Covid-19: A Tentative Estimation of Fatality Rates using Random Forest Algorithm

B. K. Praveen Kumar
School of Information Technology,
Jawaharlal Nehru Technological
University (JNTUH)

Gundamaraju Nithya
School of Information Technology,
Jawaharlal Nehru Technological
University (JNTUH)

K. Santhi Sree, PhD
Professor,
School of Information Technology,
Jawaharlal Nehru Technological
University (JNTUH)

ABSTRACT

The outbreak of the Corona Virus Disease (COVID-19) previously known as 2019 Novel Corona Virus, are known to belong to a family of viruses namely the 'Coronaviruses'. These viruses are known to affect both animals and humans. These viruses are responsible for several prevailing infections such as a common cold to life-threatening ailments like Severe Acute Respiratory Syndrome (SARS). COVID-19 is caused by a new virus belonging to this family known as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-Cov-2). This outbreak in December 2019 began in Wuhan, China. The virus spread across 114 countries so rapidly that it has been declared as a "pandemic" by the World Health Organization on 11 March 2020 itself[1] . As of now, there is no cure or vaccination to prevent this infection. The people affected by this virus will have mild to moderate respiratory illness like pneumonia and can recover by receiving supportive care under medical supervision. However, it has been observed that older people and people with a medical history of heart diseases, Diabetes, long-term respiratory diseases, and cancer are more at risk for severe illness. In this paper, the possibility of the death of the Corona infected people by considering the above-mentioned factors is observed. The sample dataset will be analyzed using a machine learning algorithm for this purpose. The goal is to predict the death probability of a patient based on his age and previous medical history with the highest accuracy.

General Terms

Machine Learning**,** Random Forest algorithm.

Keywords

Coronavirus, SARS-CoV-2, pandemic, death rate, Machine Learning, accuracy.

1. INTRODUCTION

Machine Learning can be defined as a form of data analytical methodology employed to impart the ability of decision-making to an electronic device in such a way as to predict a pattern when provided with new and unseen data. It is seen as a constituent of Artificial Intelligence which deals with the simulation of human intelligence in machines. Alternatively, Machine Learning can be defined as the study of techniques used to build an artificially intelligent machine. Machine Learning algorithms generally build an analytical model based on sample data. This data model is then used to predict a pattern in the given subsequent data instead of conventional programming. Since these algorithms involve working with a lot of data, it makes use of Statistical Principles and Computational Theory. The level of accuracy of a trained system will vary based on the number of documents used during the training phase and the coverage of the specific jargon in those documents. The system must also be retrained frequently to maintain the same level of quality. Information overload will slow down the whole system while overfitting it will make the system less accurate; in other words, the wrong selection of documents will cause a decrease in quality. This means that a limit is imposed to the level of improvement and it is often very unclear why the system has improved or how you can improve it further.

1.1.Methods Used in Machine Learning

Machine Learning has different learning algorithms to solve problems[4][5][6]. These approaches completely depend upon the type of data they input, output and the way data intended to solve. They are widely divided into 4 types of learning algorithms:

1.1.1 Supervised learning algorithm

Supervised learning algorithms construct a model of a sample dataset which understands the various associations in the data. The sample dataset contains labelled data which teaches the model how to behave and is hence named as "*Supervised Learning*". The model obtained during the learning data is used to predict a pattern or rules when given a new set of data. Classification and Regression are some important supervised learning algorithms. Classification is used to group the data with similarities into different sections to classify them whereas regression is useful to predict future behavior like the changes in stock market prices or the probability of any similar event. The prominent difference between Classification and Regression is that Regression predicts a numerical value rather than a class label.

1.1.2 Unsupervised learning algorithm

Unsupervised Learning mainly deals with the unlabeled data. Taking a set of data that contains formless data and finding the structure of data by clustering or grouping of data points comes under "*Unsupervised Learning". These algorithms learn from raw and label-less data, which is not categorized. Unsupervised Learning algorithms try to minimize the dissimilarities in the data and acts mostly on the presence and absence of the associations. They manage more complex tasks when compared to other algorithms. Unsupervised learning is mostly used in clustering techniques which are in turn useful in anomaly detection and association learning.

1.1.3 Semi-supervised learning algorithm

Semi-supervised learning is a combination of both supervised and unsupervised approaches. The learning process is not closely supervised for every single input, but we also do not let the algorithm learn on its own without providing a form of feedback. An algorithm called Generative Adversarial Networks (GANs) uses two neural networks, a generator and discriminator to generate input and outputs [6]. Since there is no need for us to provide explicit labels every single time it can be classified as semi-supervised.

1.1.4 Reinforcement learning algorithm

Reinforcement learning algorithms use rewards and punishments rather than feedback. The algorithm uses greedy approach which optimizes by either increasing reward points or reducing the punishments to reinforce performance. This reward-motivated behavior is key in reinforcement learning.

1.2.Applications

1.2.1 Spam Detection

Spam mails are identified by recognizing specific words or phrases from old mails and if the pattern is matched in the newly received ones the mail is sent to spam folder. This is used by almost every online mailing service.

1.2.2 Face Detection

From a given number of photos faces are identified and then automatically tagged. If next time, the same face is detected in a new photo, it provides us with the previous tag. Face Detection is used in social media sites and security apps.

1.2.3 Credit Card Fraud Detection

According to customers' past transactions, if there are any inappropriate purchases made, then the customer is warned immediately about the purchase and necessary action is taken.

1.2.4 Digit Recognition

A camera attached to the machine detects postal codes that are handwritten and arranges all the letters according to the geographical locations they must reach. This application requires machine learning to recognize handwritten numbers and transform them into digital signs.

1.2.5 Speech Understanding

Algorithms which can understand, and process normal speech of humans are being used widely. Examples include Apple's Siri and Google Voice Assistant.

1.2.6 Product Recommendation

Using machine learning algorithms, companies like Amazon, Flipkart try to attract more customers by personalizing the website using their previous searches and transactions.

1.2.7 Medical Diagnosis

For detecting diseases more accurately, hospitals use devices that could decide whether a person is affected with any diseases for symptoms he/she has, with the help of complete data about all diseases and symptoms.

1.2.8 Stock Trading

Algorithms which help the customers decide whether to hold or sell the stock are being used widely. They predict the future value of stocks and thus lets the customer take an informed decision

1.2.9 Customer Segmentation

With the help of past behavior patterns, the ML algorithms try to predict the number of users who will choose the paid versions from the trial versions. Amazon Prime has implemented this technique.

1.2.10 Financial Service

Companies can identify the company insights of financial sector data and can overcome the occurrence of financial fraud. It is used to identify opportunities for investment and trade. We can also prevent the financial risks prone institutions by using cyber-surveillance and take necessary actions to prevent fraud.

1.2.11 Transportation

By the travel history and pattern of travelling in various routes, machine learning helps in Transportation Companies to predict the problem in the routes and advice their customers to choose the best route.

1.2.12 Computational Intelligence

For many years computational intelligence is being developed actively. Constant improvements are being carried out on classical methods. These days computational intelligence is being used for many applications directly or indirectly.

2. PROPOSED WORK

2.1.Algorithm Used: Random Forest

Choosing the appropriate algorithm that fits the issue we are dealing with is the most important task. To work with our dataset, best approach would be supervised learning. Since the patients are classified, Classification algorithms are preferred[9]. Classification is the process of predicting the target label or category of given datasets for categorical values. It can classify an input variable as well. There are many simple classification algorithms such as regression, linear classifier etc. but since the required result needs to be as accurate as possible, we must choose efficiency over simplicity. Thus, our choice would be limited to powerful algorithms like Decision Trees and K-Means. Firstly, the choice was Decision Tree algorithm because of its simplicity. But in the trials, the accuracy obtained was 85%. So, the alternative was to go with the Random forest algorithm which is an efficient version of the decision- tree and produces liable results.

Random Forest[10] creates multiple decision-trees which resemble 'a forest' This helps in obtaining a more accurate prediction because it involves the combined efficiency of all the trees. Overfitting is a problem faced by some of the algorithms wherein the model tends to fit the new data more closely to the training data and fails to accommodate the patterns in new data. Random forest algorithm prevents overfitting as it creates several models in the form of trees from random subsets of the features. To train and test the model, dataset is split into training and the test dataset. The model will learn from the training dataset to build a model and the test dataset will be classified using the model. After a basic model is built, analysis of the relationship between the independent variables and the target variable takes place.

2.2.Flowchart:

The main purpose of the model is to predict the death possibility of infected people. So, the focus is on the accuracy of training data and the Model. Proposed work can be depicted in the flow chart as follows.

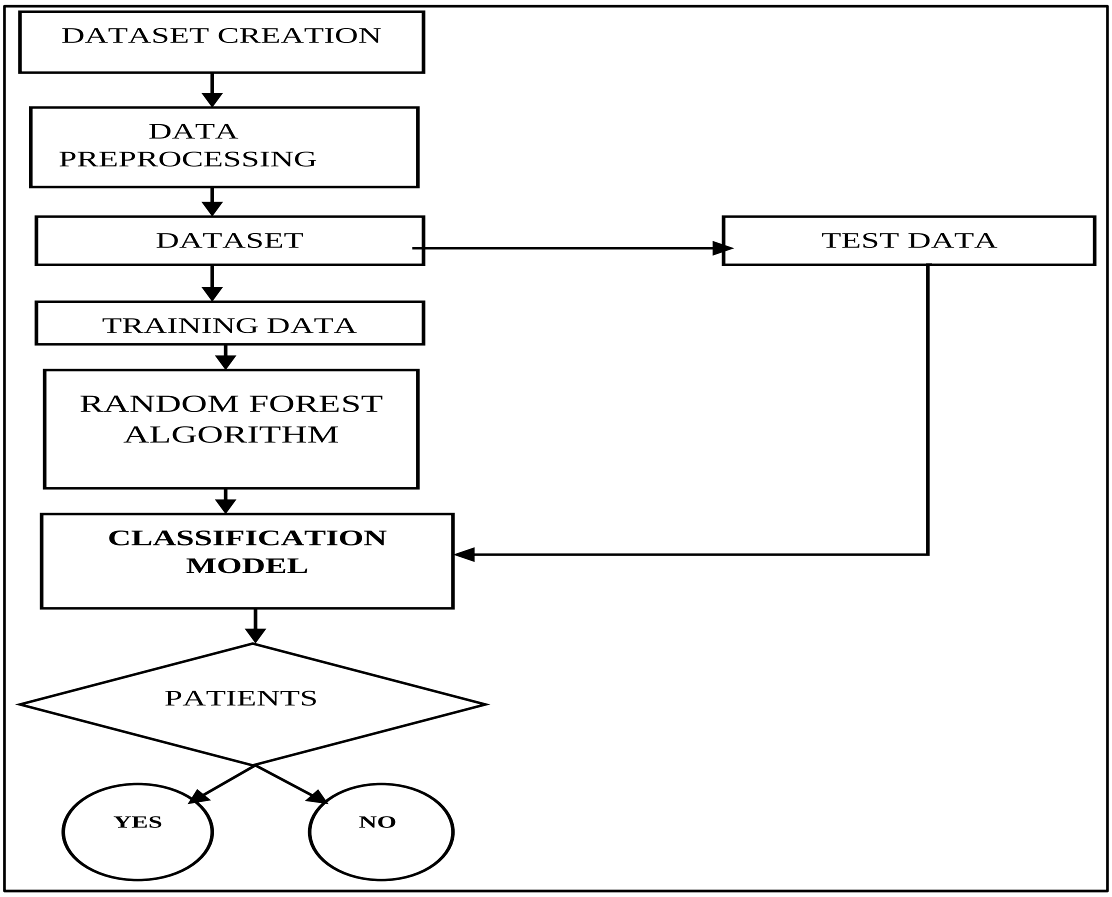

Figure-1: Flowchart

3. CORONA DATASET

The most common symptoms of COVID-19 are fever, tiredness, and dry cough. Some patients may have aches and pains, nasal congestion, runny nose, sore-throat or diarrhea. These symptoms are usually mild and begin gradually. Some people become infected but do not develop any symptoms and do not feel ill. Most people (about 80%) recover from the disease without needing special treatment. Around 1 out of every 6 people who get COVID-19 becomes seriously ill and develops difficulty in breathing[1]. In the first big analysis of more than 44,000 cases from China, deaths were at least five times more common among confirmed cases with diabetes, high blood pressure or heart or breathing problems. Even though patterns in the death rates among confirmed cases can tell us who is most at risk, they cannot tell about the precise risk in any single group. A sample dataset based on the above factors with about 250 samples with various health conditions and of varying age. The dataset prepared is based upon the observations found online[11]. We tried to base the dataset as close to the real data as possible.

Table-1: Relation between Age and Death Rate

AGE	DEATH RATES
80+	14.8%
0-79	8.0%
0-69	3.6%
0-59	0.4%
40-49	1.3%
30-39	0.2%
20-29	0.2%
10-19	0.2%
0-9	0

Table-2: Relation Between Death Rate and Previous Medical History

PRE-EXISTING CONDITION	DEATH RATE *All cases*
Cardiovascular disease	10.5%
Diabetes	7.3%
Chronic respiratory diseases	6.3%

Hypertension	6.0%
Cancer	5.6%
No pre-existing conditions	0.9%

3.1.Sample Dataset:

The original dataset which is used in the research consisted of 250 columns including almost every possible variation and replicates the real-world data as much as possible.

Table:3 A Sample of the Dataset

P_ID	Age	Travelled?	corona	sore throat	dry cough	cold	fever	Respiratory	Heart Diseases	Hypertension	Diabetes	Death?
1	90	0	0	1	1	0	0	0	0	1	1	no
2	55	1	1	1	1	1	1	0	1	1	1	no
3	76	1	0	0	0	0	0	0	1	1	1	no
4	1	0	0	0	0	0	0	0	0	0	0	no
5	41	0	1	1	1	1	1	0	0	0	0	no
6	9	1	0	0	0	0	0	1	0	1	0	no
7	25	0	1	1	1	1	1	1	0	1	0	no
8	18	1	0	0	0	0	0	1	0	0	0	no
9	5	0	0	0	0	0	0	1	0	1	1	no
10	77	0	0	0	1	0	0	1	0	1	1	no
11	70	1	1	1	1	1	1	1	1	1	0	yes
12	46	0	0	1	0	0	1	0	0	1	0	no
13	99	0	1	1	1	1	1	0	0	0	1	no
14	65	0	1	1	1	1	1	1	1	1	1	yes
15	34	0	0	0	0	0	0	1	0	1	0	no
16	63	0	0	0	1	0	0	0	1	1	1	no
17	49	1	1	1	1	1	1	1	1	1	0	yes
18	21	1	0	0	0	0	0	0	0	0	1	no
19	11	1	1	1	1	1	1	1	0	0	1	no
20	11	0	0	0	0	0	0	0	0	0	0	no

4. EXPERIMENTAL RESULTS

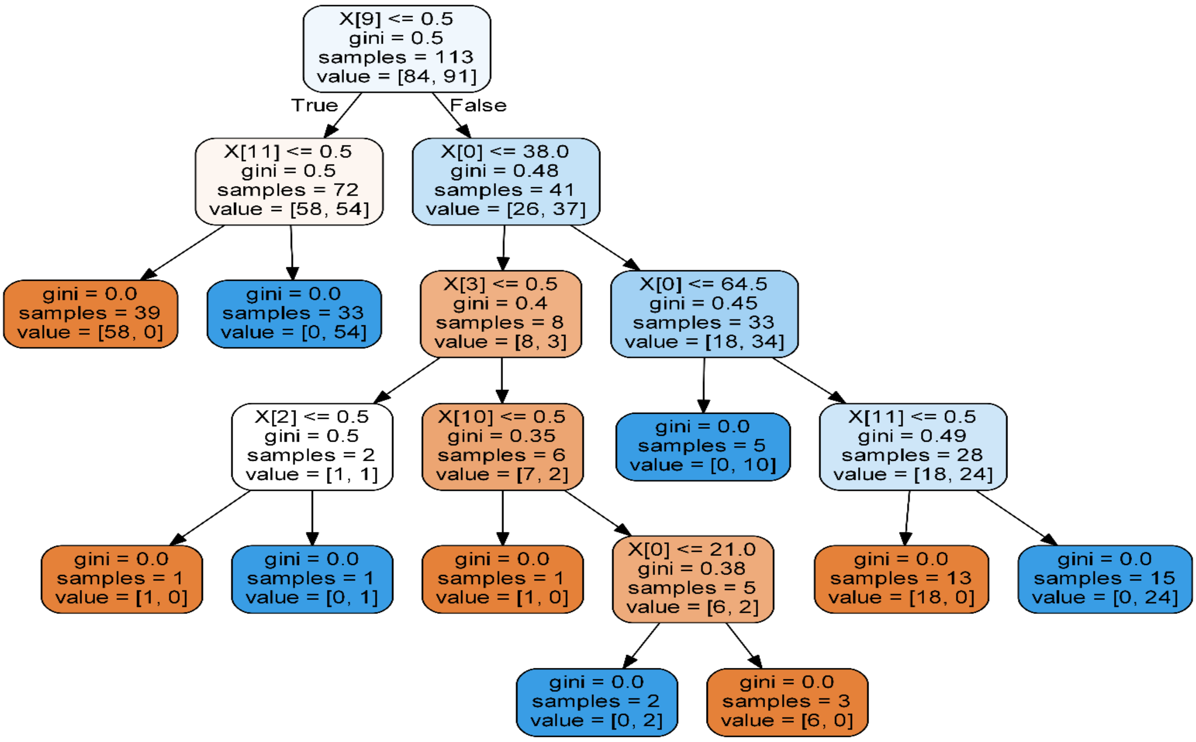

Figure 2: Experimental Result of Random Forest Algorithm Cut Down to Five Trees

5. ANALYSIS

On analysing the data, the death rate was observed to be around 5% of the total affected. This can be observed in the scatterplot below.

5.1.Respiratory Diseases:

The patients suffering from diseases like Asthma, Bronchitis etc are more likely to succumb to the virus. This might be because of their low resistance towards the virus attack.

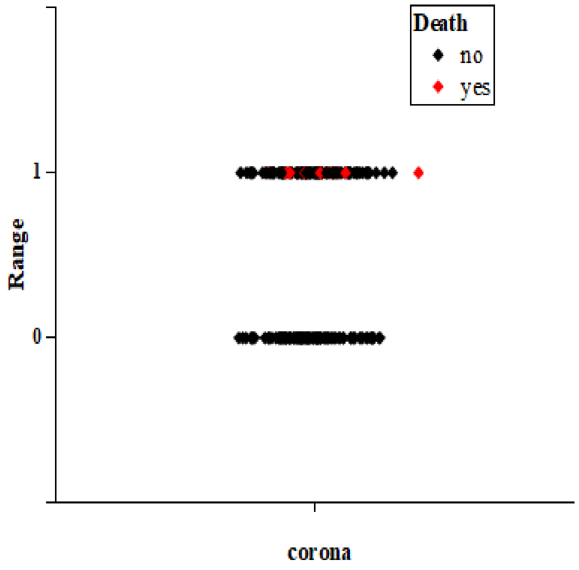

Figure 3: Corona vs Death

5.2.Age Factor:

It has been observed that older people are likely to be more susceptible to the virus than younger people. This can be explained using the graph below

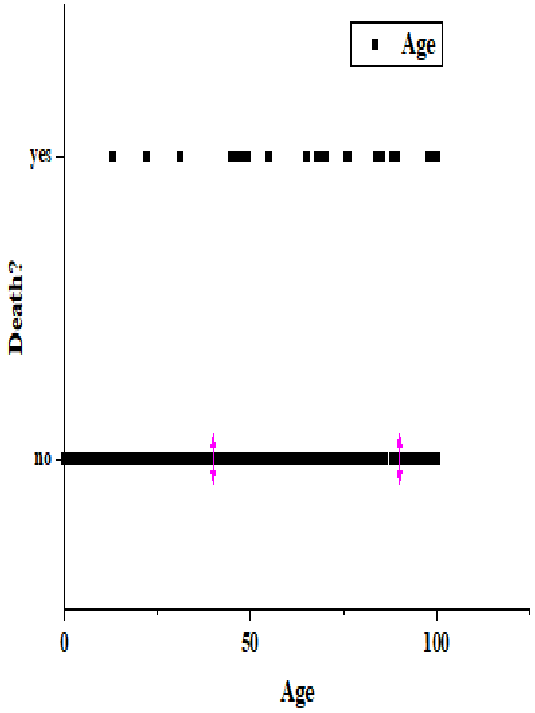

Figure 4: Distribution of Death Rate Based on Age

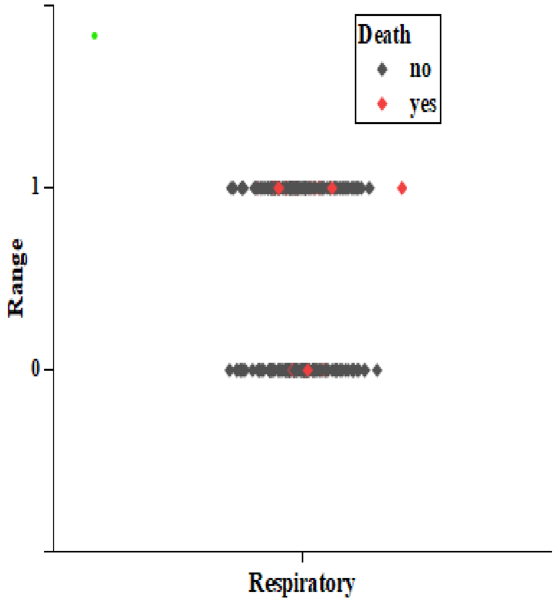

Figure 5: Impact of Respiratory Diseases

5.3.Heart Diseases:

Patients suffering from prolonged cardiovascular diseases tend to be resistant to the treatment and thus the death rate is high.

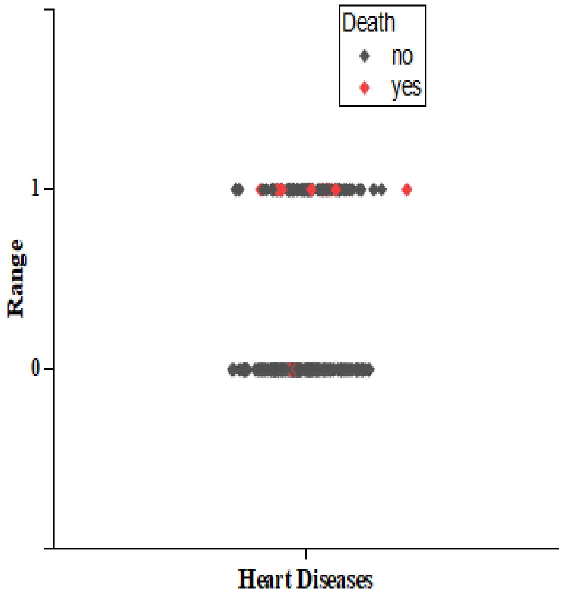

Figure 6: Impact of Cardiovascular Diseases

6. CONCLUSION

In this paper, Machine Learning techniques are incorporated to predict the risk of death for COVID-19 virus-infected people. Since this is an ongoing epidemic as of mid-March, the information regarding the number of patients or the affecting symptoms is experimentally not proven. The data was compiled based on the observations released by various sources. Then a classification model was built by applying the random forest. Using this model, the probability of the death risk for the corona virus-infected could be predicted based on their age and history of any diseases like diabetes or respiratory ailments. On analyzing the data further, the death risk was found to be higher in people with respiratory ailments or cardiovascular diseases. This might be because of their low resistance towards the virus attack. We also observed that the overall death rate accounts up to just 5% of the total patients. This shows that though the growth of infected people is exponential, the actual number of people who died due to this disease is still low. It also confirms the speculations that older people are more susceptible to death rather than the young. However, all these are just statistical observations and not scientifically proved. Regarding the algorithm, we have achieved an accuracy of 100% for the test data. This accuracy was because we took the forest size to be 100 trees. It was time-consuming but since the data was dealing with the death-risk prediction accuracy was of more importance, it was mandatory. Thus, we successfully predicted the death risk of corona affected patients and are hopeful that our model would be helpful in further studies of the novel disease.

7. REFERENCES

[1] https://www.who.int/emergencies/diseases/novelcoronavirus-2019

[2] Zou, Quan & Qu, Kaiyang & Luo, Yamei & Yin, Dehui & Ju, Ying & Tang, Hua. (2018). Predicting Diabetes Mellitus with Machine Learning Techniques. Frontiers in Genetics. 9. 10.3389/fgene.2018.00515.

[3] Hegazy, Osman & Soliman, Omar S. & Abdul Salam, Mustafa. (2013). A Machine Learning Model for Stock Market Prediction. International Journal of Computer Science and Telecommunications. 4. 17-23.

[4] Mohri, Mehryar; Rostamizadeh, Afshin; Talwalkar, Ameet (2012). Foundations of Machine Learning. The MIT Press. ISBN 9780262018258.

[5] C.M. BishopPattern recognition and machine learning,Springer, New York (2006)

[6] T.M. Mitchell,The discipline of machine learning: Carnegie Mellon University,School of Computer Science, Machine Learning Department (2006)

[7] I.H. Witten, E. Frank,Data mining: practical machine learning tools and techniques,Morgan Kaufmann (2005)

[8] Wikipedia contributors. "Machine learning." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 22 Apr. 2020. Web.24 Apr. 2020

[9] Douglas, Pamela K. et al. “Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief.” NeuroImage 56 (2011): 544-553.

[10] L. Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[11] https://www.worldometers.info/coronavirus/