Prediction of COVID-19 in Vietnam using Markov Model

Thai-Duy Dang
Honors Program, Faculty of Information Technology
University of Science, Ho Chi Minh City, Vietnam
Vietnam National University Ho Chi Minh City, Vietnam
19120491@student.hcmus.edu.vn

Chi-Hao Ha
Honors Program, Faculty of Information Technology
University of Science, Ho Chi Minh City, Vietnam
Vietnam National University Ho Chi Minh City, Vietnam
19120219@student.hcmus.edu.vn

Duy-Nam Mai
Honors Program, Faculty of Information Technology
University of Science, Ho Chi Minh City, Vietnam
Vietnam National University Ho Chi Minh City, Vietnam
19120298@student.hcmus.edu.vn

Phuong-Dinh Tran
Honors Program, Faculty of Information Technology
University of Science, Ho Chi Minh City, Vietnam
Vietnam National University Ho Chi Minh City, Vietnam
19120476@student.hcmus.edu.vn

***Abstract*—The World Health Organization (WHO) declared on March 12, 2020, the COVID-19 disease as a pandemic. In Vietnam, the first infected case was detected on January 23, 2020, and until now, the number of confirmed cases has gradually increased to reach nearly 2,000,000. To predict the COVID-19 evolution, scientists have come up with many different methods and models. The most common is using statistical and mathematical models such as generalized logistic growth model, exponential model, segmented Poisson model, Susceptible-Infected-Recovered derivative models, and ARIMA have been proposed and used. Herein, we proposed the use of the Markov model, which is a statistical system modeling transitions from one state (confirmed cases, recovered, active or death) to another according to a system of differential equations to forecast the evolution of COVID-19 in Vietnam from October 1, 2020 until now. Forecasts for the cumulative number of confirmed, recovered, active and death cases can help the authorities to set up adequate protocols for managing the post-confinement due to COVID-19. We provided both the recorded and forecast data matrices of the cumulative number of these cases through the range of the studied dates.**

***Index Terms*—Markov model, COVID-19 spreading, statistical modeling, forecast diseases**

I. Introduction

A. Motivation

Since the COVID-19 began, the Americas, Europe, South-East Asia and Eastern Mediterranean regions have the most documented cases. Globally, nationally, and at every sub-governmental level, there is a need to monitor the current caseload and predict the number of future cases to guide public health awareness, preparedness, and response. Social orders need to manage many major problems, for example, guaranteeing satisfactory supplies of individual defensive hardware, contemplations about the sufficiency of the medical care, labor force and other medical services assets, just as how to offset prohibitive security rules with keeping organizations open and the economy sound. For an original irresistible illness, it is particularly critical to predict future cases in view of what has occurred in the immediate past.

B. Challenges

Choosing the suitable data for the Markov model is the challenge we have to deal with. After considering various data sources, we decided to collect the data from the government official COVID-19 statistics website [1], which has the trustworthy data constantly updated.

C. Scope

Specifying the scope of the project is the utmost important factor in building a COVID-19 prediction model, due to the complexity and the huge scale of a global pandemic. For the sake of familiarity and ease of searching, we decided to build the model based on the COVID-19 cases data from Vietnam, taken from the government official COVID-19 statistics website [1]. After initial data preprocessing, the dataset we ended up with contains 109 dates of Vietnam's COVID-19 cases, death, and recovery, spanning from October 1, 2020 to January 18, 2022. We will use this data to build a COVID-19 prediction model using the Markov model in the span of one month so that we can have an adequately accurate view of the future situation of COVID-19 in Vietnam.

II. Data

A. Value of Data

Our data provide forecasts for the cumulative number of confirmed, recovered, active and death cases, which is important for both monitoring and control the COVID-19 spreading in Vietnam. The data was based on the predicted values from October 2020 to present, the authorities can benefit from these data to set up adequate protocols for the post-confinement. Moreover, other researchers can use this data for comparison and further meta-analysis of the COVID-19 spreading in Vietnam.

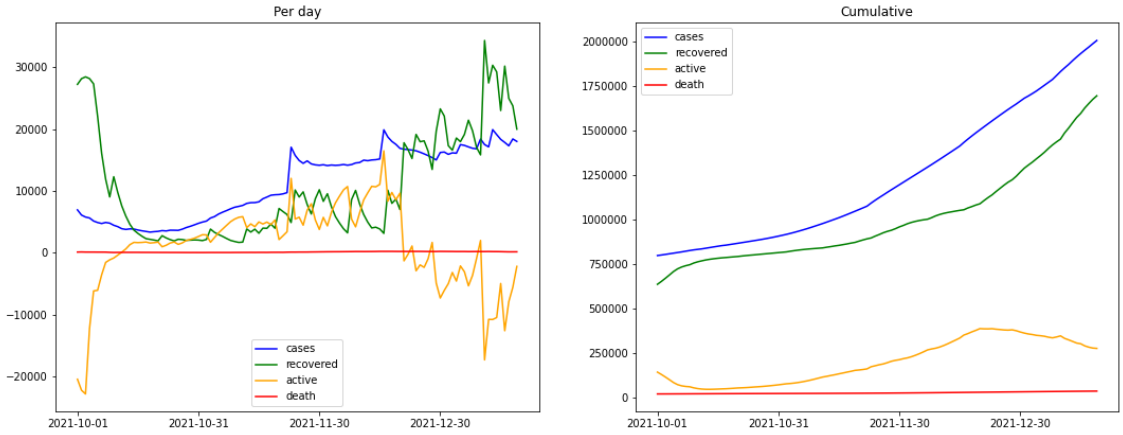

Fig. 1. COVID-19 situation in Vietnam, from Oct 1, 2021 to Jan 18, 2022

B. Data Description

The data were collected from the public APIs of the government official COVID-19 statistics website [1]. It describes the epidemic situation in Vietnam from October 2021 to the present. Figure 1 shows the trends of the COVID-19 pandemic according to the data collected.

In this paper, we only consider the part of data which represents the epidemic situation since October 2021 because the the fourth wave of epidemic broke out strongly from July and had a much more serious impact than the previous three outbreaks in Vietnam. Therefore, we have to remove records from before July so as not to affect the model. We also remove the data from July to the end of September 2021, because since the beginning of October, the epidemic situation has entered a new phase in which people have to open up and live with the epidemic. The number of active cases since October tends to increase again. Therefore, we need to remove the previous data to reduce the influence of the previous period on the future prediction.

III. Method

A. Basic Concept

In our project, we use Markov models to predict COVID-19 spreading in Vietnam. First of all, we will briefly explain this model. Markov model is a stochastic method for randomly changing systems where it is assumed that future states do not depend on past states. These models show all possible states as well as the transitions, rate of transitions, and probabilities between them.

Markov models are often used to model the probabilities of different states and the rates of transitions among them. The method is generally used to model systems. Markov models can also be used to recognize patterns, make predictions, and learn the statistics of sequential data. Applications of Markov modeling include modeling languages, natural language processing (NLP), image processing, bioinformatics, speech recognition, and modeling computer hardware and software systems.

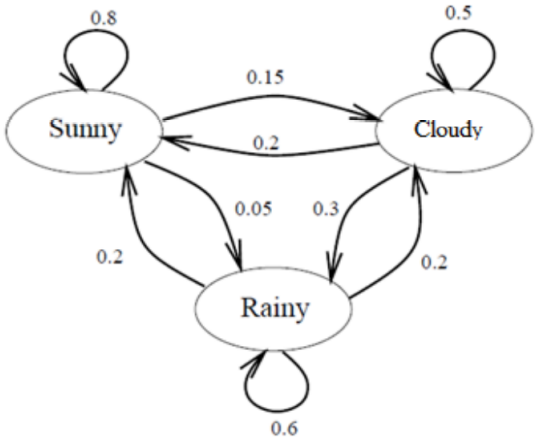

Fig. 2. Example of Markov model

B. SIRD Model

Many modeling techniques are used in cost-effectiveness but by far the most common is Markov models. They come under a myriad of names and methods originating in different fields. We will cover a classic model for infectious diseases extensively used in epidemiology. They are not called Markov models, but they are in fact Markov models—dynamic Markov models. In our project, we use an extended version of the SIR model—a dynamic Markov Model for Infectious Diseases [2], that is the SIRD model, which stands for Susceptible-Infected-Recovered-Death.

The SIRD model is called “dynamic” because people move from one state to the other at different rates over time. In other words, the transition probabilities are not fixed, they change over time. The transition probabilities are themselves

a function of other parameters in the model. The rates at which people go from one state to the other are derivatives, the model is actually a system of differential equations. In stochastic dynamic Markov models, the parameters are also given probability distribution. We will use simulations to understand the key insights of the model.

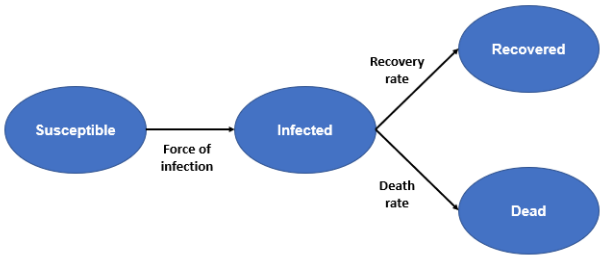

Fig. 3. A transition diagram of SIRD model

- The force of infection depends on the proportion of the population who is infected and the transmission rate. We will denote it by λ (lambda).
- The recovery rate is the rate at which people recover from the infection. We will denote it by γ (gamma)
- The death rate is the rate at which people die because of the infection. We will denote it by μ (mu)

The force of infection (the equivalent of a transition probability) is not a constant. It is defined as $\lambda(I) = \frac{\beta I}{N}$, where β is the transmission rate, I is the number of infected, and N is the total population. This is a realistic way of modeling the chances a person is infected. The larger the proportion infected $\binom{I}{N}$ the more likely it is people can become infected. β, the transmission rate, is a function of the rate of contact and probability of transmission given contact. In this simple model, we capture all these features with the parameter β, but we could define $\beta = rc * p(T)$

C. SIRD Equation

The math of the model can be complicated because of derivatives, but derivatives are just rates of change. Over time the rate of change in the number of susceptible people is going to go down as a proportion of the force of infection (negative slope):

$$\frac{dS}{dt} = -\lambda(I)S = -\beta\frac{I}{N}S \tag{1}$$

$\frac{dS}{dt}$ is the derivative of S with respect to time t. It's a change in the number of susceptible over a unit of time, where t could be days, weeks, or months (technically, continuous).

The change in the number of infected (I) per unit of time depends on the susceptible becoming infected and the infected recovering. Our approach now is modeling the transition probabilities. We assumed they were fixed numbers given by the transition matrix. So the change in the infected people over time is:

$$\frac{dI}{dt} = \beta\frac{I}{N}S - \gamma I - \mu I \tag{2}$$

The first term is the same as the above equation, it is susceptible people becoming infected as a proportion of the proportion infected. The second term, γI, are the people recovering after being infected. That's the most important equation.

Moreover, the recovery rate and death rate has the equation as below:

$$\frac{dR}{dt} = \gamma I; \frac{dD}{dt} = \mu I \tag{3}$$

One more thing, at any point the total number of people (N) must be made up of:

$$N = S + I + D + R \tag{4}$$

To summarize the SIRD model, we can see that there is a system of differential equations, after we excluded the fifth constraint:

1. $\frac{dS}{dt} = -\lambda(I)S = -\beta\frac{I}{N}S$
2. $\frac{dI}{dt} = \beta\frac{I}{N}S - \gamma I - \mu I$
3. $\frac{dR}{dt} = \gamma I$
4. $\frac{dD}{dt} = \mu I$
5. $N = S + I + D + R$

D. Solving the SIRD Model

The features of the model such as how quickly and how many people will become infected, how long will they be infected and how can we changed the situation, ... can be inferred from the system of differential equations if we studied it analytically.

A key concept in the SIRD model is the equilibria, which is a state of equilibrium where S, I, R and D stop changing overtime. It essentially equivalent to all the differential equations are equal to zero.

These stability scenario could mean that there are no more infections (disease-free equilibrium) or the disease becomes endemic (infections steadily persist; endemic equilibrium). In our model, we do not factor in the vaccines, so we will assume there are no vaccine developed yet. Therefore the only possible equilibrium is no more infections, because at some point everybody will be infected ("disease-free" equilibrium). We can control the outcome of the infection by making β zero, but that means changing the parameter, which is not an outcome of the model, but an input.

Because of the nature of the dataset we have collected, we will have to redesign the model to work with discrete time. To achieve the goal, there is only one thing to do, which is to replace derivatives in the differential equations to discrete changes. For example $\frac{dS}{dt}$ will become $\frac{\triangle S}{\triangle t}$, the changes in S from one period to the other is just $S_{t+1} - S_t$. To make things simple, time will increase by 1 unit, therefore $\frac{dS}{dt}$ will become

$\frac{S_{t+1}-S_t}{1}$, where $t+1$ is the next period of the current period t.

After the modification, we can rewrite our system of differential equations as follow:

1. $S_{t+1} - S_t = -\beta S_t$
2. $I_{t+1} - I_t = \beta S_t - \gamma I_t - \mu I$
3. $R_{t+1} - R_t = \gamma I_t$
4. $D_{t+1} - D_t = \mu I_t$

IV. Experimentation

A. Description

In our experiment, we will apply the model and simulate the future situation of COVID-19 in Vietnam. The simulation is carried out using the Python programming language on computer. In general, we've taken the following steps to carry out the experiment:

- Obtaining data
- Cleaning data
- Computing necessary parameters (β, γ and μ)
- Establishing appropriate data structures and carry out the simulation

1) Obtaining data: As mentioned in the previous section, we obtained the data using the public API from the official government website [1].

2) Cleaning data: Besides the truncation we presented, we also filtered out to include only the requires fields, which are the count of cases in each day and cumulative number of cases in each day of the following categories: active (currently being infectious), recovered, death and confirmed cases. Then, the time-series data is smoothed using the moving average technique with window size of 7 days.

3) Computing necessary parameters: Since the parameters β, γ and μ are unknown, the best we can do is to approximate them using available data. Let β_t, γ_t, μ_t be the force of infection, recovery rate, and death rate in day t, respectively. Let s_t, c_t, i_t, r_t, d_t be the number of susceptible individuals, new confirmed cases, active cases, new recovered cases, and new death cases in day t, respectively. Then, for each day t we have the following formulae:

- $\beta_t = c_t / s_t$
- $\gamma_t = r_t / i_t$
- $\mu_t = d_t / i_t$

The number of susceptible individuals s_t is not included in the data, but can be computed by the following formula:

$$s_t = N - C_t$$

where C_t is the cumulative number of confirmed cases in day t and N is the population of Vietnam. We assume N is fixed and is equal to 98,678,083.

Then, the parameters β, γ and μ are approximated by taking the average of β_t, γ_t and μ_t, respectively. In our experiment, the parameters are derived to be $\beta = 1.145\text{e}{-3}$, $\gamma = 0.0637$, $\mu = 8.969\text{e}{-4}$.

Category	Predicted	Absolute changes	Average change per day
Confirmed cases	2,338,762	$+330,480$	$+11,054$
Active cases	185,082	$-89,890$	$-3,017$
Recovered cases	2,112,377	$+413,749$	$+13,875$
Death cases	41,303	$+5,821$	$+195$

TABLE I
TOTAL CASES, ABSOLUTE AND AVERAGE CHANGES PREDICTED FROM JAN 18 TO FEB 17.

4) Carrying out the simulation: We define a state in day t to be a tuple of (s_t, i_t, r_t, d_t). The state in day $t+1$ is computed by using the SIRD model formula with the computed parameters β, γ and μ. We choose the starting day to be on January 18 and we repeatedly applying the model to predict for 30 days until February 17. The resulted sequence of states are then transformed (i.e. recomputing C_t from s_t) and combined with the original data.

B. Results

Figure 4 shows the predicted pandemic situation in Vietnam for the next 30 days. Figure 5 provides a better view of the predicted cumulative number of death cases. We can see that the model predicts trends for the categories with the rates of change are fairly similar to that of the available data.

Table I shows the 30-day absolute and average changes in each category predicted by our model, together with the total numbers. According to the result, by mid-February, the number of cases in Vietnam will increase by about 330,000 cases, that is approximately 11,000 cases a day, reaching 2.3 million cases in total. Nearly 14,000 people are relieved of the disease per day, which contributes to the declination of the number of individuals that need treatment (i.e. the active cases). This is a good sign since less stress will be placed on the public health system. By that time, it is predicted that there will be about 180,000 cases still in treatment. Death rate will still be high, with nearly 6,000 cases in one month, 195 cases per day.

C. Analysis

Our SIRD model is an extension to the classical SIR model, thus it inherits most of the characteristics of the SIR model: it is simple, deterministic and predictive. There is one crucial drawback in our model when we use it to predict the pandemic situation, however, is that we have left out vaccination. Leaving out this state means that the model assumes the pandemic will progress until all individuals in the population are infected, and then either become immune to the disease or become dead. Figure 6 shows the model prediction when we run the simulate for a wider range of time (365 days). We can see that the cumulative number of confirmed cases and recovered cases grow without bound at stable rate. This is because of the equilibrium of the model. The states S, I, R, and D will match towards the equilibrium where no state produces any more change. Since vaccination is not included in the model, R is not limited. The current percentage of vaccination in Vietnam is relatively high (about 80% of the population has been administered at least one dose), R

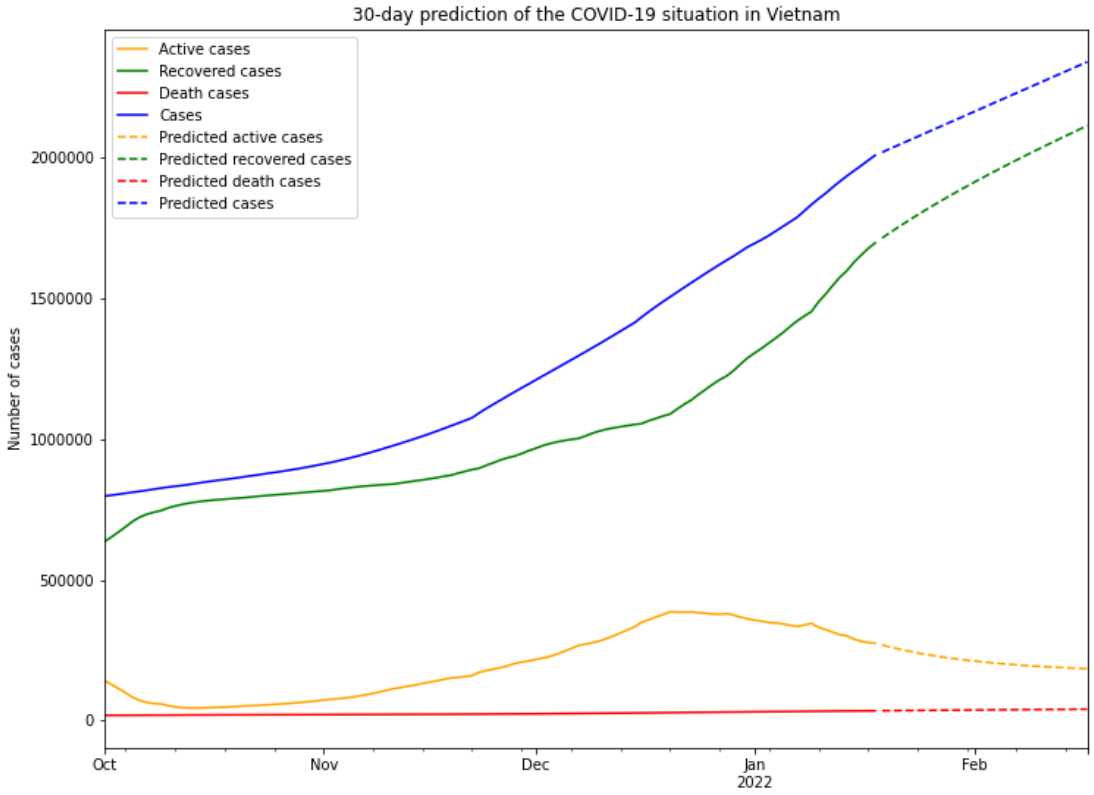

Fig. 4. 30-day prediction of COVID-19 situation in Vietnam from Jan 18 to Feb 17. The number of confirmed cases and recovered cases are expected to increase with normal rate. The number of active cases keeps declining as when it started in late December. The graph for predicted death cases is shown in Figure 5 for a better view.

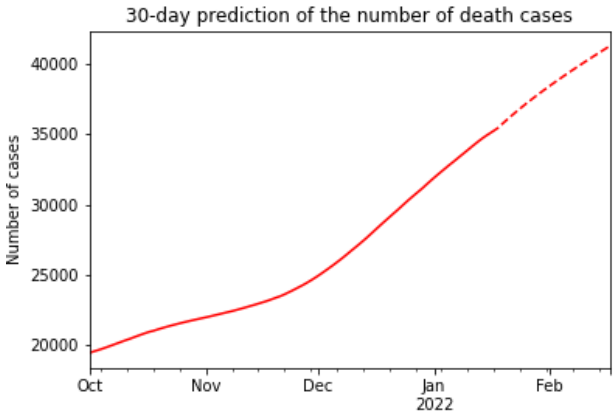

Fig. 5. 30-day prediction of the number of death cases. Notice that the predicted rate of change is relatively the same as the previous two months.

will be tightly limited if vaccination was included. Thus, this prediction fails in the long term.

Nevertheless, in short-term prediction, the model still provides a reasonable result with trends that are approximations of recent data. The result might be beneficial and serve as a basis for the government to reference and produce or adjust policies.

V. Conclusion

Epidemic can spread to a large number of people extremely fast, often exponentially. Worse with a new, highly contagious virus, everybody in the population is susceptible. Flattening the curve—a term that is very widely used—is extremely important in fighting the pandemic because if people become more and more infected at the same time we are running out of hospital capacity to treat them, mortality will be skyrocketed. The curve can be flattened by changing the contact rate, the probability of transmission given contact and improving the recovery rate. Those can be achieved by starting lockdown, wearing mask and making improvement on medications. Flattening the curve does not mean that the disease is over. With no cure, we are extending the epidemic. Flattening the curve implies a longer outbreak, but with a significantly less impact.

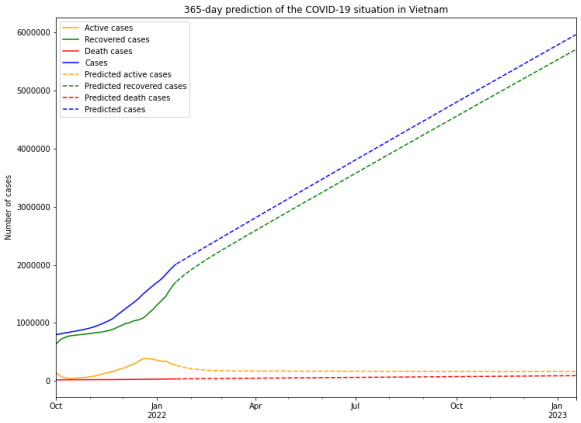

Fig. 6. Prediction of COVID-19 situation in Vietnam in one year. Notice how the number of cases grows without bound and nearly surpasses 60 million cases. Since the vaccination rate is currently over 80%, this prediction is unlikely to happen.

In this paper, we presented the SIRD model—an extension to the classical SIR model—and we carry out the simulation on the COVID-19 data in Vietnam to predict the pandemic situation in the near future. Although the model cannot produce precise prediction in the long term due to the lack of consideration for vaccination, it generalizes well for near-future forecast, using approximations of recent trends to predict the situation in the upcoming few days. With the advantage of being simple and easy to evaluate, the model can be a useful tool for government to make policy adjustment for a short period of time.

REFERENCES

[1] National Cyber Security Center. COVID-19 Data. https://covid19.ncsc.gov.vn (accessed January 12, 2022)

[2] Marcelo Coca Perraillon, “Markov Models for Infectious Diseases”, Univerisity of Colorado, Anschutz Medical Campus

[3] Abdelghafour Marfak, Doha Achaka, and Asmaa Azizi, “The hidden Markov chain modelling of the COVID-19 spreading using Moroccan dataset”, October 2020

[4] Toan Luu and Duc Huynh, “Data for understanding the risk perception of COVID-19 from Vietnamese sample”, June 2020

[5] Anh-Duc Hoang, Ngoc-Thuy Ta, Yen-Chi Nguyen, ..., “Dataset of expat teachers in Southeast Asia’s intention to leave due to the COVID-19 pandemic”, August 2020

[6] Quang D Pham, MD, Robyn M Stuart, PhD, Thuong V Nguyen, MD, ... “Estimating and mitigating the risk of COVID-19 epidemic rebound associated with reopening of international borders in Vietnam: a modelling study”, July 2021

[7] Quang Van Nguyen, Dung Anh Cao, and Son Hong Nguyen, “Spread of COVID-19 and policy responses in Vietnam: An overview”, November 2020

[8] Moutaz Alazab, Albara Awajan, Abdelwadood Mesleh, ..., “COVID-19 Prediction and Detection Using Deep Learning”, May 2020