



Data Science in Life Science

SS20

Quentin Quarantino



Copyright © Quentin Quarantino

PUBLISHED BY PUBLISHER

BOOK-WEBSITE.COM

Licensed under the Creative Commons Attribution-NonCommercial 3.0 Unported License (the “License”). You may not use this file except in compliance with the License. You may obtain a copy of the License at <http://creativecommons.org/licenses/by-nc/3.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

First printing, March 2019

Contents

I	Part 1	
1	Introduction	7
2	Topic 1: Spreading Models	9
2.1	Background	9
2.2	Data and Methods	9
2.3	Discussion & Results	10
3	Topic 2: Data-based Time Series Prediction	13
3.1	Background	13
3.2	Data and Methods	13
3.3	Results	14
3.4	Discussion	14
4	Topic 3: Risk Factor Analysis	15
4.1	Background	15
4.2	Data and Methods	15
4.3	Results	15
4.4	Discussion	15
5	Topic 4: Diagnostic	17
5.1	Background	17

5.2	Data and Methods	17
5.3	Results	17
5.4	Discussion	23
6	Topic 5: Origin Analysis	25
6.1	Background	25
6.2	Data and Methods	25
6.3	Results	25
6.4	Discussion	26
	Bibliography	29
	Articles	29
	Books	29
	Index	31

Part 1

1	Introduction	7
2	Topic 1: Spreading Models	9
2.1	Background	
2.2	Data and Methods	
2.3	Discussion & Results	
3	Topic 2: Data-based Time Series Prediction	13
3.1	Background	
3.2	Data and Methods	
3.3	Results	
3.4	Discussion	
4	Topic 3: Risk Factor Analysis	15
4.1	Background	
4.2	Data and Methods	
4.3	Results	
4.4	Discussion	
5	Topic 4: Diagnostic	17
5.1	Background	
5.2	Data and Methods	
5.3	Results	
5.4	Discussion	
6	Topic 5: Origin Analysis	25
6.1	Background	
6.2	Data and Methods	
6.3	Results	
6.4	Discussion	
	Bibliography	29
	Articles	
	Books	
	Index	31



1. Introduction

In this weeks project each group member was assigned one overarching topic pertaining to the current Covid-19 epidemic. The current epidemic is globalized, with severe consequences to social, health and economic order. As of today 212 countries are affected, with a total of 4,215,274 confirmed cases and a death toll of 284,672 [9]. Within each topic a short introduction to the general concept is given. The understanding of these concepts is then deepened by real world code examples, showing a glimpse of what is possible in each topic in regards to Covid-19.



2. Topic 1: Spreading Models

2.1 Background

Modeling the spread of infectious diseases is not only an essential tool in understanding the transmission rates and the trajectory of future cases but also has a significant influence on the appropriate guidelines to control the course of an epidemic. The approaches towards modeling the spread can range from computational models (e.g. agent-based) to mathematical modeling (e.g. compartmentalized models). In this short introduction we will focus on a SIR-model which is part of the compartmentalized subgroup. These models follow a deterministic pattern where each subpopulation is divided into groups. In SIR-models each letter stands for one group: $S = \text{susceptible}$, $I = \text{infectious}$ and $R = \text{recovered/death}$. Then it follows that for each time independent point t the rates for each subgroup can be calculated by:

$$\begin{aligned}dS/dt &= \nu N - \beta SI/N - \mu S \\dI/dt &= \beta SI/N - \gamma I - \mu I \\dR/dt &= \gamma I - \mu R\end{aligned}$$

with γ denoting the time rate of death/recovery, β denoting the number of new infections one case causes per time point t , μ denoting general death rate and ν denoting being the birthrate. hey

2.2 Data and Methods

This code is based on the kaggle notebook from <https://www.kaggle.com/lisphilar/covid-19-data-with-sir-model>. It uses python and a basic framework of libraries e.g pandas, sklearn, datetime etc.. The main data used is from the World Health Organization showing novel corona infections by country. Furthermore supplementary data is used to include the age pyramid for each country. The WHO Data set is preprocessed to include the variables: Date, Country, Province, Confirmed, Infected, Deaths and Recovered. A first visualization shows the global rate of infected, deaths and recovered people (Figure 2.1). Next the growth factor is calculated, which is given by: G_n/G_{n-1} with $G = \text{confirmedcases}$. Countries with growth factor higher then one have an increasing number

of cases. In contrast growth factor lesser then one shows a declining number of cases. The actual analysis is done for 5 countries: Italy, Japan, India, USA and New Zealand. Giving one case as an

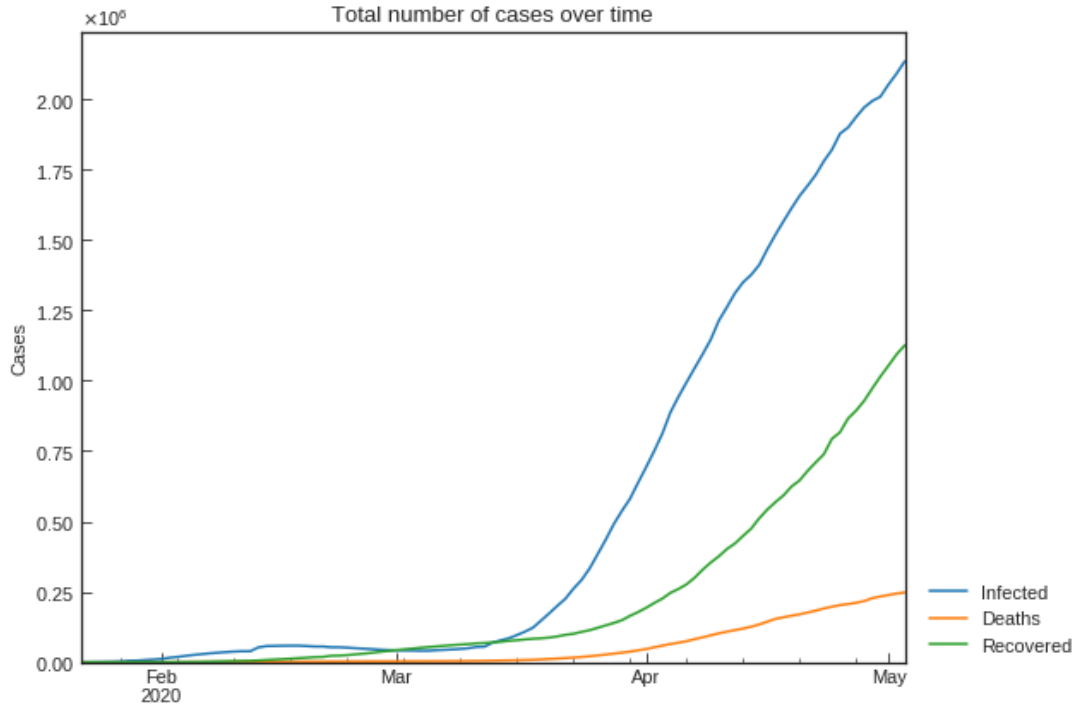


Figure 2.1: Global infections, deaths and recovered people

example as a first step a S-R trend is plotted (Figure 2.2). It shows the trend of susceptible against recovered people. 5 change points can be identified. Next the SIR-F model parameters are estimated for each change point. As a last step the changes in the p value are contrasted with measures taken by the country. While these results are interesting SIR modeling also has its limitations.

2.3 Discussion & Results

Even though the SIR-Model is one of the most basic infectious disease models available it can show promising results with careful consideration for parameter selection and data processing steps. In the case of Italy 3 measures could be shown to reduce the p value: quarantine of person contacted with positive patients, school closure and lock-down. It also has to be considered that the SIR model is based on very basic assumptions. For example the number of susceptible people is treated as fixed as well as the rates of change.

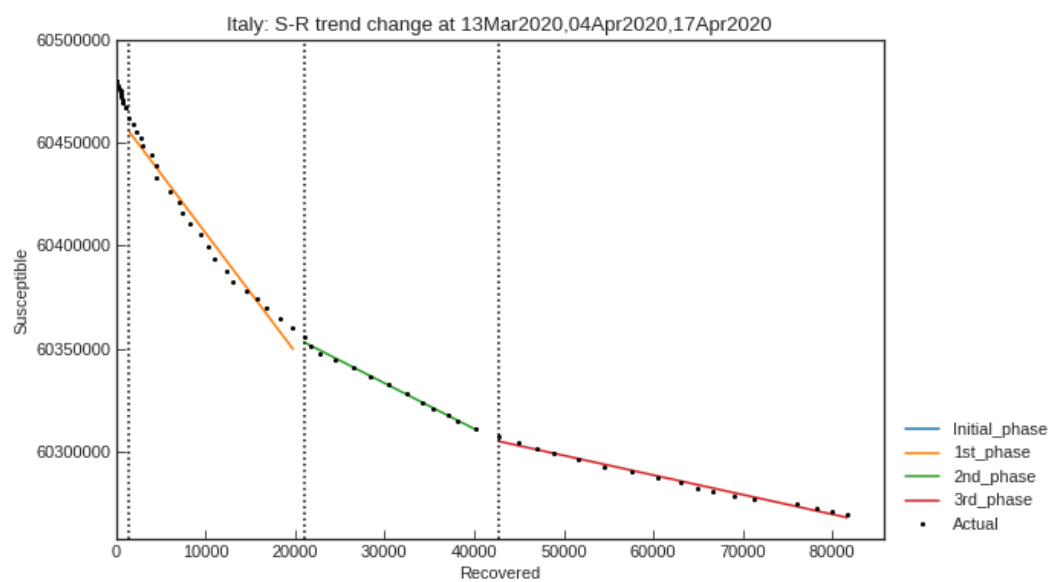


Figure 2.2: Trend of susceptible people versus recovered. 5 distinct change points can be identified.

3. Topic 2: Data-based Time Series Prediction

3.1 Background

Many governments around the world are building their political decisions around the number of current confirmed cases of people infected by COVID-19. Nonetheless, not only the current number of confirmed cases is from greater interest, but also how the virus spreads in the future. One way of forecasting the spread of the virus is by using data-based time series prediction.

3.2 Data and Methods

Therefore machine learning models are calibrated using publicly available data sources like the WHO health report. Time series forecasting can be framed as a supervised learning problem. Other than agent-based spreading simulation such as the SIR model, the models used here do not simulate a population. The forecasting is performed using python's numpy and sklearn libraries. At first the data is downloaded. Since no data points are missing no preprocessing else than converting integers into date times and reorganizing dataframes is performed. The data contains for a wide range of countries the number of infected people per day starting January 22. A support vector machine model is implemented to forecast the number of infected people. The parameters that have been set can be seen in figure 3.1. The test and test training data sets are generated by splitting them without shuffling them, such that the time series is preserved.

```
[ ] 1 # svm_confirmed = svm_search.best_estimator_  
2 svm_confirmed = SVR(shrinking=True, kernel='poly', gamma=0.01, epsilon=1, degree=5, C=0.1)  
3 svm_confirmed.fit(X_train_confirmed, y_train_confirmed)  
4 svm_pred = svm_confirmed.predict(future_forecast)
```

Figure 3.1: Parameters set for SVM Model.

The model has been trained using the first 75 days since January 22. In figure 3.2 it can be seen that the model over estimates the number of confirmed infections by over 1.5 Million.

3.3 Results

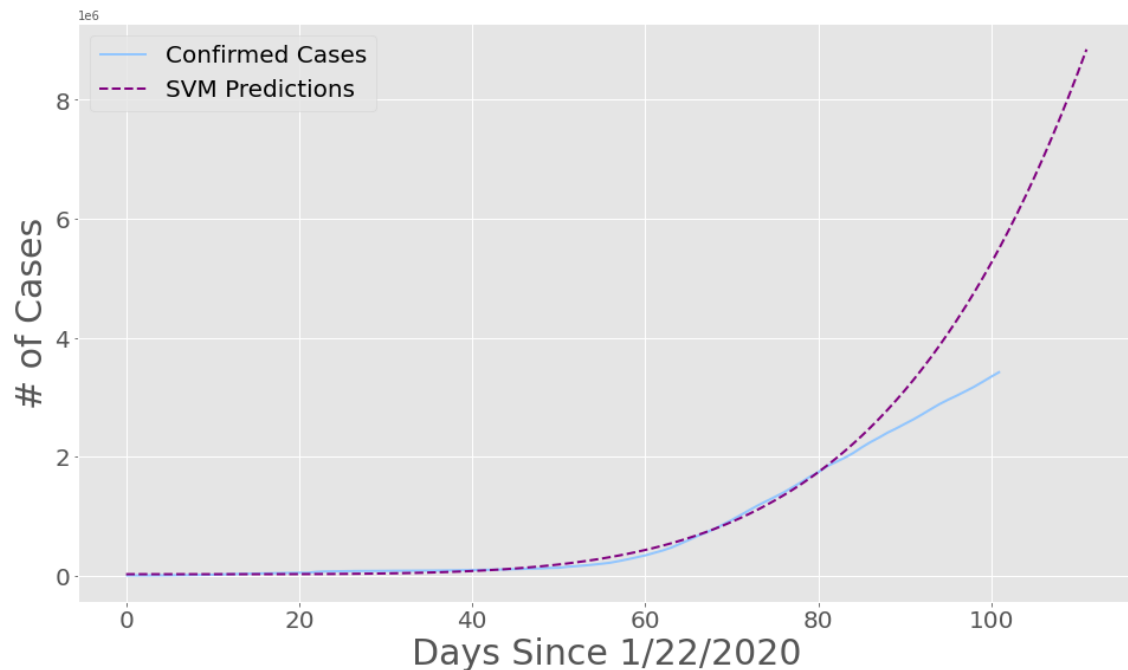


Figure 3.2: Comparison between the observed and the estimated number of infected people.

3.4 Discussion

The shown results are quite underwhelming. This fact has several reasons. One pandemic curves usually increase at the beginning exponentially but then flatten down e.g. because of restrictions in society to decrease the spread of the virus. The model is trained using data from the beginning of the crises where the number of cases rapidly grow. Based on this assumption the estimated number of infected people overshadows the confirmed number. Nonetheless due to different test capacities around the world the estimated might be closer to the real number than it seems to be the case shown in figure 3.2. Still the used model was quite simple and no testing was shown how the parameters were found. A more complex model may gives a better insight to the spread of the virus.



4. Topic 3: Risk Factor Analysis

4.1 Background

The potential dangers of 2019-nCoV have prompted a number of studies on its epidemiological characteristics. It is essential to estimate the number of infections (including those that have not been diagnosed), to be able to analyze the spread of the diseases. To better assess the epidemic risk of 2019-nCoV, among the key parameters to be approximated are the basic reproduction number R_0 and the incubation period. Initially we estimate the cumulative number of cases in China outside Hubei province after 23 January, using a time-dependent compartmental model of the transmission dynamics and then we use that number as an input to the global transportation network to generate probability distributions of the number of infected travellers arriving at destinations outside China. Finally using a Galton–Watson branching process to model the initial spread of the virus.

4.2 Data and Methods

The analysis is performed using the python libraries namely numpy, matplotlib, kiwisolver, scipy and cycler. We computed the risk of the of the individual countries with the selected possible parameters like connectivity and R_{loc} where R_{loc} is the local reproduction number of the infection, Getting all the combination of the variables from the data surrounds the neighbour of the china to generate the Heat map.

4.3 Results

Heat map generated gives the information about the outbreak risks as functions of Θ and R_{loc} , when $C = 200,000$. The arrows show the directions corresponding to the largest reductions in the risk, which is shown in the figure 4.1

4.4 Discussion

By combining three different modelling approaches helps to assess the risk of 2019-nCoV outbreaks in countries outside of China. This risk depends on three key parameters: the cumulative number of

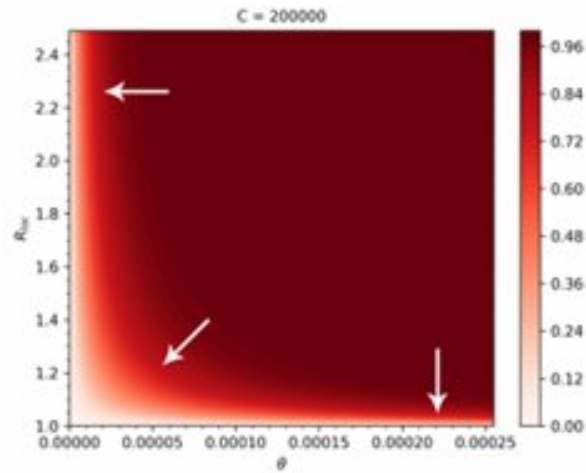
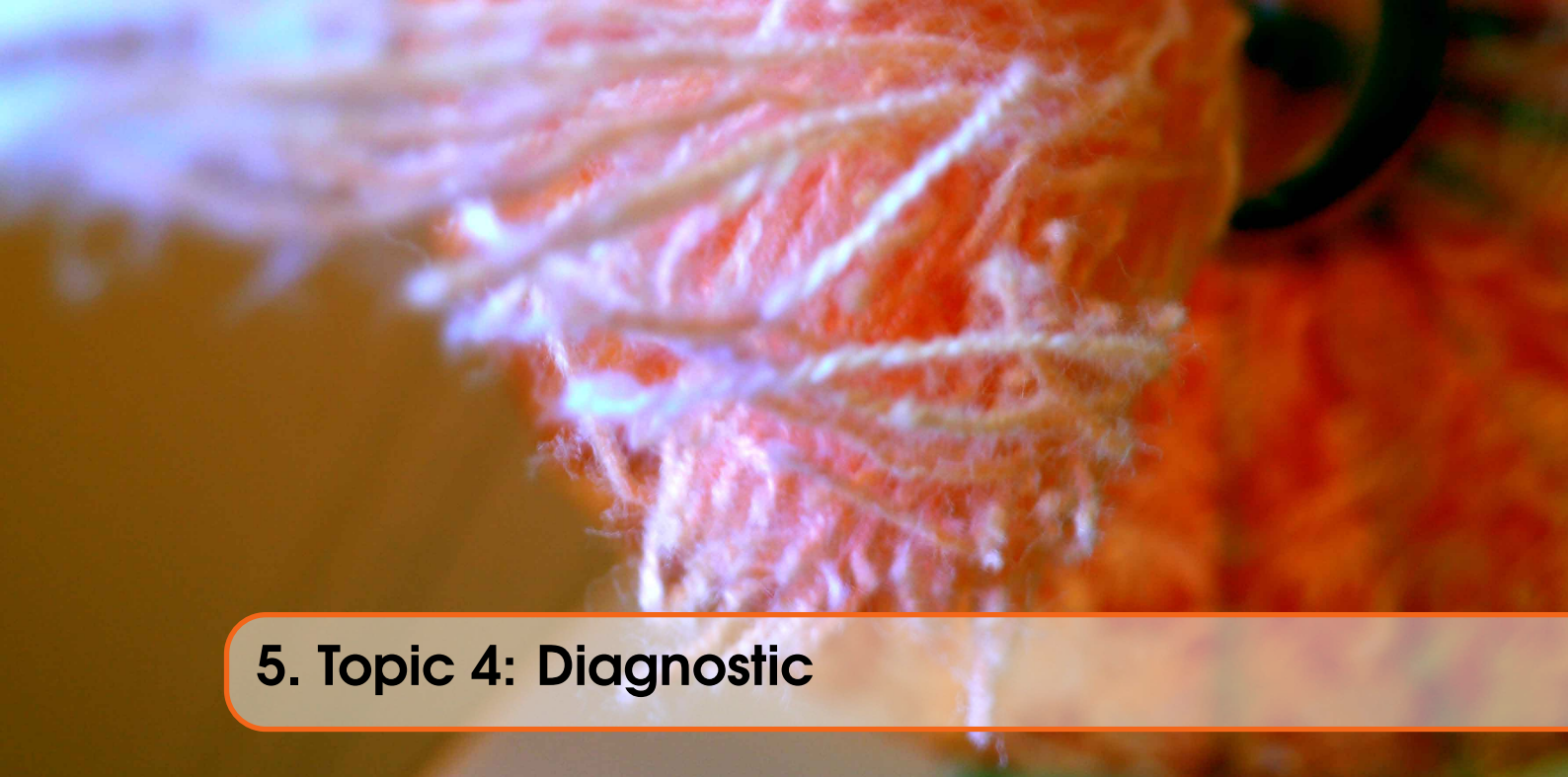


Figure 4.1: Heatmap of the outbreak risks as functions of Theta and Rloc

cases in areas of China which are not closed, the connectivity between China and the destination country, and the local transmission potential of the virus in countries with low connectivity to China but with relatively high R_{loc} , the most beneficial control measure to reduce the risk of outbreaks is a further reduction in their importation number either by entry screening or travel restrictions. Knowing R_{loc} and the generation interval are needed not only to have a better quantitative risk estimation, but also for guidance as to which types of control measures may reduce the outbreak risk the most effective.



5. Topic 4: Diagnostic

5.1 Background

The objective of diagnostics is to help effectively diagnose COVID-19 disease. Diagnostics based on RT-PCR-analysis is not very secure due to a high number of false positives. Diagnosis using X-Ray / CT scan images has objective to help effectively diagnose COVID-19 disease with the help of X-Ray/CT scan images in order to improve speed accuracy and scale of diagnosis.

5.2 Data and Methods

X-Ray Detection method reproduced here is done by training a deep learning model using x-ray images (see Figure 5.1) with TensorFlow and Keras in Python to predict whether a patient has COVID-19. The full list of required tools are here (see Figure 5.2)

5.3 Results

So at first X-Ray data were downloaded from the source and python scripts were downloaded. In the next step, anaconda was installed as it contains a lot of preinstalled packages. In separate environment all the packages listed (see Fig 5.2) were installed with needed help tools and also other needed packages needed (like cuda toolkit and cudnn) to run tensorflow were installed (see Figure 5.3) Then the step augmentation of given X-Ray images was performed for both classes covid positive and normal respectively (see Figure 5.4, 5.5) . In this step using 70 covid and 28 normal X-Ray data were 5088 covid and 2424 normal augmented data generated. In the next step the model was trained and tested using augmented data. The augmented data were divided in train and test data. 80% (6009 data) of augmented data were used as train data and were included in model and 20% (1503 data) of data were used as test data for predictions (see Figure 5.6, 5.7). The model was validated for 1503 test data 100 times with 46 repetitions . Using confusion matrix specificity, sensitivity and accuracy values were estimated and plotted.

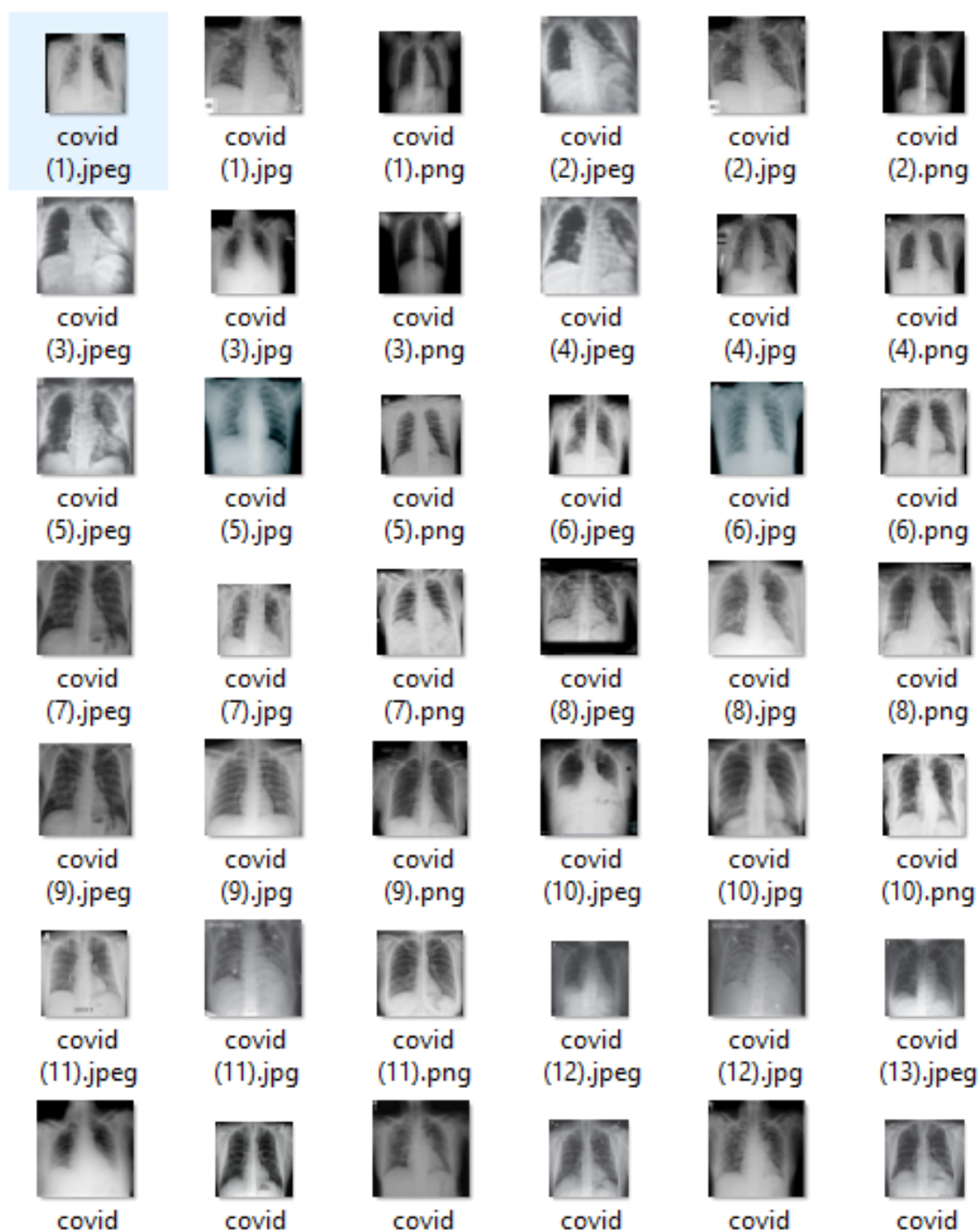


Figure 5.1: Xray data

```
abs1-py==0.9.0
astor==0.8.1
cachetools==4.0.0
certifi==2019.11.28
chardet==3.0.4
cyclr==0.10.0
gast==0.2.2
google-auth==1.11.3
google-auth-oauthlib==0.4.1
google-pasta==0.2.0
grpcio==1.27.2
h5py==2.10.0
idna==2.9
imutils==0.5.3
joblib==0.14.1
Keras==2.3.1
Keras-Applications==1.0.8
Keras-Preprocessing==1.1.0
kiwisolver==1.1.0
Markdown==3.2.1
matplotlib==3.2.0
numpy==1.18.2
oauthlib==3.1.0
opencv-python==4.2.0.32
opt-einsum==3.2.0
pandas==1.0.2
Pillow==7.0.0
protobuf==3.11.3
pyasn1==0.4.8
pyasn1-modules==0.2.8
pyparsing==2.4.6
python-dateutil==2.8.1
pytz==2019.3
PyYAML==5.3
requests==2.23.0
requests-oauthlib==1.3.0
rsa==4.0
scikit-learn==0.22.2.post1
scipy==1.4.1
six==1.14.0
sklearn==0.0
tensorboard==2.1.0
tensorflow==2.1.0
tensorflow-estimator==2.1.0
tensorflow-gpu==2.1.0
tensorflow-gpu-estimator==2.1.0
termcolor==1.1.0
urllib3==1.25.8
Werkzeug==1.0.0
wrapt==1.12.1
```

Figure 5.2: List of required tools


```

45/46 [=====>.] - ETA: 29s - loss: 9.0710e-04 - accuracy: 1.00
46/46 [=====] - 1686s 37s/step - loss: 8.8856e-04 - accuracy
: 1.0000 - val_loss: 1.0893e-04 - val_accuracy: 1.0000
[INFO] evaluating network...

```

	precision	recall	f1-score	support
covid	1.00	1.00	1.00	1018
normal	1.00	1.00	1.00	485
accuracy			1.00	1503
macro avg	1.00	1.00	1.00	1503
weighted avg	1.00	1.00	1.00	1503

```

[[1017  1]
 [  0 485]]
acc: 0.9993
sensitivity: 0.9990
specificity: 1.0000
[INFO] saving COVID-19 detector model...
(myenv) PS E:\DSinLS20>

```

Figure 5.7: Screenshot of Validation process

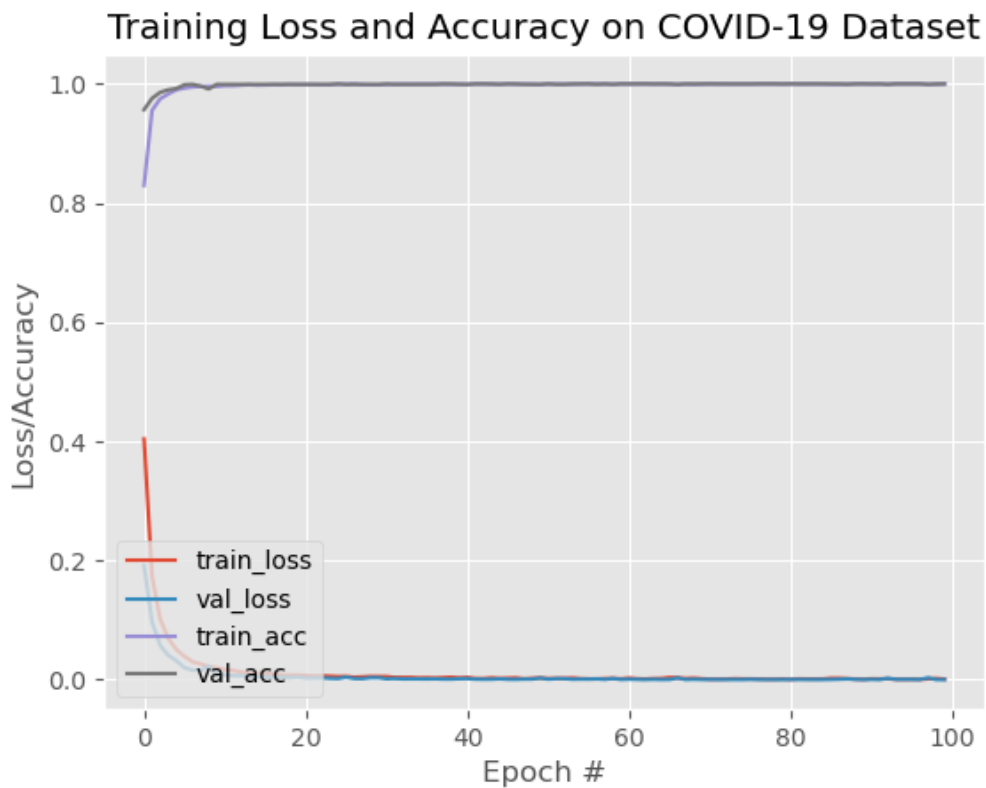


Figure 5.8: Plot of Validation

5.4 Discussion

Approach based on deep Learning described here is a very promising tool for Covid-19 detection in lungs. But on the other hand it is very time consuming. All the steps of this pipeline are very time consuming. Augmentation of images took 2.5 hours. Training and testing using model took about 40 hours. The accuracy, specificity and sensitivity are very high (see Figures 5.7, 5.8) and prove that this approach is very useful.



6. Topic 5: Origin Analysis

6.1 Background

Phylogenetic analysis aims to reconstruct phylogenies both for a group of species and also for the individuals within those species. For pathogens, we are interested in the evolution of the diseases at the genetic level, and what this can tell us about their past and present diversity [8]. An important fact about the Coronaviridae family is that its members tend to “jump” from one species to another. When the transmission occurs from a non-human host to a human host it is called zoonosis [3]. The determination of the most recent common ancestor of the human SARS-CoV-2 and the zoonotic transmission can provide important information about biological features, key mutations and properties of the virus. A detailed understanding of how an animal virus jumped species boundaries to infect humans will help in the prevention of future zoonotic events. [1].

6.2 Data and Methods

We will compare the genetic sequence of SARS-CoV-2 with other viruses of the Coronaviridae family in different hosts. The following analysis is based on a Github repository of Simon Burgermeister [2]. Six complete genomes were considered, whose names and hosts are listed in Table 6.1. The sequence data (fasta files) were downloaded from the NCBI Virus public library [4]. To compare the genetic sequences, a multiple sequence alignment needed to be performed. Clustal Omega is a software that uses seeded guide trees and HMM profile-profile techniques to generate alignments between multiple sequences. Unfortunately, my local computer was not able to compute the alignment due to RAM exceedance. Therefore, I submitted a request to the online version of Clustal Omega [7]. Based on the resulting alignment, a distance matrix was calculated with the *TreeConstruction* package from Biopython. Afterwards, the same package was used to create the phylogenetic tree based on the UPGMA algorithm.

6.3 Results

The resulting phylogenetic tree (Figure 6.1) shows that our human SARS-CoV-2 sequence is most similar to the SARS-like coronavirus sequence of the *Rhinolophus* (horseshoe bat) with a similarity

Accession number	Host	Description
MN996528	H. Sapiens	Human SARS-CoV-2
NC_019843	H.Sapiens	Human MERS-CoV
JQ065048	Anatidae	Ducks, geese and swans
MG772934	Rhinolophus	Horseshoe bats
NC_034972	Apodemus chevrieri	Rodent
KX38909	Gallus gallus	Chicken
MT084071	Manis javanica	Pangolin

Table 6.1: Considered Coronaviridae strains and hosts.

of 96%. The host with the next similar sequence is the *Manis javanica* (Pangolin) with a similarity of 0.89% between their genomes. The human MERS-Cov genome and the SARS-CoV-2 genome share only a sequence similarity of 0.74%.

6.4 Discussion

As many early cases of COVID-19 were linked to the Huanan market in Wuhan [10], it is possible that an animal source was present at this location. Given the similarity of SARS-CoV-2 to bat SARS-CoV-like coronaviruses, it is likely that bats serve as reservoir hosts for its progenitor. Although the similarity of 96% to the coronavirus sequence hosted by the *Rhinolophus*, Andersen et al. [1] identified that its spike protein diverges in the receptor binding domain (RBD), which suggests that it may not bind efficiently to the human ACE2 receptor. Furthermore, it is assumed in this and other studies [5, 6] that an intermediate host was probably involved.

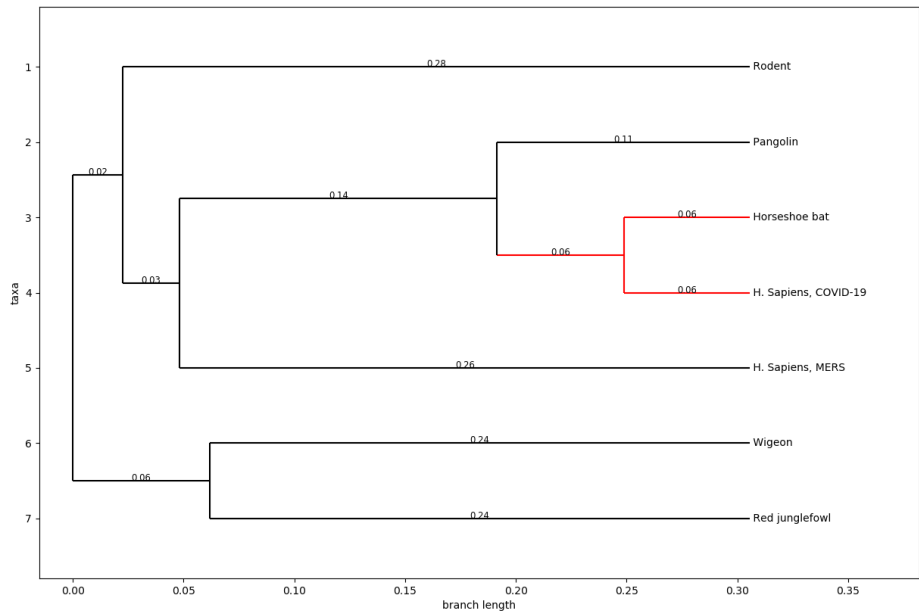


Figure 6.1: Phylogenetic tree of the origin detection analysis.



Bibliography

Articles

- [1] Kristian G Andersen et al. “The proximal origin of SARS-CoV-2”. In: *Nature medicine* 26.4 (2020), pages 450–452 (cited on pages 25, 26).
- [3] Kuldeep Dhama et al. “SARS-CoV-2: Jumping the species barrier, lessons from SARS and MERS, its zoonotic spillover, transmission to humans, preventive and control measures and recent developments to counter this pandemic virus”. In: (2020) (cited on page 25).
- [4] Eneida L Hatcher et al. “Virus Variation Resource–improved response to emergent viral outbreaks”. In: *Nucleic acids research* 45.D1 (2017), pages D482–D490 (cited on page 25).
- [5] Tommy Tsan-Yuk Lam et al. “Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins”. In: *Nature* (2020), pages 1–6 (cited on page 26).
- [6] Zhixin Liu et al. “Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2”. In: *Journal of medical virology* 92.6 (2020), pages 595–601 (cited on page 26).
- [7] Fábio Madeira et al. “The EMBL-EBI search and sequence analysis tools APIs in 2019”. In: *Nucleic acids research* 47.W1 (2019), W636–W641 (cited on page 25).
- [10] Fan Wu et al. “A new coronavirus associated with human respiratory disease in China”. In: *Nature* 579.7798 (2020), pages 265–269 (cited on page 26).

Books

- [8] Michel Tibayrenc. *Genetics and evolution of infectious diseases*. Elsevier, 2017 (cited on page 25).



Index

B

Background 9, 13, 15, 17, 25

D

Data and Methods 9, 13, 15, 17, 25

Discussion 14, 15, 23, 26

Discussion & Results 10

R

Results 14, 15, 17, 25