MODELING COVID-19 TRANSMISSION USING SIR AND GENETIC ALGORITHM

Authors:

1. Amir Hesham - Sec.1 - BN: 10

2. Fady Mohsen - Sec.1 - BN: 46

3. Farah Osama - Sec.2 - BN: 3

4. Malak Nasser - Sec.2 - BN: 29

5. Hazem Raafat - Sec.1 - BN: 14

6. Muhannad Abdullah - Sec.2 - BN: 31

7. Mariam Ahmad - Sec.2 - BN: 10

8. Camellia Marwan - Sec.2 - BN: 5

MTH-2245: PDEs and Special Functions

Prof. Dr. Samah El-Tantawy

1. Abstract

In the last few years, the world has witnessed unprecedented and dramatic loss due to the horrifying pandemic, COVID-19, that invaded many countries around the globe leaving behind a tragedy that has led to numerous death and infection cases, with estimated numbers of 6.59 million deaths and 631 million cases of infection. That's why modeling the spread and genetic mutations of the Novel disease, Covid-19, has been of interest to researchers in the past few years. There are multiple models that have been made to predict as accurately as possible the spread of the disease (eg, SIR model, SEIR model,..etc) most of these models use partial differential equations and ordinary differential equations in order to describe the spread of the virus, as vast as this field of research is, we are concerned in this study mainly on using genetic algorithm as a proposed approach, for fitting the curve of the SIR model, aimed towards obtaining the parameters (β and γ) of the Kermack-McKendric SIR model(Susceptible, Infectious and Recovered), as it would be afterward applied to the model automatically, i.e using genetic algorithm as an automation tool to work more efficiently with the SIR Model and the parameters would be automatically generated. as the genetic algorithm is used to select the fittest generation by mocking the selection phenomenon that occurs in nature. By the end of this research, you'll find that our suggested approach, does in fact decrease the tedious work of finding the parameters of the SIR model specifically, or any model in general, by doing the task that used to take the day in just countable minutes, thus it's a much easier process to predict any further mutations of the virus that changes core features of its, or even new viruses.

Keywords

COVID-19, SIR model estimation, Predictive modeling Optimisation, Genetic Algorithm

2. Problem Definition

Across the generations, humans have suffered from a lot of severe Pandemics, and the most relevant in today's day and age is COVID-19, which is a part of the CORONAvirus family.

Unfolding From Wuhan, China. And spreading like wildfire all across the globe, COVID-19 was acknowledged by the WHO (World Health Organisation) as a pandemic in March 2020 as shown in fig.1. It's seen as a pandemic due to its Epizootic-like nature and its severe effects on many life sides such as; health and environments.

According to [3] COVID is still spreading steadily. As fig.1 shows Coronavirus daily new cases, the total number of cases has reached 19 million around the world, with 700,000 deaths in 213 different countries and territories[3].

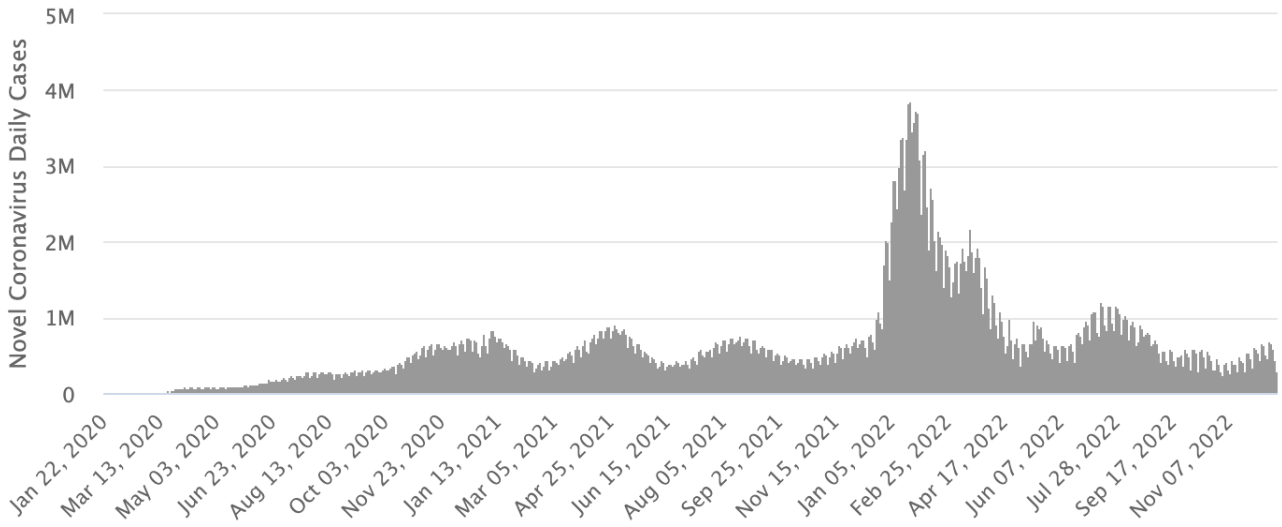


Figure 1: shows Coronavirus Daily new case [1].

It's known that the disease is passed on through respiratory droplets of the infected person, along with Flormite-mediated and nosocomially acquired infections, which is a key source of viral diffusion. However, [4] believes that the detailed mechanisms of the disease are yet to be known.

So it has become greatly significant to simulate and model the virus infection development process to accurately predict the effect of the virus spread over time and how it will grow and mutate, and whether it will be more fierce or mild and this will give us a better view of the way we should handle the virus in the future and mitigate the impact of spread.

This can be reached only by an efficient model describing the behavior of the disease over time with some parameters that affect it. So, Scientists have been using the SIR model for modeling the infectious of COVID-19, in which they model the total population classified into three categories S, I, and R:

Susceptible who are free of the disease but are susceptible to the infection by the Infectious, Infectious are the ones tested positive for the disease, and Recovered people.

The SIR model can provide us with insights and predictions of the spread of the virus in communities that the recorded data alone cannot. And it's represented by a few ODEs, including 2 main parameters which are the infectious rate (β) and the recovery rate (γ) but depending on manual intervention to calibrate the parameters of the model in order to fit the curve of the real-world data is such a hindrance, and a huge problem that we managed to solve.

So here, our solution comes to finding the most suitable assisting optimization tool; the GA (Genetic Algorithm) which is the most widespread and typical example of Evolutionary Algorithms (EAs) and it deals with highly nonlinear optimization issues that need global optimal solutions. In our model, the genetic algorithms play a significant role since they will be utilized to perform curve fitting on the real-world data to determine the values of (β) and (γ), allowing us to derive the SIR model's parameters.

Our proposed approach reduces the manual involvement which was required in tuning the SIR model. Thus, we can easily narrow down the gap between the simulation and the real-world data.

This Proposed solution will only consider COVID-19 interactions with individuals in a confined space, which has assumed conditions that will be discussed later. There is, however, a belief among the authors that modeling COVID-19 in that way will open the door for further studies about viruses in general in the future [11].

3. Literature Review

One of the newly fresh solutions that enable us to keep tabs on the technological world, is to use genetic algorithms as an assisting and optimization tool. The GA(Genetic Algorithms) is the most widespread and typical example of Evolutionary Algorithms (EAs). It was first proposed by John Holland in 1975. The GA adopts Darwin's theory of natural selection and evolution. Since then, it has been widely applied in different disciplines [5].

The approach of genetic algorithms (GAs), biologically inspired optimization algorithms, comes from the evolutionary process. When dealing with highly nonlinear optimization issues that need global optimal solutions, GA is frequently used as a technique. Parents and children are always included in GA as a subpopulation. Every member of a population is assumed to have a specific fitness. Only people who are highly fit survive each generation (Survival of the fittest) [11].

The GA is normally created by selecting a group of potential solutions known as chromosomes. These chromosomes form a population. The genes that make up each chromosome are subdivided into smaller parts.

The number of genes in a chromosome, or its length, affects how dimensional a problem is. To prepare a new population in a fashion that mimics natural evolution, a series of evolutionary operators (selection, crossover, and mutation) are performed at each iteration of the GA to increase variety in the present population. Each chromosome is assessed using a set of evaluation criteria (object functions) to establish its quality and suitability. In each iteration, the best person (highest-rated solution) is kept.

The worst candidates (unfit solutions) are candidates to be replaced by the newly generated offspring. This allows the average fitness value to increase dramatically throughout the iterations [5], [17].

With the unprecedented spread of covid-19 in the last few recent years, several various predictive models for covid-19 have come into play, each one includes a different set of data and parameters and different ways to estimate how they work together then implement the methodology and evaluate its efficiency. M. M. Malik mentions that [9] - similar to all types of analytics - the key issue in healthcare analytics is how we can use data effectively to generate insights to improve processes and aid decision-making [9].

Current developments are suggesting two ways [8] to deal with the problem of modeling a problem with such a large number of contributing factors: Curve Fitting Based Models, and several attempts have tried to solve this problem using evolutionary mechanisms like GAs (Genetic Algorithms) which are what we are trying to achieve.

Los Alamos National Laboratory[8] used a Curve-fitting technique in their predictive model, named: Confirmed and Forecasted Cased Data Model [8]. As a fact, from their model, the best guess was for California state as of April 08, 2020, which were 4082 deaths (compared to 2974 actual deaths, dated May 14, 2020). The Institute for Health Metrics and Evaluation (IHME) [7] – an independent global health center – used a curve-fitting model to project numbers of hospitalizations and deaths in the U.S. (including state-wise data) through August 2020. Their predictions vary over time as they employ a curve-fitting model [7].

Undoubtedly genetic algorithms are the most widespread and typical methodology of electronic arts (EAs). This paper [5] has used the genetic algorithm to model by identifying a set of candidate solutions, in their application called chromosomes. These chromosomes represent a whole big population. Each chromosome contains 2 small units called genes. The dimensionality of the problem is determined by the length of the chromosomes or the number of genes in the chromosomes. Some number of iterations happens depending on the dimensionality. Each iteration in the genetic algorithm shows diversity in the current population by some main operators including selection, crossover, and mutation. This diversity is used to introduce other new populations so that it simulates natural evolution. The purpose of these iterations is to identify whether the chromosome is fit or unfit based on some criteria. The fittest solution is then preserved for the next iteration. The unfit solutions are then replaced by other newly generated offspring. These analytics provide a dramatic increase in the fitness of the generations.

A study Y. & Sun, H. S., P. (2020, June 18) was conducted in 1995 proposed by Kennedy and Eberhart came up with the idea that a group of birds that move together can benefit from all other members because as a bird is flying and looking for food, all the other birds can share their experience and aid the rest of the flock get the best hunt. This method is called Particle Swarm Optimization (PSO). It's used as an optimization algorithm to evaluate the different parameters of the system [12].

The classical SEIR model has four essential elements which are S (susceptible), E(exposed), I (infectious), and R (recovered). Thus, N=S+E+I+R means the total number of the population [12].

The PSO algorithm provides a heuristic way for estimation and calculation of the parameters of the SIR model. It helps us in finding the optimal solution in a multi-dimensional solution space that is pretty close to the real global optimal solution. And according to the application of the data of Hubei province, the accuracy is acceptable. [12].

A genetic algorithm (GA) is an optimization strategy that draws inspiration from Darwinian evolution. In a genetic algorithm, an initial population of candidate solutions, each represented by a sequence of values called a genome, undergoes random mutations as they breed and reproduce. The fundamental working principle of a GA is that, through Darwinian natural selection, only the top-performing solutions are allowed to procreate and pass on their genes. After a predetermined number of generations, the evolution process is finally terminated when a predetermined stopping condition is confirmed. [18].

To simulate COVID-19 diffusion we will be using a SEIQRDP model in which at a specific time t the population is split between different compartments.The susceptible part of the population, individuals yet to be infected, is represented by compartment S(t). while P(t) represents the protected population. E(t) covers people that have been exposed to the virus. I(t) show individuals that are currently infectious. Q(t) represents quarantined individuals. R(t) represents the part of the people that have recovered from covid and are not involved in the virus spread anymore
N=S(t)+E(t)+I(t)+Q(t)+R(t)+D(t)+P(t) is the total population at time t[18],[15].

In our case, applying a GA to find the best fitting approach for the SEIQRDP model parameters.The best-performing set of parameters is the ones that produce and give rise to the curves that match the best original real-world data, by minimizing the normalized root mean squared error. [18].

The COVID-19 pandemic presents unprecedented and horrific challenges to the world health system in this day and age. Researchers reported a lot of COVID-19 vaccines introduced by various institutions and different companies around the globe. However, research developing an integrated framework for choosing and ranking the optimal potential vaccine against COVID-19 is minimal. So we are here to fill

the gap by using a hybrid methodology based on ELimination Et Choice Translating REality III (ELECTRE III)–Genetic Algorithm (GA) and Technique of Order Preference Similarity to the Ideal Solution (TOPSIS) approach to select the optimal SARS‑CoV‑2 vaccine. [6].

Identifying and picking out the appropriate criteria for COVID-19 candidate vaccines is essential and significant for reliable and well-founded outcomes. We examined several criteria sets proposed by the World Health Organization (WHO), Centers for Disease Control and Prevention **(CDC)**, and U.S. Food and Drug Administration (**FDA)**. Those indicators are safety(**SAF**): implying that data collected from the animal and human studies support no risk of enhanced disease in vaccines, efficacy(VEF): which refers to the percentage reduction in the occurrence of the disease in a vaccinated group compared to an unvaccinated group under normal conditions, effectiveness**(EFF**): monitors the benefits of vaccination for the community, and stability(STA): refers to maintaining optimal storage conditions [6].

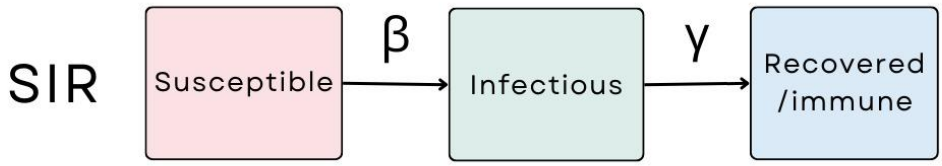
So After performing a detailed state-of-the-art and taking into consideration expert opinions, an evaluation model is developed, and we are able to evaluate and assess the available vaccine alternatives and select the most convenient one according to many factors including subjective and qualitative judgments and many different complex factors. So in the end, MCDM methods can be effectively employed to pick the most suitable COVID-19 vaccine appropriately and accurately [6]

Publication	**Model**	**Curve fitting approach**
Shaobo He, et al. [12]	SEIR Model	Particle Swarm Optimization (PSO).
M. A. Bahloul, et al. [18]	SEIQRDP Model	Genetic Algorithm (GA).
Forestal RL, et al. [6]	Hybrid approach, TOPSIS	Multiple-criteria decision making (MCDM).

Table (1): shows a comparison between the different mentioned studies.

4. Mathematical Modeling

For modeling the infections of COVID-19 disease, we need to use the Kermack-McKendrick SIR model as shown in Fig. 2 in which we model the total population. Kermack-McKendrick studied and developed a model to understand how the progress of the epidemic is by narrowing down the scope of the population or by making some assumptions. The assumptions are fully illustrated in the upcoming section. The total population is classified into three categories **S**, **I** and **R**: where **Susceptible (S)**, are healthy individuals that are free of the disease but are susceptible to the infection by the **Infectious (I)**, Infectious are the ones who tested positive for the disease, and then for simplicity after a period of time considered **Recovered (R)**, as it's assumed that they gained permanent immunity [13], [14].



Figure(2) shows The Relation Between SIR Elements.[2]

$$S(t) + I(t) + R(t) = N : N \text{ total population number} \qquad Eq.1$$

The assumptions made are that the Population **(N)** is a homogeneous mix of the infected I and susceptible S, where mere contact causes the spread of the disease, for simplicity, the population is placed in a relatively small confined space, there are no immigrants or births in the population. Also assuming that once a person is Recovered R, they are no longer in the scope of the population being studied because they won't get reinfected (or are dead). The infectious people are assumed to be spreading the disease for a constant time t before they are removed from the pool and considered Recovered (i.e there's a fixed infection rate per day and fixed recovery time). In this particular scenario, the number of susceptible people would decay monotonically over time to zero.

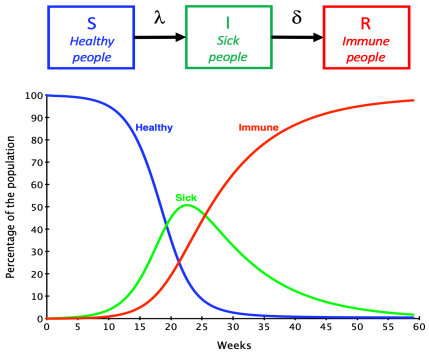


Figure 3: The Relation between weeks and percentage of population in SIR Model

4.1 Modeling the Susceptible group S

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI \qquad Eq.2$$

where β is a positive constant that describes the infectious rate in the homogenous confined space. The negative sign indicates that the susceptible group is decreasing over time as they ultimately turn into infectious people. Multiplying S and I models the continuous interactions between susceptible people and infected people [13].

4.2 Modeling the Infected group I

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta SI - \gamma I \qquad Eq.3$$

where γ is a positive constant that describes the recovery rate of the system. The idea is that there are a number of people interacting with the infected ones and in turn, they become infected themselves as well, however, there is a portion of the infected people who are also recovering and joining the recovered group [13].

4.3 Modeling the Recovered group R

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I \qquad Eq.4$$

Only using the γ constant with a positive sign makes sense because the recovered group are the ones extracted from the Infected group and they can't get infected once again.

After modeling the system as three categories, now it's needed to find a near-beginning value that represents the early growth rate.

Focusing on near beginning conditions for the Infected group we find the following in Eq. 5:

$$\left.\frac{\mathrm{d}I}{\mathrm{d}t}\right|_{t=0} = \beta S_0 I_0 - \gamma I_0 \qquad Eq.5$$

If the above quantity is observed to be less than or equal to zero, this means that the infection is declining, otherwise, it means that the infection is still growing in an epidemic way.

As we discussed, we are interested in how this value behaves near the beginning:

$$\beta S_0 - \gamma <? 0 \qquad Eq.6 \qquad \text{then,} \qquad \frac{\beta S_0}{\gamma} <? 1 \qquad Eq.7$$

This is the quantity we are searching for. It's an indication of how the infection is growing at the near-beginning conditions. We will call this value R_0 .

$$R_0 = \frac{\beta S_0}{\gamma} \qquad Eq.8$$

The genetic algorithm (GA) plays a great role in this model. GAs are mainly concerned with optimization and for that reason, they will be used in this model to extract the rate of infection and the rate of recovery from any given curve or real data.

The SIR model derived above is relatively simple, it consists of two first-order ordinary differential equations. This is a result of the fact that it describes the evolution of the disease within a small confined and limited scale: it does not take into account the dynamics of the individuals at the microscopic scale. The model derived is macroscale.

Because of that, it is often thought that the model is accurate only to describe the evolution of a disease in a population located in one "small" site, where contacts between individuals are extremely frequent, such as in a very dense city for instance [19], [11],[13].

If one wants to describe the propagation of a disease into a region, a country, or the whole world, the setting is different, and it is then natural to take into account in the model spatial, microscopic effects: movement of individuals (migration, diffusion), the spatial distribution of the population.

Starting from the SIR model, we can build a simple microscale model for the spatial spread of an epidemic by adding diffusion terms. These terms will reflect the spatial distribution of individuals at the microscopic scale. Doing so, we're led to study the following system of PDEs:

$$\begin{cases} \partial_t S(t,x) = d_s \Delta S(t,x) - \alpha SI, & t>0,\ x \in \Omega \\ \partial_t I(t,x) = d_I \Delta I(t,x) + \alpha SI - \mu I, & t>0,\ x \in \Omega \\ \partial_v S(t,x) = d_v I(t,x) = 0, & t>0,\ x \in \partial\Omega. \end{cases} \qquad Eq.9$$

In this system, the domain $\Omega \subset$ RN is an open connected bounded set of class C2, and ν is the unit outward normal vector field to Ω.

The functions S(t, x), and I(t, x) represent the densities of susceptible and infectious individuals respectively, at time t > 0 and at position x $\in$ Ω. The individuals move randomly following a Brownian

motion on the domain, which is reflected in the equations by the presence of Laplace operators. The quantities dS and dI > 0 are the diffusivities of the susceptible and infectious individuals respectively. They represent the amplitude of the Brownian motions of the individuals.

The Neuman boundary condition accounts for the fact that the individuals that reach the boundary bounce back into the domain following a Descartes reflection law. Finally, one has to complete this system with an initial condition **(S0(·), I0(·))** representing the initial spatial distribution of susceptible and infectious individuals. Observe that, if S0, I0 is constant over Ω, then the new PDEs system boils down to the previously derived ODEs [19], [11].

4.4 Curve-fitting using Genetic Algorithms

After having an overview of the SIR model and how it represents the total population, we've found that by using the genetic algorithm we can better predict the SIR graphs to match the real-world data.

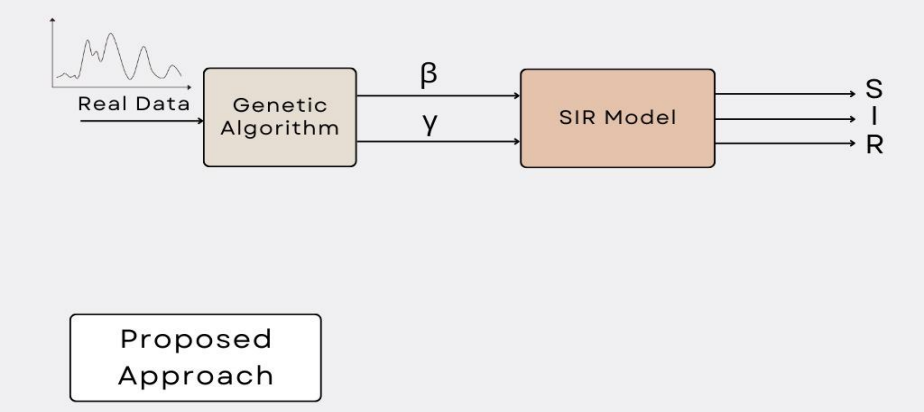


Figure (4): Out proposed approach of SIR Model

The simple input/output diagram displays the proposed approach of the SIR model as shown in Fig.4.

The parameters obtained by a fitting algorithm are used to construct the variable curves that fit the initial data.

Those curves are then extrapolated to a longer period, thus forecasting the evolution of the epidemic.

And to calibrate our model's parameters and fit the data originating from a specific region, we have chosen the evolutionary genetic algorithm (GA) for solving our optimization problem.

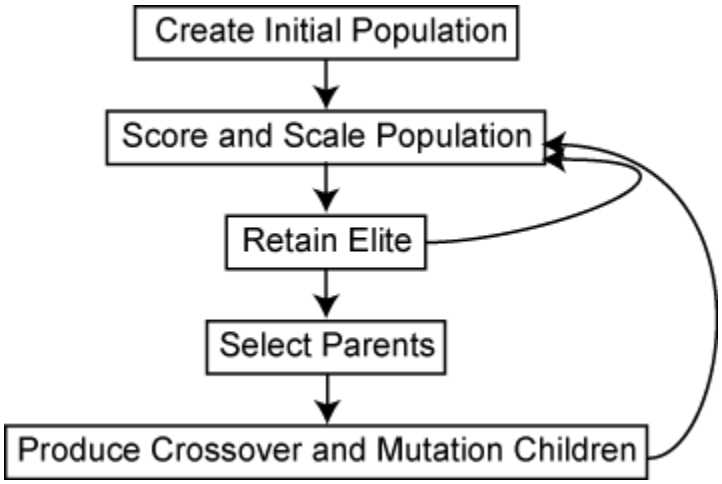


Figure (5): Genetic Algorithm Mechanism

A genetic algorithm (GA) is an optimization approach that is inspired by the process of natural selection. Natural selection is the concept that the fittest individuals reproduce and create offspring who inherit traits from their parents. The initial set of candidate solutions called the **initial population**, each represented by a sequence of values forming a genome, evolves by breeding and reproducing while being subjected to random **mutation**. The key mechanism in a GA is that only the **best-performing** solutions get to survive and reproduce and pass on their genes. The evolution process is finally stopped after several generations when a defined stopping condition is verified.

5. Methods

5.1 The SIR Model

To briefly wrap up, the SIR Model is our primary tool for modeling the susceptible, infected, and Recovered subjects in a constant population. Mathematical models, in general, need mathematical optimization, in order to find the best solution possible for problems, it's also considered a valuable tool to estimate the quality of a model, based on how well its solution agrees with real-world data, and some crucial things that need to be set carefully are the parameters of your model, as they are used to calculate the output, so, they need to be as accurate as possible. In our case, the SIR model parameters are the γ and β.

In the scope of epidemic models or predictive models, differential equations that make up the model can be solved using a forward approach, if you have the pre-defined parameters, giving you observable data and measurements that describe how the model behaves.

It can be concluded that the individual takes 10 days to recover from infection according to the value of γ, we can from here proceed with the numerical solution to solve the system of quotations thus producing a solution.

However, we don't always have the luxury of knowing the γ and β values, and finding them is such a laborious process that involves many trial and error trials. And that's what inspired the usage of the genetic algorithm to solve this problem, as we will be using the genetic algorithm in order to automate the process of finding the β and γ so that if the virus mutates, the parameters can be easily obtained, or even if another virus broke out, that has completely different features than Covid-19, there will be an easier and much faster way to obtain the parameters for the predictive model.

5.2 Genetic Algorithm

In this section, we will be giving insight into how we will be using the genetic algorithm as an automation tool to easily obtain the parameters of the model, specifically here the SIR model.

Step 1: the initialization of the population and its size, (eg. population X has 4 chromosomes, where an individual is considered a chromosome), our goal is to obtain the β and γ from the following equations of the SIR model.

$$\frac{dS}{dt} = -\beta SI$$

$$N\frac{dI}{dt} = \beta SI\,N - \gamma I$$

$$\frac{dR}{dt} = \gamma I.$$

The genetic algorithm will generate then five possible sets, within the feasible region.

Step 2: determining the fitness of curves, The fitness or the objective is a way to describe how accurate the output curves of the genetic algorithm are from the real data, a smaller fitness distance indicates that the predicted curve is closer to the real data. Thus, the individuals who produced smaller fitness function values, have higher fitness scores, and higher possibilities to be chosen for the next generation.

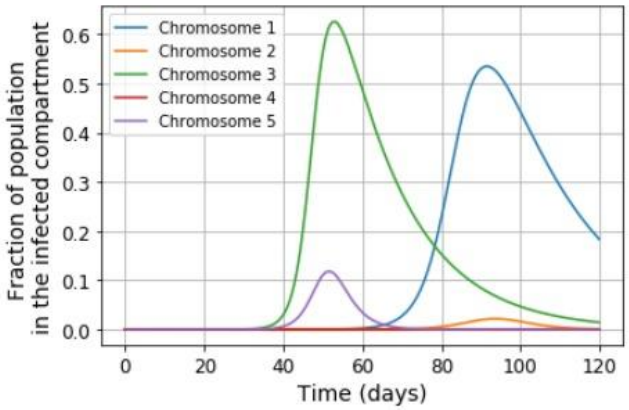
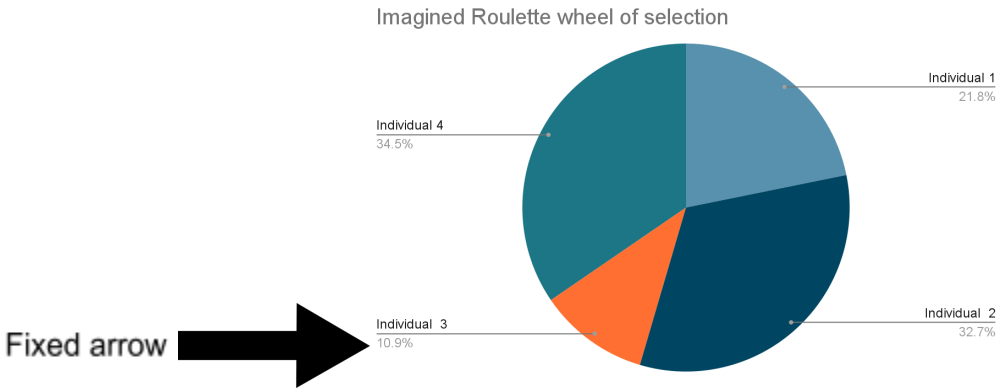


Figure (6): The number of individuals in the infected compartment for each t ∈ [0, 120] for each chromosome [2].

Step 3: The selection process it's deciding the fittest individuals for the next generation, however, there's a random part in the selection process, i.e. it's not as simple as choosing the highest fitness scores, as the genetic algorithm puts into consideration the possibility of random selection, just like what happens in nature.
The selection process was thought of as a Roulette wheel, then the fittest generations will take larger sections of the wheel so that they are more likely to be selected, yet it's still possible that individuals with lesser fitness scores would be chosen
It can be illustrated in the following figure.



Figure(7): Imagined Roulette Wheel of Selection

Step 4: The Crossing over and mutations. This step is the production of the offspring from the selected parents, crossing over and mutations are done to produce the offspring, each of the cross-over and the mutations has their possibilities so that the offspring will either be a product of crossover or mutations.

These phases will be repeated until we reach a generation that produces an output, as accurate to the real data we have as possible, and we derive the β and γ from them. As it will be more clear in the experimental work, the Genetic algorithm takes very little time to produce the β and γ that we can spend hours and even days to derive for each case of virus mutation, as their features change and for other viruses.

7. Experimental Work

7.1 Describing the data

Table 2 provides The data we are going to use as official data on the COVID-19 epidemic in Malaysia. Powered by CPRC, CPRC Hospital System, MKAK, and MySejahtera. The data includes the number of new cases, active cases, and recovered cases from 25 January 2020 to 21 December 2022

	cases_new	cases_import	cases_recovered	cases_active
count	1050.000000	1050.000000	1050.000000	1050.000000
mean	4769.954286	36.639048	4715.462857	55298.667619
std	6646.229856	96.153953	6595.671687	74654.422327
min	0.000000	0.000000	0.000000	1.000000
25%	634.000000	3.000000	286.250000	5444.750000
50%	2360.000000	6.000000	2359.500000	27501.500000
75%	5302.500000	15.000000	5193.750000	62801.500000
max	33406.000000	719.000000	33872.000000	323785.000000

Table (2); Shows the description of the used data
using Pandas Python Library

7.2 Exploring the data

This data shows that the number of new cases is very high relative to its population and with other countries, both mean and Standard deviation show that the number of new cases per day is high and

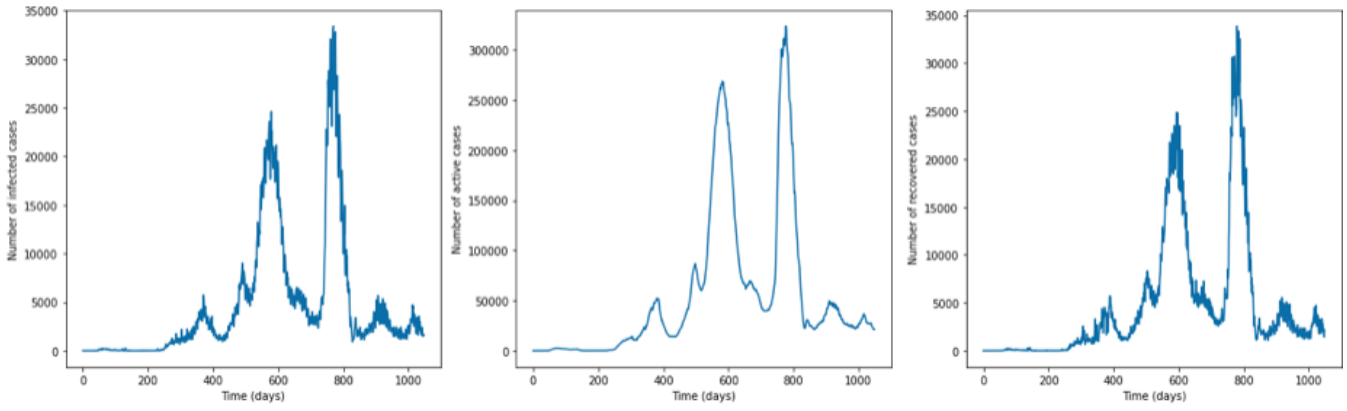


Figure (8): Shows a visualization for the number of infected cases,
active cases, and recovered cases, respectively, with time.

The recovered cases in the table show that the number of cases that become well and recovered is almost in a range of results per day.

To assess the curve-fitting capabilities of the developed genetic algorithm, The following to processes shall be considered:

- Calibration of the GA

  A predetermined pseudo SIR infection curve with pre known parameters will be used to assess the curve-fitting by letting the GA approximate that curve and compare the resulting parameters β and γ with the predetermined ones and calculating the error.

- Testing against real-world data

  the GA curve fitting against an infection curve observed from real-world data. The data of COVID-19 new cases in Malaysia will be used to generate that curve, then the GA will apply curve-fitting in order to automatically get the estimated infection rate (β) and the recovery rate (γ) in order to be able to accurately describe the infection transmission with the SIR model.

7.3 Calibration of the genetic algorithm

To test the curve-fitting ability of the genetic algorithm, we will choose random calibration values for the SIR parameters.

1. Consider an infection curve I_1 having a rate of infection of β_1 and a rate of recovery of γ_1. Figure(11) Shows an arbitrary plot that corresponds to the arbitrary parameters (on the left).

2. We generate a random population of five members. Let each member be Ω where:

$$\Omega = \{\beta,\ \gamma\} : \beta \in [0.1,\ 1],\ \gamma \in [0.05,\ 1]$$

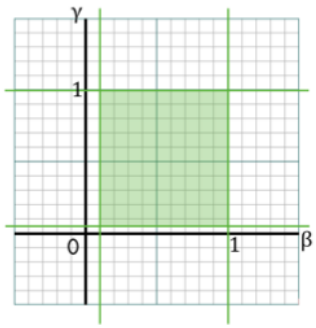


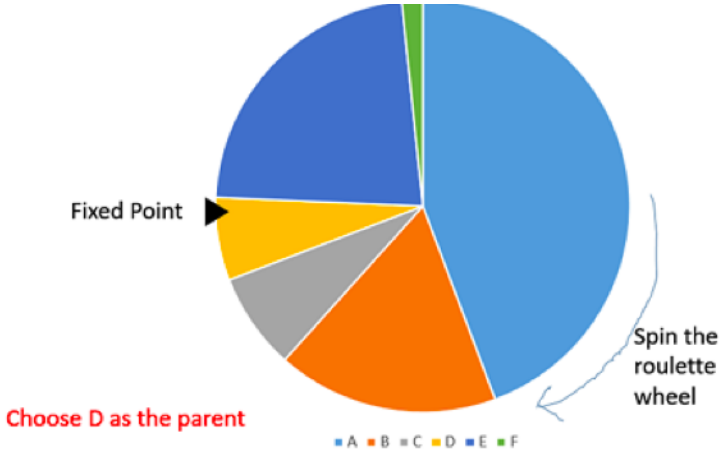
Figure (9): Shows an arbitrary plot that corresponds to the arbitrary parameters (on the left) [2].

Note that both β and γ should be within their feasible range described above.

3. The selection process of the genetic algorithm uses the following fitness function. This fitness function tries to minimize the L2 norm distance between the curve with the parameters of Ω and the calibration curve in order to achieve the curve-fitting after a number of generations.

$$\begin{cases} \text{minimize} & ||c_r - I^{\Omega}||_{L_2} \\ \text{subject to} & \Omega \in F \subset \mathbb{R}^n, \end{cases}$$

4. We are following the Roulette Wheel selection process where two parents are chosen to create the new offspring. Figure(X) shows the Roulette Wheel selection between six members.



Chromosome	Fitness Value
A	8.2
B	3.2
C	1.4
D	1.2
E	4.2
F	0.3

Figure (10) Visual representation of the Roulette Wheel selection with six individuals in a population [2].

5. After the selection process, we are using the traditional crossover and mutation processes in order to create the new generation. Then, the cycle is repeated until the termination condition is met which is the number of generations given.

6. Figure(11) shows the plotting of both the calibration curve (on the left) and the final curve generated by the genetic algorithm (on the right) and it's easily observed that the developed approach is doing a good job fitting the calibration curve.

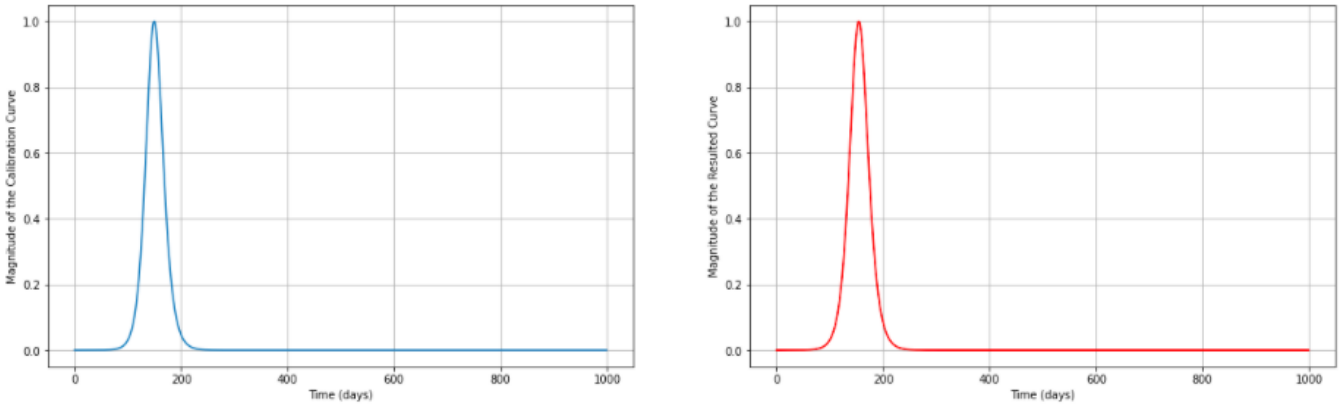


Figure (11): shows the plotting of both the calibration curve (on the left) and the final curve generated by the genetic algorithm (on the right).

7.4 Testing against the real-world data

Now that we are confident that our developed approach produces a good estimation of the parameters of a given arbitrary curve, we can use the same technique mentioned in (7.3) to extract the parameters β and γ of the real-world data described in (7.1).

- When using noisy real-world data against our genetic algorithm, it should be noted that the genetic algorithm aims for the global

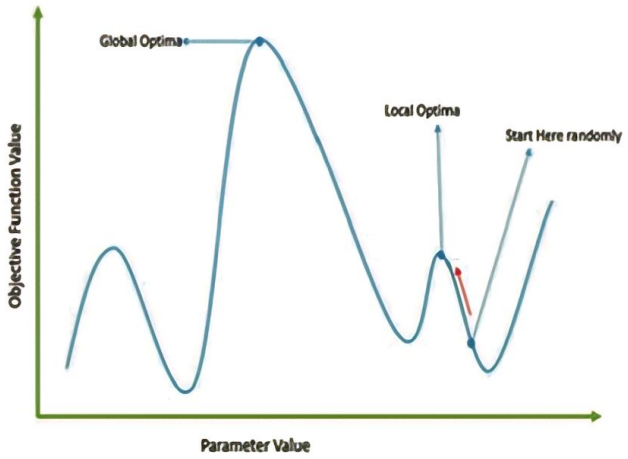


Figure 12: The relation between the parameter and the value and objective function value [2].

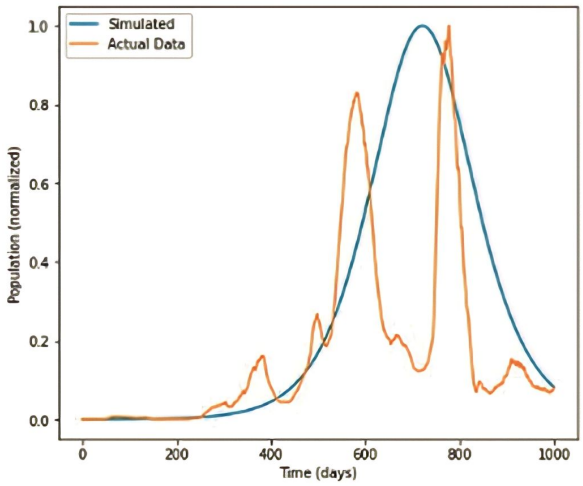


Figure 13: shows a test 100 generation run against real-world data with a time range of 1000 days.

Figure (13) shows a test 100 generation run against the real-world data with a time range of 1000 days. It demonstrates our approach's ability to find the global optima. The results are considered fast and enough.

8. Results and Discussion

In order to study the behavior of infection using the SIR model we need to extract the parameters needed by the **SIR** model (the rate of infection β and the rate of recovery γ) from the given data using a genetic algorithm.

It should be noted that due to the unpredictable probabilistic nature of a genetic algorithm, the results are not perfectly reproducible between repeated simulations. At the same time, since the number of generations simulated is large, the result can approach the same behavior statistically (large number theorem).

To determine the error of the curve fitting results using the genetic algorithm, a pseudo SIR model with predetermined parameters is created, and by letting the GA do the curve fitting process with **1000** generations the following results are observed in Table. 3 and . Figure 14:

	Test	**Simulated**	**Error (%)**
Infection Rate β	0.5	0.559	- 11.86
Recovery Rate γ	0.4	0.461	-15.14

Table (3): Calibration of the curve-fitting of GA

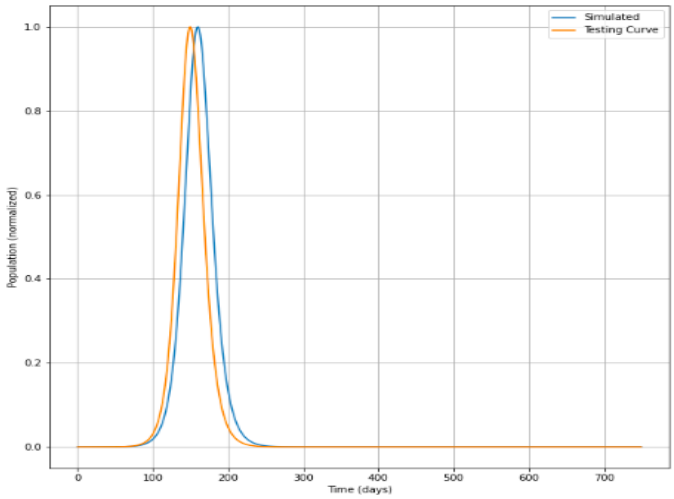


Figure (14): The Simulated curve against the test curve.

In Fig. 15 It's observed that the genetic algorithm fits the curve of infection of the real-world Malaysia dataset pretty decently.
The estimated parameters of Malaysia COVID-19 data corresponding to Fig. 15:

Infection Rate β	0.6910783845725373
Recovery Rate γ	0.673648944074867

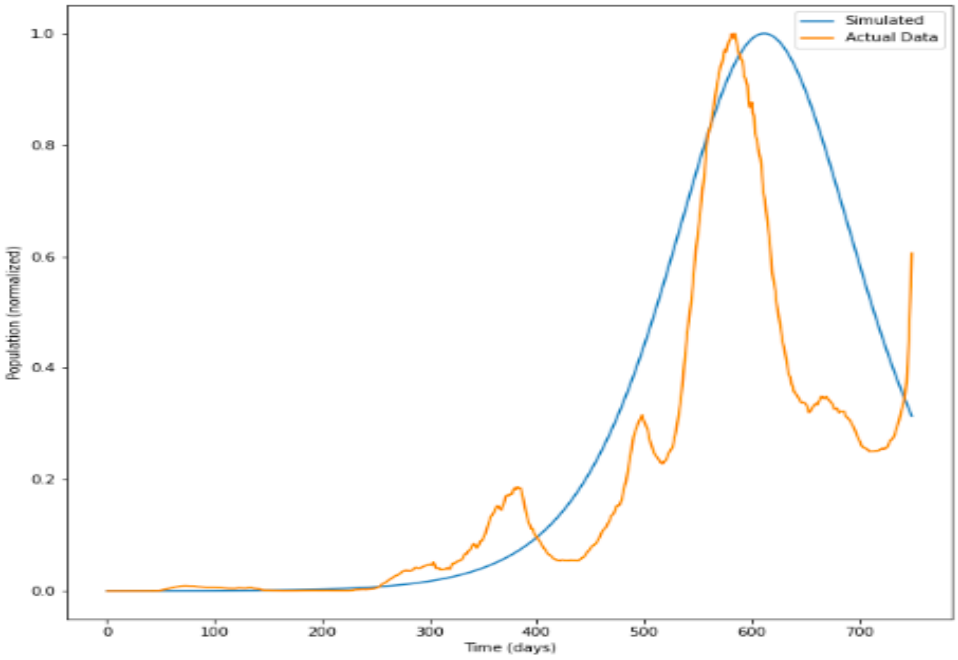


Figure (15) Fitting the curve of infection in the Malaysia dataset using the genetic algorithms

It's also worth mentioning that the results of the **curve fitting** (parameters tuning) can be enhanced by increasing the number of generations in the genetic algorithm.

Now that it's possible to easily determine/tune the parameters of a given dataset, the behavior of the infection can be easily modeled and studied using the typical **SIR** mathematical model.

Conclusion

It's been concluded by the end of this study, that our proposed method curve fitting, which is the genetic algorithm, can be successfully used as an automation and optimization tool, to be used in a model like the SIR model, it's also been found that the genetic algorithm reduces the manual intervention that is required while tuning the SIR model in order to fit the observed new cases, by using the mutated parameters of the genetic algorithm.
In a matter of seconds, genetic algorithms can easily optimize the parameters used to narrow down the gap between the simulation and the real-world observation, saving days worth of code running and comparing it with the observation.

Needless to say, there's still room for further improvement in this field of research, as the genetic algorithm's efficiency at discovering optimal parameters in model simulation leading to minimal human intervention is quite vast, plus, such an approach can be applied to any activity or case study with a large number of trial and error is needed in order to match a pattern and optimize functions heuristically. Thanks to this approach, in the future we can take precautions for any upcoming pandemics.

Future work

Apparently, genetic algorithms pulse a lot of discoveries in social sciences, open ended evolution, artificial life and AI. It can produce high-quality solutions for various problems in the future, as it is able to surmount a wide range of problems faced by traditional algorithms, because of its applicability in various research domains.

Genetic algorithms are more likely to produce global optimal solutions unlike the traditional ones, as well as it can handle the real-world problems arising from different fields which are multi-modal in a more efficient manner because of its large solution space.

So for the sake of humanity, we can lean on our proposed method furthermore in the future as an automation and optimization tool to model any incoming pandemic more efficiently with very high accuracy rates, as it can auto-tune the parameters of any model without making such a human effort, and with a very insignificant error.

10. Appendix

1. Code on Google Colab Notebook
2. Introduction Presentation

11. References

[1] *Coronavirus Cases: Statistics and Charts - Worldometer*. (n.d.). https://www.worldometers.info/coronavirus/coronavirus-cases/

[2] Spataru, D. (2021) *Using a Genetic Algorithm for Parameter Estimation in a Modified SEIR Model of COVID-19 Spread in Ontario*. Guelph. https://atrium.lib.uoguelph.ca/xmlui/bitstream/handle/10214/26379/Spataru_Daiana_202109_MSc.pdf?sequence=1

[3] Ghosh, S., & Bhattacharya, S. (2020). A data-driven understanding of COVID-19 dynamics using sequential genetic algorithm-based probabilistic cellular automata. *Applied Soft Computing, 96*. https://doi.org/10.1016/j.asoc.2020.106692.

[4] Acosta-González, E., Andrada-Félix, J., & Fernández-Rodríguez, F. (2022). *On the evolution of the COVID-19 epidemiological parameters using only the series of deceased. A study of the Spanish outbreak using Genetic Algorithms, Mathematics, and Computers in Simulation. 91-104*. https://doi.org/10.1016/j.matcom.2022.02.007.

[5] Al-Ahmad, B., M. Al-Zoubi, A., & Abu Khurma 2, R. (2021). *An Evolutionary Fake News Detection Method for COVID-19 Pandemic Information.* https://www.mdpi.com/2073-8994/13/6/1091/htm

[6] Forestal RL, Pi SM. A hybrid approach based on ELECTRE III‐genetic algorithm and TOPSIS method for selecting optimal COVID‐19 vaccines. (2021, Nov 15). *Journal of Multi‐Criteria Decision Analysis.*, 6.https://doi.org/10.1002/mcda.1772

[7] The Institute for Health Metrics and Evaluation (IHME). (2020, June 20). *The Institute for Health Metrics and Evaluation (IHME) COVID-19 Model:* COVID-19 Model. https://covid19.healthdata.org/united-states-of-america

[8] Los Alamos National Laboratory. (2020, June 20). *Confirmed and Forecasted Cased Data Model*. LANL COVID-19 Cases and Deaths Forecasts. https://covid-19.bsvgateway.org/

[9] Malik, M.M., Abdallah, S., & Ala'raj, M. (2016, December 24). *Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review*. https://link.springer.com/content/pdf/10.1007/s10479-016-2393-z.pdf

[11] Rangarajan, S., Sammakia, B., & Porank, S. (2021, April 28). Genetic Algorithm-based Predictive model for COVID-19. https://assets.researchsquare.com/files/rs-107714/v2_covered.pdf?c=1631865061

[12], Y. & Sun, H. S., P. (2020, June 18). SEIR modeling of COVID-19 and its dynamics. 15. https://doi.org/10.1007/s11071-020-05743-y

[13] El Mehdi Lotfi. 2014. *Partial Differential Equations of an Epidemic Model with Spatial Diffusion*, (Feb). https://doi.org/10.1155/2014/186437.

[14] et al, Srikanth R. 2021. "Genetic Algorithm-based Predictive model for COVID-19 – Theoretical Model based on an evolutionary approach." https://doi.org/10.21203/rs.3.rs-107714/v2.

[15] et al , M.T.Rouabah.November 18, 2021, Genetic algorithm with cross-validation-based epidemic model and application to the early diffusion of COVID-19 in Algeria .https://www.sciencedirect.com/science/article/pii/S2468227621003513

[16] Iba, T., Levy, J.H., Connors, J.M. *et al.* The unique characteristics of COVID-19 coagulopathy. *Crit Care* 24, 360 (2020).

https://doi.org/10.1186/s13054-020-03077-0

[17]McCall, J. (2004). *Genetic algorithms for modeling and optimization*. doi:10.1016/j.cam.2004.07.034 | Elsevier Enhanced Reader

https://www.sciencedirect.com/science/article/pii/S0377042705000774

[18]M. A. Bahloul, A. Chahid and T. -M. Laleg-Kirati, "Fractional-Order SEIQRDP Model for Simulating the Dynamics of COVID-19 Epidemic," in *IEEE Open Journal of Engineering in Medicine and Biology*,

https://ieeexplore.ieee.org/abstract/document/9178435

[19] Romain Ducasse. Qualitative properties of spatial epidemiological models. 2020. Ffhal-02571610f

https://hal.archives-ouvertes.fr/hal-02571610/document