raman and ML

Data Wrangling

Exploratory Data Analysis

Modeling

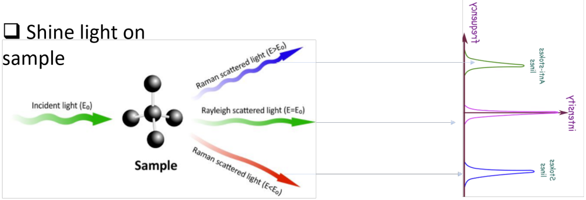
Summary

Acknowledgement

Principle of Raman effect

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

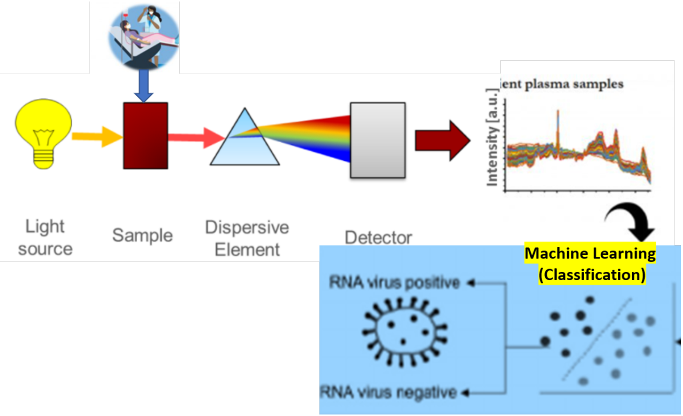
❑ Shine light on sample


Incident light (E_0)
Sample
Raman scattered light ($E>E_0$)
Rayleigh scattered light ($E=E_0$)
Raman scattered light ($E<E_0$)
frequency
intensity
Anti-stokes lines
Stokes lines

❑ Most of the light is unaffected

❑ Small percentage of light undergoes frequency shift

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



Side Deck - Capsone 1

Introduction

Current diagnostic method

Raman spectroscopy

Raman and ML

Data Wrangling

Exploratory Data Analysis

Modeling

Summary

Acknowledgement

Drop features with all values equal to 0

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

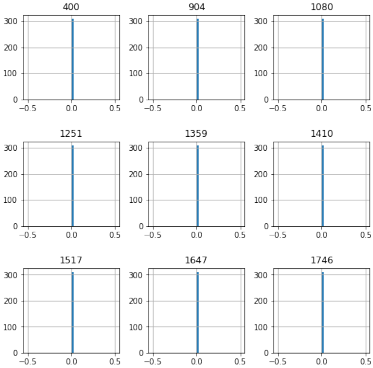


Figure: Single valued features (dropped).

Remaining features have normal distribution

- No concern on feature distributions.
- Most close to normal.
- little skew on several features.

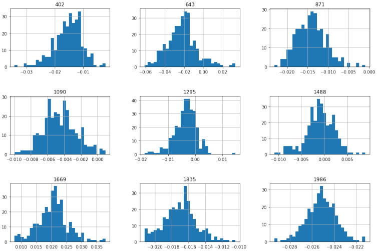


Figure: Close to normal feature distributions.

Dataset is balanced

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

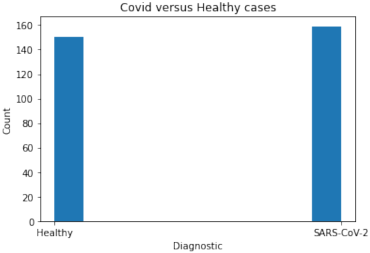
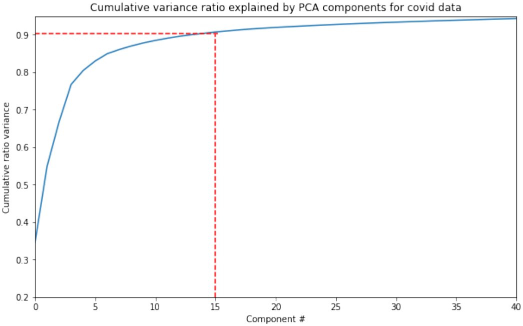


Figure: ≈ 50 : 50 COVID to Healthy class ratio.

Principal component analysis – feature reduction

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

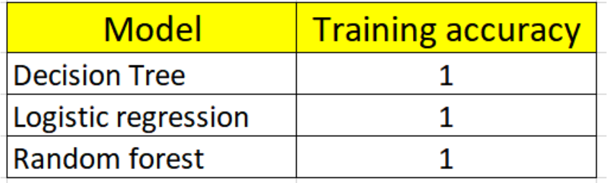


Springboard

Model training and accuracy

Three models considered:

- Decision tree
- Logistic regression
- Random forest



Model	Training accuracy
Decision Tree	1
Logistic regression	1
Random forest	1

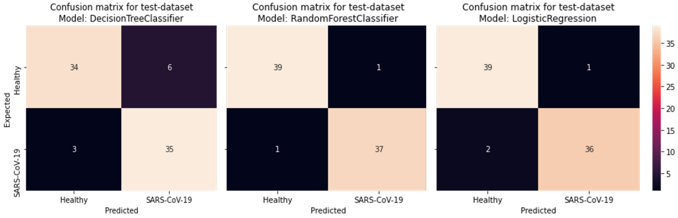
All models seem to over-fit.

Need to be tested with the test split.

Random forest classifier is the best performing model

Though all models persisted with good accuracy:

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



Accuracy = 0.88

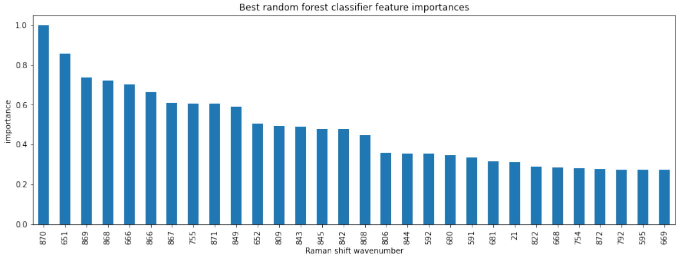
Accuracy = 0.97

Accuracy = 0.96

Important features with high predictive power

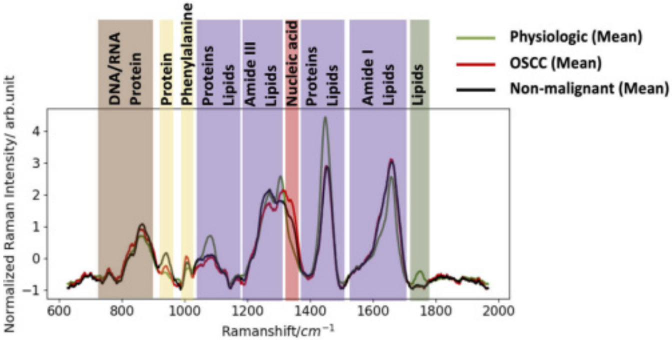
15 features out of 901 are the most important.

- 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

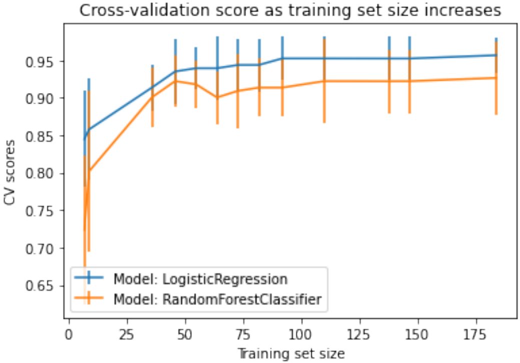
- Virus is an RNA/DNA protein
- Band [700, 900] is prominent for RNA/DNA
- Band corresponds to predicted important features



Springboard

Do we need more data to enhance model performance?

Model accuracy saturates well before the end of available data.



No need for more data.

Summary

- We obtained Raman spectroscopy for COVID detection experimental dataset from Kaggle.
- To get insight We applied data cleansing, wrangling, and exploring techniques.
- We compared and contrasted the performance of Logistic regression, decision tree, and random forest classification models
- We find Random forest to be the best with diagnostic accuracy of 97%

Acknowledgement

Springboard mentor: Yuxuan Xin

for time generous and insightful discussions