Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Applied Soft Computing Journal 96 (2020) 106692

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

A data-driven understanding of COVID-19 dynamics using sequential genetic algorithm based probabilistic cellular automata

Sayantari Ghosh [a], Saumik Bhattacharya [b,*]

[a] Department of Physics, National Institute of Technology Durgapur, India
[b] Department of E & ECE, Indian Institute of Technology Kharagpur, India

ARTICLE INFO

Article history:
Received 4 August 2020
Received in revised form 25 August 2020
Accepted 26 August 2020
Available online 29 August 2020

Keywords:
Epidemiological model
Probabilistic cellular automata
Genetic algorithm
Real data modelling

ABSTRACT

COVID-19 pandemic is severely impacting the lives of billions across the globe. Even after taking massive protective measures like nation-wide lockdowns, discontinuation of international flight services, rigorous testing etc., the infection spreading is still growing steadily, causing thousands of deaths and serious socio-economic crisis. Thus, the identification of the major factors of this infection spreading dynamics is becoming crucial to minimize impact and lifetime of COVID-19 and any future pandemic. In this work, a probabilistic cellular automata based method has been employed to model the infection dynamics for a significant number of different countries. This study proposes that for an accurate data-driven modelling of this infection spread, cellular automata provides an excellent platform, with a sequential genetic algorithm for efficiently estimating the parameters of the dynamics. To the best of our knowledge, this is the first attempt to understand and interpret COVID-19 data using optimized cellular automata, through genetic algorithm. It has been demonstrated that the proposed methodology can be flexible and robust at the same time, and can be used to model the daily active cases, total number of infected people and total death cases through systematic parameter estimation. Elaborate analyses for COVID-19 statistics of forty countries from different continents have been performed, with markedly divergent time evolution of the infection spreading because of demographic and socioeconomic factors. The substantial predictive power of this model has been established with conclusions on the key players in this pandemic dynamics.

1. Introduction

With its outbreak in Wuhan, China, Coronavirus disease-2019 (COVID-19) has spread across the world within a few months. Due to its explosive growth and considerable rate of fatality, World Health Organization (WHO) declared COVID-19 as a pandemic and a global health emergency [1]. According to the available statistics in June, 2020, the total number of infections by SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2), the causative agent of this disease, is approaching 19 million around the world, causing around 700,000 deaths in 213 countries and territories, with no effective vaccination available in the market so far. Beyond respiratory discomforts including pneumonia, dry cough, cold and sneezing [2,3], it has been reported to cause liver and gastrointestinal tract maladies, kidney dysfunction and heart inflammation, in cases of severe infection [4–6]. This highly infectious disease transmits from person-to-person through respiratory droplets produced by infected person. Fomite-mediated and nosocomially acquired infections are also being identified as important sources of viral diffusion [7–9]. A typical incubation time from exposure to symptoms has been reported for COVID-19, while infection transmission from asymptomatic individuals has been observed as well [10–12].

Immediately after the detection of human-to-human transmission, the government agencies of various countries started implementing several mitigation strategies to control the epidemic. The measures thus taken include social distancing, restrictions on domestic as well as international travel, cancelling social events, shutting down of public as well as commercial activities etc. which can effectively reduce the possibilities of physical human contact. Moreover, contact tracing, aggressive testing as well as hospital or home quarantine for infected individuals and suspected cases have also been executed to track and prevent further spread. However, these strategies are directly contributing to enormous economical loss. The optimum estimation of this novel disease dynamics is emerging out as a challenging problem in this context. The immense disruption caused by COVID-19, resulting into overwhelming disorder in the health, economy and lives of billions of people around the globe, has brought the necessity for accurate modelling of infectious diseases into the focus. The effect and effectiveness of this complex interplay between differing

* Corresponding author.
E-mail addresses: sayantari.ghosh@phy.nitdgp.ac.in (S. Ghosh), saumik@ece.iitkgp.ac.in (S. Bhattacharya).

https://doi.org/10.1016/j.asoc.2020.106692
1568-4946/

length-scales and time-scales with the applied control strategies can only be understood and predicted with the help of precisely designed quantitative models.

1.1. Models for understanding COVID-19 statistics

With a tremendous effort from researchers around the world, a spectrum of various mathematical and computational approaches is being used to understand and predict COVID-19 statistics, addressing its different perspectives. On a rudimentary sense, the studies being pursued can be segmented in two categories: (i) data science and machine learning approaches and (ii) differential equation based mathematical modelling techniques. The first group of studies trusted mostly on data mining from national/international repositories (e.g., WHO, country specific data centres etc.) or popular social media platforms to forecast the active cases and mortality data [13–17]. The major goal of these studies are to estimate and predict the time evolution of the disease using specific computational concepts, like Monte Carlo decision making, fuzzy rule induction, deep learning etc [18–22]. Some of these studies also explored impact of disease control interventions, like, travel restrictions [23], patient quarantining and isolation [24], medical facilities [25], social distancing and administrative responsibility [15] on epidemic spreading rate. Though these models are quite effective, being entirely dependent on data, the efficiency of these studies can be heavily inclined towards the data quality. As comprehensively reviewed by [26], several data-dependent models are prone to suffer from high risk of bias, which is very much probable for imprecise short time series data.

With the evidence of giving effective predictions for past pandemics [27–29], the traditional approaches of the mathematical theory of epidemiological dynamics also have driven several researchers to study COVID-19 dynamics. Theoretical modelling based approaches have been long associated to understand and predict the outbreak probabilities and seriousness of a disease, and provide key information to control the intensity [30–33]. Most of the mathematical models that are being used to investigate the COVID-19 dynamics [34–37] are based on variants of classical deterministic model of susceptible-infectious-recovered (SIR) that was introduced by Kermack and McKendrick [38]. Constituting a set of nonlinear ordinary differential equations (ODE), the SIR model compartmentalizes the population where susceptible subpopulation declines over time, constantly getting infected (by infectious subpopulation), and then recovered from (and gaining immunity to) the disease over time. Being powerful and computationally favourable tool to analyse epidemic, variants of this methodology are common in understanding real epidemic data [39,40]. Though these models capture the disease transmission dynamics, being deterministic, they suffer from the assumption of homogeneous mixing, forgoing the spatial information.

For modelling real-world dynamics of a disease that spreads from close-contacts only, the tool needs to accommodate neighbourhood information. Moreover, the platform requires to take into account of stochasticity of real dynamics, spatial infection spread and inherent heterogeneity in population, which are some major limitations of the mentioned works. Thus, the identification of research gap points out in a direction of designing a methodology that addresses the above mentioned issues to understand and predict neighbourhood-dependent person-to-person probabilistic transmission of COVID-19, that should be powered with extensive computational tools for parametric optimization.

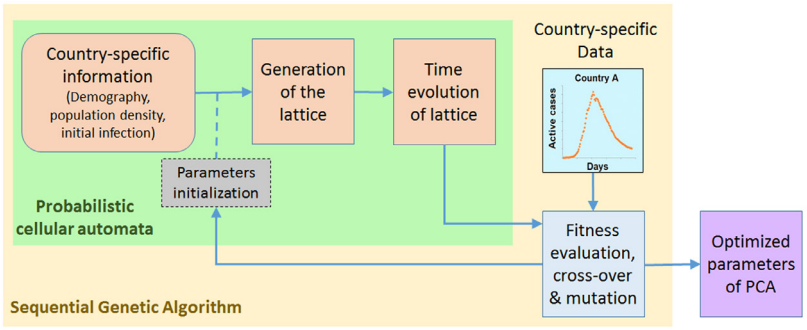
1.2. Motivation and contributions

In this study, we propose probabilistic cellular automata based dynamical model, optimized through sequential genetic algorithm for an accurate assessment of the extent of COVID-19 dynamics. The major motivation of using cellular automata (CA) is its ability in depicting extremely complex macroscopic outcomes, while being based on local interactions that trusts on the interaction of a multitude of single individuals [41,42]. This methodology is capable of giving a direct correspondence to the physical system and also rectifies the major drawbacks of ODE models by (i) tracking individual contact processes, (ii) giving room for introducing probabilistic individual behaviour, and (iii) capturing neighbourhood as well as global spatial information. Because of these reasons, CA based approaches have been successfully used as a competent substitute method to simulate physical, biological, environmental and social contagion-like spreading [43–46]. For studying past epidemics as well as interpreting COVID-19, some studies have proposed cellular automata as an alternative method [47–50]. However, to capture and interpret the behaviour of real data through CA needs a large-scale parameter optimization that could be time consuming as well as sub-optimal. Thus, though being extremely flexible and powerful, CA has not been yet optimized to understand and interpret COVID-19 data for countries worldwide. To explore this, in this study, genetic algorithm (GA) has been employed, which is a well-known method for generating the optimal parameter subset through stochastic search procedures based on the principle of the survival of the fittest [51–55]. Cross-over and mutations, two key properties of genetic algorithm help to optimize the parameter set efficiently in limited steps. Cellular automata coupled with genetic algorithm has been used before to explore evolutionary aspects of game theoretical problems [56], but to the best of our knowledge analysing and developing understanding from real pandemic data like COVID-19 using optimized CA platform has not been attempted yet. The main contributions of this work are as follows:

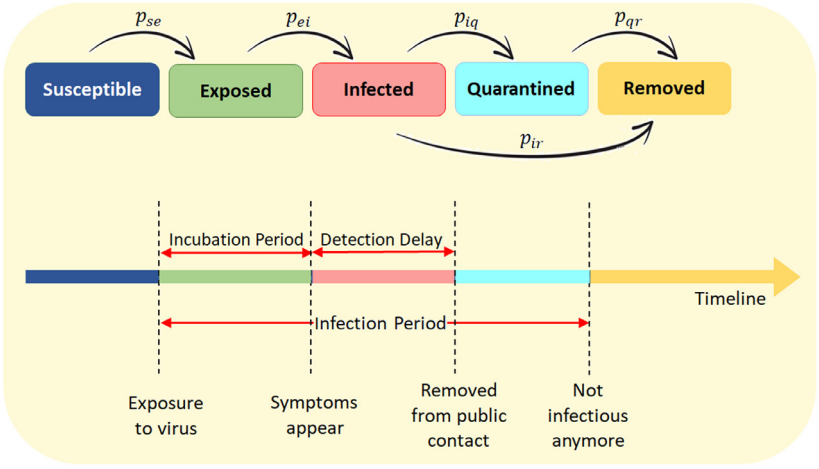
- To build a CA model which is probabilistic, so that it can take into account of demographic variations, neighbourhood diversity and uncertainties of real dynamics.
- To create an easily implementable framework where optimization using GA will be done sequentially for all parameters associated with the transition rules of the CA model for real data interpretation.
- To interpret and understand COVID-19 disease transmission dynamics with an optimized CA framework, which can be extended for prediction as well.

Through this, on one hand, one can track the individual contact process through time and space; on the other hand, a self-adapting process of evolutionary strategies has been created by designing the chromosome with parametric genes and establishing fitness function that maximizes over the generations. The main limitations of the state-of-the-art algorithms and the major contributions of the proposed method are listed in Table 1 for a clear understanding. The main rationality behind this approach is that it is extremely difficult to find the optimal parameter of the complex spatial epidemiological model using random search or analytical techniques. The proposed GA based framework helps to search the parameter space more efficiently for the optimal performance of the entire algorithm.

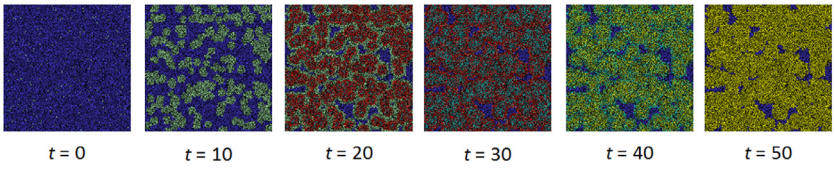
The rest of this article is organized as follows: Section 2 includes the proposed concepts of epidemiological model, probabilistic cellular automata and the sequential genetic algorithm used in this work. In Section 3, the results has been elaborately discussed where the optimized CA model has been employed for



(a)



(b)



(c)

Fig. 1. An overview of the dynamics: (a) Object process diagram of the proposed model; (b) The schematic diagram of the disease transmission dynamics in form of a modified SEIQR model. Transition probabilities p_{se}, p_{ei}, p_{iq}, p_{ir} and p_{qr} are pointed out. The associated state transition delays are indicated on the timeline of the disease dynamics. (c) Time evolution of the spatial lattice during spread of the infection in a population. The colours of the respective subpopulations, (i.e., susceptible, exposed, infected, quarantined and removed) are same as depicted in (a).

Table 1
Comparison of the proposed method with the state-of-the-art COVID-19 models.

Basic methodology	Differential equation models	Data science approaches
References	[33–37,39,40]	[13–22]
Limitations	a) Homogeneous Mixing b) Most models are considered as deterministic	(a) No way to track person to person transmission. (b) No neighbourhood consideration.
Contribution	Proposed method, (a) accommodates heterogeneity in population (b) includes stochasticity and probabilistic dynamics (c) estimates optimum epidemic dynamics parameters. (d) considers neighbourhood and demography explicitly. (e) performs robust prediction with limited data.	

simultaneously understanding as well as analysing active infections, total infections and total death caused by COVID-19 for several countries, considering the demographic and spatial population density variations. Section 4 is comprised of concluding remarks.

2. Proposed methodology

An object process diagram of the proposed method has been depicted in Fig. 1(a). The methodology starts with the infection spreads following the SEIQR epidemiological model in a random human population over a 2D grid, initialized on a country-specific basis. The parameters of the epidemiological model is continuously optimized using proposed sequential genetic algorithm to match the real country-specific infection spread data. The proposed methodology is consisted of three distinct parts— (A) epidemiological model that governs the infection spreading, (B) probabilistic cellular automata (PCA) to model the dynamics of the pandemic spread and (C) optimization of the parameters associated with PCA using genetic algorithm (GA) to fit real-world data.

2.1. Epidemiological model

In the epidemiological model, the entire population is partitioned in five distinct parts. At the very beginning, every person was healthy but they are vulnerable to the infection. These people are denoted as susceptible (S) subpopulation. At time instance $t = 0$, some people in the population got exposed to the infection from some known or unknown source. These exposed people do not have any particular symptom of the infection, but they can spread the infection to the susceptible people. These asymptomatic people are referred as exposed (E) subpopulation. At time instance $t = 0$, there were also some people who had clear symptoms of the infection and they also had the potential to spread the infection among susceptible people. This symptomatic people are considered as infected (I) subpopulation. After an incubation period, some of the exposed people show the symptoms of the infection and they move to subpopulation I. Because of the health facilities and testing time, the infected people are detected with some average delay, and put to quarantine. The people who are quarantined cannot spread the infection to other people, though they themselves remain in the infectious stage. These people are denoted as quarantined (Q) subpopulation. Both the quarantined people and the infected (but not detected) people would come out of the infectious stage eventually, and after that they no longer contribute in the infection spreading dynamics. These people are denoted as removed (R) subpopulation in the model. This removed subpopulation contains two kinds of people— one who have recovered from the infection completely and they neither infect nor get infected in future, and the other kind of people who have died due to the severity of the infection. Schematic diagram related to the transitions, probabilities and timelines corresponding to the dynamics of infection are shown in Fig. 1(b). In the analysis, normalized subpopulations have been considered, and the respective normalized subpopulation is denoted using the same lowercase character. For example, the normalized susceptible and infected subpopulations are denoted by s and i respectively. As shown in Fig. 1(c), this epidemiological time evolution has been implemented on a 2D lattice using PCA as discussed below.

2.2. Probabilistic cellular automata

Let L be a finite subset of $\mathbb{Z}^2$ at time instance t, denoted as $L \sqsubset \mathbb{Z}^2$ which defines a regular 2D lattice. Every point on this lattice $\mathbf{x} \in L$ can acquire finite number of *states* A. In this particular problem, the set A can be defined as $A = \{0, s, e, i, q, r\}$, where the terms s, e, i, q and r denote the particular possible states of infection as discussed in Section 2.1, and 0 denotes no human occupant or an empty space. At time $t = 0$, n_i^0 points are randomly selected on L and assign the state a_i where $i \in A$. The total initial population is defined as $N = \sum_{i \in A \setminus 0} n_i^0$. At any instance of time t, n_i^t, $i \in A \setminus 0$ denotes the total number of the people in respective state a_i.

For neighbourhood criteria, modified-Moore neighbourhood or d-neighbourhood has been used. A finite subset $\Omega_d \sqsubset \mathbb{Z}^2$ is defined, containing the origin $\mathbf{0} = (0, 0)$, and the cardinality of Ω_d is $4d(d + 1)$. General probabilistic cellular automata (PCA) is a stochastic process that describes sequence of mappings $\Lambda_t^a : L \to a$, $a \in A$, where any particular state $\Lambda_t^a(\mathbf{x})$ of $\mathbf{x} \in L$ at a particular time instance t is dependent on the previous states of the d-neighbourhood of $\mathbf{x}$, denoted as $x + \Omega_d = \{\mathbf{x} + \boldsymbol{\omega} : \forall \omega \in \Omega_d\}$ with certain probabilities. More precisely, in COVID-19 infection spread, $\Lambda_t^E(\mathbf{x})$ will be decided by $\Lambda_{t-1}(\mathbf{x} + \omega)$, $\forall \omega \in \Omega_d$. The other mappings $\Lambda_t^a(\mathbf{x})$, $a \in A \setminus E$, depends on the sequence of states $\Lambda_\kappa^a(\mathbf{x})$, $0 \leq \kappa < t$.

2.2.1. Transitional probabilities

The transition probability $p_{a_i a_j}^t$ denotes the probability of transition at time t from state a_i to state a_j, where $a_i, a_j \in A$. Without any loss of generality, $p_{a_i a_j}^t$ is denoted as p_{ij}^t and transition from state a_i to a_j as a_{ij} in the rest of the discussion for a simpler notation. In cases, where $a_i \neq a_j$, p_{ij}^t is referred as state transitional probability, and if $a_i = a_j$, p_{ii}^t is called as self transitional probability.

If a state transition a_{ij}, $i \neq j$, happens in $\mathbf{x}$ at time t following the transition probability p_{ij}^t and the transition state a_{ij} has a transitional delay τ_{ij}, then

$$p_{ij}^t = \begin{cases} 0 & \text{if } t < t_{ui} + \tau_{ij} \\ p_{ij} & \text{if } t \geq t_{ui} + \tau_{ij} \end{cases}$$

where t_{ui} is the time instance when transition a_{ui}, $u \neq i$ happened. In this infection diffusion model, only the state transitional probabilities p_{se}^t, p_{ei}^t, p_{ig}^t, p_{qr}^t and p_{ir}^t are considered to be nonzero at certain instance of time, and for all the other transitional probabilities, τ_{ij} is set to infinity, where p_{ij} and τ_{ij} are user defined parameters. However, for the transition a_{se}, t_{ui} and τ_{ij} are set to zero, and for $\mathbf{x} \in L$, let us define $p_{se}^t = p_{ij} = 1 - p_{ss}^t$ and the self-transition probability $p_{ss}^t = (1 - p_i)^{i_{t-1}}(1 - p_e)^{e_{t-1}}$ where i_{t-1} and e_{t-1} are the number of cells in states i and e respectively in the Ω_d neighbourhood of $\mathbf{x}$ at time $t - 1$. The probabilities p_e and p_i are defined as 'infection probabilities' which can be considered as the probabilities that a susceptible person become exposed to the infection when that person meets an exposed or an infected person respectively.

An empty cell does not contribute in the infection spread, and thus, self transitional probability $p_{00}^t = 1$, $\forall t$. Among the total removed population r_t at time instance t, a population fraction $p_\beta r_t$ is considered that recover from the infection at time instance t and acquire long-term immunity towards the disease, and a population fraction $(1 - p_\beta) r_t$ is considered to be deceased. The removed population r_t is not considered further in the infection dynamics and it is taken that $p_{rr}^{t'} = 1$, $t' > t$.

Table 2
Descriptions of the parameters used in the proposed work.

Notation	Description
L	Spatial lattice
A	Set of possible states on lattice
$A \setminus 0$	Set of epidemiological states
n_i^t	Total number of people at state a_i at time t
Ω_d	d-neighbourhood of $\mathbf{x} \in L$
Λ_t^a	Mapping $L \rightarrow a$ at time t
p_{ij}^t	Probability at time t that $\mathbf{x} \in L$ moves from a_i to a_j
τ_{ij}	Transitional delay for $\mathbf{x}$ to move from a_i to a_j
e_t, i_t	Number of exposed and infected people in the d-neighbourhood of $\mathbf{x}$ at time t
p_e, p_i	Probabilities that an exposed or an infected person spreads the infection to a susceptible person when they meet
Θ	A gene containing all the parameters of PCA method
B	Binary encoded representation of Θ
$G(\Theta)$	The PCA model with parameter Θ
$\mathbf{y}$	Time series of an epidemiological state in a country
$\hat{\mathbf{y}}$	Time series estimate of epidemiological state from PCA
e_{ji}	Estimation error of jth gene in ith generation
N_g	Total number of chromosome in genepool
F	Number of parents selected for mating from N_g
p_β	Fraction of r_t that recovers from the disease
ρ	Fraction of parents F that lives in the next generation

2.3. Parameter optimization using GA

Though PCA has potential to model the probabilistic transition of states on a spatial lattice, the main challenge to use it for modelling a real-world scenario is to find out the optimal parameters for the PCA. As the searching space for the proposed PCA model is very large, it is practically impossible to search for the optimal parameter setting manually to analyse the characteristics of the infection spread from a real data. Thus, genetic algorithm (GA) has been applied to find out the optimal parameter set given a real time-series data.

Let us assume a discrete time signal $y[\mathrm{n}]$, $0 \leq \mathrm{n} \leq (T-1)$ associated with the real world infection spread. The PCA model is denoted by $G(\Theta)$, where $\Theta = [\theta_1, \theta_2 \ldots \theta_h]$ denotes the set of parameters used for the PCA model. If $\hat{y}[\mathrm{n}]$, $0 \leq \mathrm{n} \leq (T-1)$ is the time evolution of the desired variable in the model $G(\Theta)$, then the objective is to find an optimal parameter set Θ^* such that $\hat{y}[\mathrm{n}] \rightarrow y[\mathrm{n}]$, $\forall \mathrm{n}$. To apply GA, each θ_i, $1 \leq i \leq h$, is encoded as a string of binary digits b_i [54,55] assuming the θ_i has a bound $|\theta_i| < \zeta_i$, $1 \leq i \leq h$. This binary string is referred as *gene*, and the concatenated genes in the order of the appearance of respective θ_i in Θ is called the *chromosome*. For example, if B is the chromosome corresponding to parameter set Θ, $G(B)$ is equivalent to $G(\Theta)$. A collection of N_g number of chromosomes of estimated parameters, often referred as *gene pool*, are evaluated at every time step (called as *generation*). In our work, the error of each chromosome has been evaluated using l_1 norm distance. At ith generation, the error of the jth chromosome B_{ji} is computed as

$$e_{ji} = \|\mathbf{y} - \hat{\mathbf{y}}_{ji}\|_1 = \sum_{\mathrm{n}=0}^{T-1} |y[\mathrm{n}] - \hat{y}_{ji}[\mathrm{n}]|$$

where $\hat{\mathbf{y}}_{ji}$ is the estimated output of $G(B_{ji})$ in the vector form and $\hat{y}_{ji}[\mathrm{n}]$ is the value of $\hat{\mathbf{y}}_{ji}$ at time instance 'n'. At each generation, GA finds out $min(e_{ji})$, $\forall j$ and tries to make $e_{ji} \rightarrow 0$ as $i \rightarrow \infty$. In the proposed framework, some of the parameters are related to probabilities having a range 0 to 1, and some of the parameters are associated with time (in days) which are discrete integers, and greater than or equal to zero in our case. Thus, the parameters are initialized randomly keeping their domain restrictions intact.

For mating, two chromosomes, often referred as *parents*, are selected from the gene pool considering their '*fitness*'. Among two selected parents, a crossover point or a splice point is selected at b_i, $1 \leq i \leq h$ in both chromosomes and a crossover [55] happens that produces two offsprings. In our approach, fitness f_{ji} of each chromosome has been defined as the inverse of their respective errors at a particular generation. At each generation, F number of best chromosomes are selected from the gene pool having the maximum fitness for mating. Following the idea of [52], ρF number of parents are kept to the next generation along with the new chromosomes to ensure that the error in the next generation is always less than or equal to the current generation. Selecting ρF number of chromosomes from the parents, $N_g - \rho F$ number of children are produced from mating to keep the size of the gene pool constant. After the offsprings are generated, in the parameter space, s genes are randomly selected and small perturbations are added individually to mimic mutation.

As shown by several researchers [57], the homogeneity in the gene pool increases with the generations, and as the perturbations due to mutation are typically small, the reduction of error becomes a problem after a few generations. Thus, to restrict homogeneity in the gene pool, a small number of offsprings μ are selected from the total $N_g - \rho F$ number of generated offsprings, and replaced them with randomly generated chromosomes to maintain diversity. This step is called as 'diversification' of gene pool.

In our problem, the parameters Θ of the PCA model $G(\Theta)$ are the state transitional probabilities p_{ei}, p_{iq}, p_{ir}, p_{qr}, infection probabilities p_e and p_i, state transition delays τ_{ei}, τ_{iq}, τ_{qr}, τ_{ir}, neighbourhood d, and death probability p_β as mentioned in Section 2.2. As optimizing these many parameters simultaneously might be challenging and require huge amount of resources, we propose a variant of GA with sequential evolution mechanism where instead of optimizing the solutions simultaneously, the parameters are optimized sequentially. Let us define a set of generations as an *era*. For the first era containing a small number of generations, a traditional GA methodology is followed as discussed this far to have a set of initial parameters. From the next era onward, two parameters are fixed and optimized sequentially in that era. Mutation and crossover are restricted to those two respective genes, whereas parent selection is done based on the performances of the entire chromosomes. This newly proposed sequential optimization of parameters of PCA using GA is defined as PCA-GA. The proposed approach can optimize a large number of parameters using limited resources efficiently. All the notations used in PCA-GA are briefly summarized in Table 2.

Proposed PCA-GA has a complexity which can be approximated as $O(N_g T_g O(f))$ where N_g is the number of population, T_g is the total generation and $O(f)$ is the complexity to measure the fitness in the GA. For a large enough N_g, T_g is considered as a comparatively smaller constant and thus, the complexity of the entire algorithm is mainly governed by N_g and $O(f)$. The complexity of estimating the fitness can be approximated as $O(f) = O(T + 8N\tau T)$ for Moore neighbourhood criteria, where N is the total population on the 2D grid. The length of the original time series data T, and τ, the maximum of τ_{ij}, are both constant, and thus $O(f)$ can be represented as $O(N)$.

Though GA has been selected as a strategy to optimize the parameters of the proposed PCA model, it is evident that because of the generalized construction of the proposed framework, other meta-heuristic methods could also be employed to search the parameters of the spatially driven SEIQR model which is the main focus of this work. However, presence of mutation and diversification in GA help to search for better solutions as the search space is extremely large.

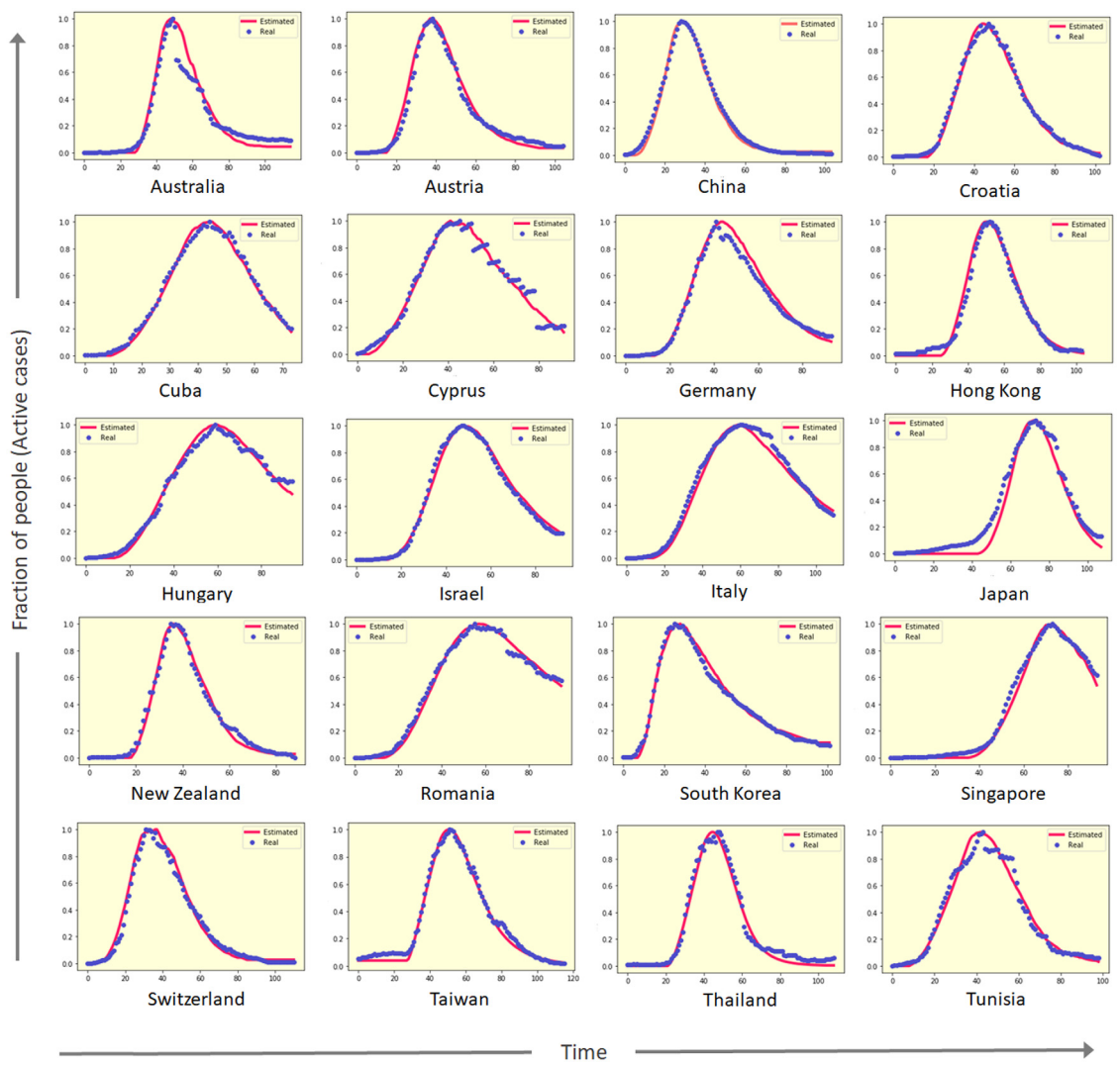


Fig. 2. Time series data for active cases (blue) of COVID-19 pandemic in different countries where the peaks of the infection spread of the first wave have been passed, and estimated active cases (red) from proposed PCA-GA method.

3. Results

To validate the effectiveness of the proposed framework, using PCA-GA, the actual statistics of COVID-19 spreads till 20th June, 2020 in different countries is used. For finalizing the data-set from available data of 213 countries, several aspects have been considered. At first, 102 countries had been dropped due to less number of reported cases (less than 1000 reported cases till 20th June 2020). Out of the remaining countries, some countries, like Iran, Greece, Paraguay etc., are removed due to data inconsistency, and finally 40 countries are randomly selected ensuring the following points:

- At least 2 countries from each continent got selected to maintain demographic diversity in our data.
- Care has been taken to maintain significant variation in population density, which we believe as a major factor contributing in disease transmission.
- It was ensured that countries from three distinct stages of COVID-19 infection are considered: (i) where the infection is significantly diminished, (ii) where the peak infection has been reached but substantial infection still persists, and (iii) where consistent growth in infection is occurring.

With these widely variant spectrum of time series data, we proceed for quantitative calibration and interpretation through the proposed methodology. All data samples are taken from the website worldometers.info.[1]

To point out the major contributing factors in dynamics of infection spread, for every country under consideration, three available time series, namely daily active cases, total number of infected cases and total number of deaths are accumulated. Out of these three series, the daily active cases time series is used for model formulation, and the rest are considered for model validation. It is important to mention that the population q_t is the relevant observable here, as infected people as i_t and e_t remain latent and undetected in the population. The reported daily active case data is associated with lifetime of the infection, and are used in this study to check the effectiveness of the proposed framework as follows. By applying PCA-GA on the daily active case data of a particular country, the parameters Θ^* that gives the minimum l_1 error is extracted. For validation of the optimized parameters and understanding the robustness of the algorithm, results generated by using $G(\Theta^*)$ for the total infected states and deceased states are then compared with the real-world data.

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

Here it must be noted that the optimal parameters Θ^* remain unaltered and no further optimization is performed.

3.1. Experimental setup

For all the simulations, PCA is initialized with a fixed lattice size of 100 × 100 with $n_e = 50$ and $n_i = 4$. The population n_q and n_r are set to zero at $t = 0$. The susceptible population n_s has been initiated depending on the population density of a country as follows: among the countries considered in our study, for the country with lowest population density (Canada), $n_s = 2500$ has been selected, and for the country with highest population density (Singapore), $n_s = 6000$ has been fixed. For any other country, n_s has been assigned within this range using logarithmic scaling based on the population of that country. As each of the parameters of PCA-GA has physical relevance, the sequential searching process has been initiated by following restrictions of ranges. It is important to note that in our problem, genes associated with probabilities are initiated in the range [0,1] and clipped during the optimization process accordingly. The state transition delays τ_{ei} (incubation period) and τ_{iq} (testing delay) are considered to be within the range (0,30). The transition delay τ_{ir} and τ_{qr} (corresponding recovery periods) are initialized in the range (20,100). All the simulations are executed in a system with Intel Core i7 8700K processor, 64 GB RAM and 8 GB NVIDIA GeForce RTX 2080 8 GB GPU using Python and numpy packages.

3.2. Estimation of parameters using active cases

The daily active cases can be defined as the $c_t = c_{t-1} + q_t - r_t$ where c_t is the number of active cases at time instance t having the initialization $c_0 = 0$. In Fig. 2, the active cases of 20 different countries are shown along with the respective estimated active cases using PCA-GA model. For the countries shown in Fig. 2, the first peak of the infection is already crossed and a steady fall in the infection spread is observed. It can also be seen that some of the active cases of the countries like China, Israel, Switzerland, follow smooth bell-shaped curves, whereas for some countries, like Australia, Cyprus, Hungary etc., the times series data deviates from bell-shaped curves with substantial degree of noises. In all the cases, PCA-GA has successfully captures the trend of the time series data estimating the parameters of the epidemiological process. To measure the goodness of the model estimation, three different metrics has been used to measure the quality of the estimated values. The root mean square (RMSE) distance, correlation distance and chi-square distance [58–60], denoted as d_l, d_c and d_χ respectively, are computed between the real data and the estimated values from the PCA-GA model to evaluate the effectiveness of the optimized model. For two vectors $\mathbf{u}$ and $\mathbf{v}$, we define

$$d_l = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(u_i - v_i)^2},\ d_c = 1 - \frac{(u-\bar{u}).(v-\bar{v})}{\|(u-\bar{u})\|_2\|(v-\bar{v})\|_2},$$

$$d_\chi = \sum_{i=0}^{T}\frac{(u_i - v_i)^2}{v_i}$$

where T is the length of each vector, u_i and v_i are the ith elements of $\mathbf{u}$ and $\mathbf{v}$ respectively and (.) denotes dot product of two vectors. As shown in Fig. 3(a), the proposed model performs well in modelling the real data. When evaluated over all the countries considered in this work, the proposed model fits the data well, and for only 0% −12.5% cases the fittings were poor depending on the evaluation metric. It is important to mention that all the distance measures are evaluated on normalized data.

In Fig. 2, an interesting point to notice is that the peak of the active cases are located at markedly differing time instances, and the other properties, like variance, skewness etc., of the observed distributions are also varying drastically. The fundamental differences between the fitted curves are quantified with the help of boxplot of the parameters in Fig. 3(b)–(c) by analysing basic statistical properties. The reported boxplots are specifically for the countries selected in Fig. 2. It can be noted that p_e, p_i and p_{ei} exhibit a wide variability in Fig. 3(b). During our analysis, a strong positive correlation with population density for p_e and p_i has been also observed. This can be thus inferred that the variation in population density in the considered countries causes the wide range of these parameters. It can be also concluded that high density of population increases the probability of transmission of the disease. The considerable difference in the mean magnitudes of the infection associated probabilities (p_e, p_i and p_{ei}) and recovery-related probabilities (p_{iq}, p_{ir} and p_{qr}) indicate the sharper rise and slower fall of active cases curves, which results into a skewed distribution in most of the cases (see Fig. 2). In Fig. 3(c), it is also shown that τ_{ei}, which is identified as the incubation time in the model, exhibits a range of 3–14 days with a mean at 7.3, which perfectly aligns with the observed cases all around the world [61]. In this figure, a wide variability in the range of τ_{ir} and τ_{qr} is observed, which points out the substantial difference in health infrastructure of these countries.

Here it must be mentioned that, while performing this statistical analysis with all 40 countries, some countries were detected showing consistent outliers (not included in Fig. 3(b)–(c)) in terms of four transitional parameters: p_{ir}. p_{qr}, τ_{ir} and τ_{qr}. While analysing the active case distributions of these outliers, it was found out that the time series data for all these countries have a saturating trend where the daily active cases do not show an average descent with time. Some of such cases are shown in Fig. 4. Even for these data which have drastically different qualitative trend compared to countries shown in Fig. 2, the proposed PCA-GA framework has successfully captured the trend of the real time series data accurately.

There are also certain countries, like India, Brazil, Chile, Mexico, etc., for which the infection spreading started later than the countries like China or Italy, and the active daily cases are still growing almost exponentially. As shown in Fig. 5, PCA-GA is able to estimate the time series data for these countries where the infection is spreading rapidly. Dynamics of COVID-19 spread in these countries are of particular interest as the prediction of the peak positions in these countries might help immensely to understand the maximum socioeconomic impact of the disease at a time in that geographical location.

3.3. Validation of the proposed model

While analysing a complex dynamics like the spread of a pandemic, it is not always sufficient to model the input real data only. It is required that the optimized model should be robust and can provide meaningful interpretations without further retraining or parameter tuning for real-world applications. To validate the robustness and the effectiveness of the proposed algorithm, the optimized model is now employed for three different tasks. At first, the robustness of the optimized model is checked by estimating the total number of infected cases, followed by total number of death cases without any further training, tuning or supervision. Finally, to further validate the efficiency of the model, its performance has been evaluated for the prediction task by training the model with partitioned data and evaluating on its future predictions without any further optimization.

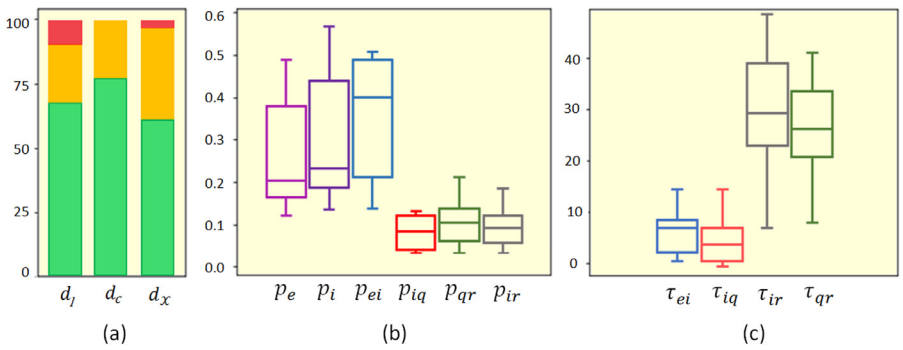


Fig. 3. Parameter estimations and goodness of model estimation: (a) RMSE, Correlation and χ^2 distance, d_l, d_c and d_χ for all 40 countries considered in this work in terms of goodness of agreement with model estimations shown in percentage. The colours green, orange and red signify level of agreement. Values between (0:0.05) for d_l, (0:0.01) for d_c and (0:1) for d_χ are considered as good (green). Values between (0.05:0.08) for d_l, (0.01:0.1) for d_c and (1:3) for d_χ are considered as moderate (orange). Values above moderate are considered as poor (red). For all three metrics $65-75\%$ countries have shown good agreement with model estimation; (b) and (c) represent boxplot for the best-fit parameters of state transition probabilities and state transitional delays respectively, for all the 20 countries shown in Fig. 2. The height of the boxplots represents the interquartile range (IQR). The dark line inside the box represents the median. The lower and upper whisker extend to the lowest and highest values within 1.5 IQR of the first and third quartile, respectively.

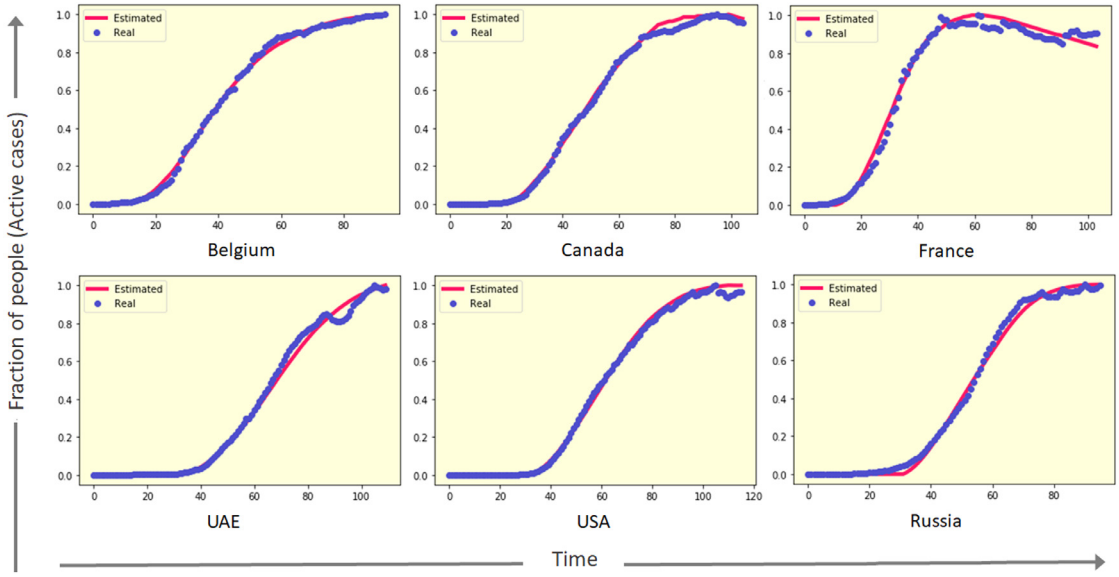


Fig. 4. Time series data for active cases (blue) of COVID-19 pandemic in different countries where the cases are saturating, and estimated active cases (red) from proposed PCA-GA method.

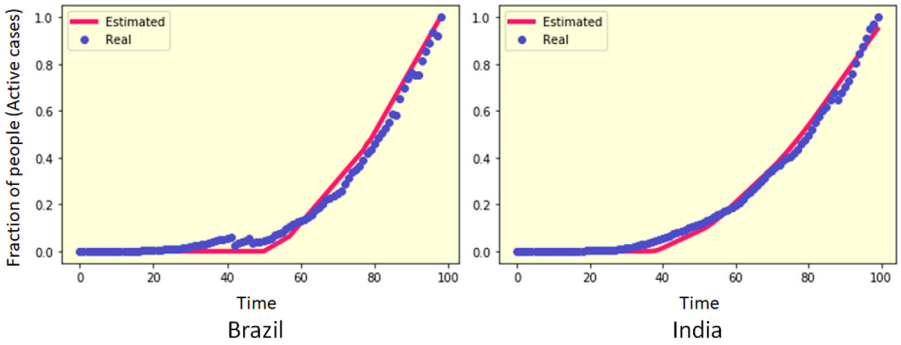


Fig. 5. Time series data for active cases (blue) of COVID-19 pandemic in different countries where the cases are increasing exponentially, and estimated active cases (red) from proposed PCA-GA method.

3.3.1. Total number of infected

The total number of infected cases z_t at time instance 't' can defined as $z_t = \sum_{i=0}^{t} q_i$. This cumulative sum indicates the total number of people who suffered from the disease at any point of time. For a country, where the first wave of the infection has passed, e.g., Croatia, Italy, etc., z_t follows a sigmoid function approximately, whereas for the countries like India, Mexico etc., where the infection has not reached the peak, z_t follows an exponential function. As PCA-GA is optimized using the time series information of daily active cases c_t, z_t is used to validate

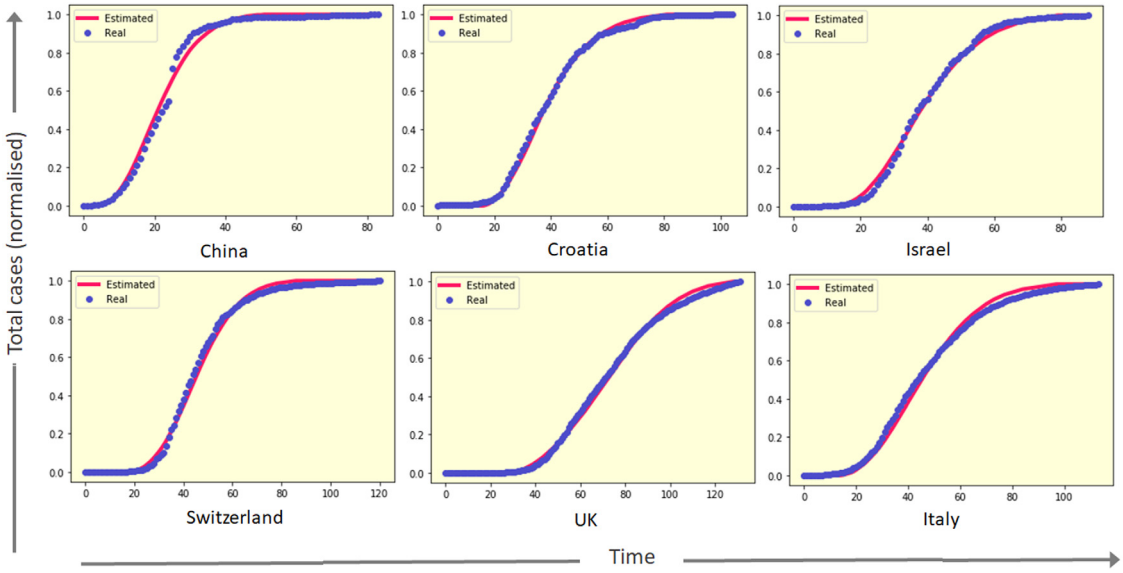


Fig. 6. Total infected cases (blue) of COVID-19 pandemic in different countries, and estimated total cases (red) from proposed PCA-GA method.

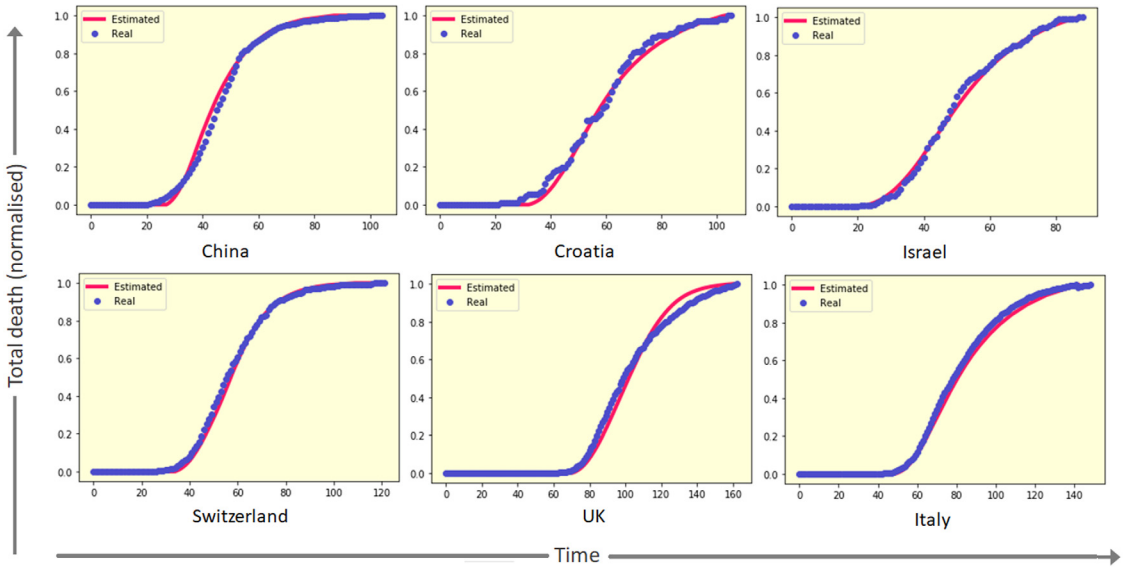


Fig. 7. Total deaths (blue) of COVID-19 pandemic in different countries, and estimated total deaths (red) from proposed PCA-GA method.

the parameters learnt by the sequential GA framework in the following way. Once a particular country is selected, Θ^* is estimated using PCA-GA with the actual c_t. Next the $\hat{z}_t$ for $G(\Theta^*)$ is calculated without any further fine-tuning of the parameters, and compared $\hat{z}_t$ with actual z_t. In Fig. 6, the total cases (blue) of six such countries are shown along with the best-fit results obtained from PCA-GA (red) which depict an excellent agreement with the data. It must be mentioned that for all three dynamical stages of infection spreading as discussed in Section 3.2, i.e., where the first wave of infection has passed, where the active cases are almost saturated currently or where the active cases are increasing rapidly, our estimated $\hat{z}_t$ closely matches z_t without any further parameter optimization. When evaluated over all 40 countries for the number of infected people, the proposed method gives average d_l, average d_c and average d_χ as 0.037,0.006 and 0.53 respectively, which exhibits the robustness of the model.

3.3.2. Total death cases

To further validate the 'goodness' of the estimated parameters, the parameter set Θ^* optimized over the daily active cases of a particular country is taken and the identical parameter values are used to compare the estimated total deaths with the actual total deaths of that country. Death in the population is the prime concern in case of the COVID-19 pandemic, and as mentioned in Section 2.2.1, daily deceased population is a fraction of r_t in our model. So, the total estimated death cases can be defined as $\hat{d}_t = (1 - p_\beta) \sum_{i=0}^{t} r_i$ where p_β and r_i for $0 \leq i \leq t$ are given by Θ^* and $G(\Theta^*)$ respectively. Fig. 7 demonstrates the comparison of the actual total death cases d_t with estimated total death cases $\hat{d}_t$ for Θ^*, the identical set of parameters used for estimating active cases as well as total cases previously. The same countries shown in Fig. 6 have been selected to show the robustness of the estimated parameter Θ^* using the proposed technique. Excellent agreement with data has been found for this case as well; when evaluated over all 40 countries for the total number of death cases, the proposed method gives average d_l, average d_c and average d_χ as 0.041,0.006 and 0.48 respectively.

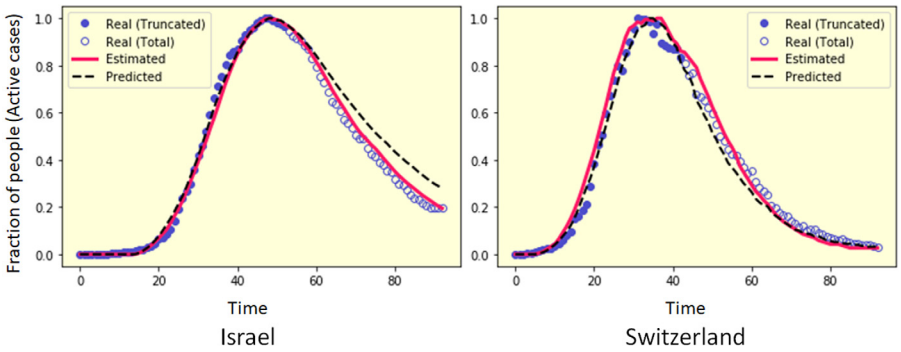


Fig. 8. Prediction of daily active cases from truncated data. For Israel and Switzerland, real data up to 54 and 43 days has been used to predict the daily active cases for 100 days. For prediction, the average of 50 independent PCA-GA simulations are considered.

3.3.3. Prediction related to infection spread

Prediction of future events is always challenging in data modelling [62]. For the final stage of validation of the methodology, the predictive power of the model has been tested. As the impacts of this pandemic becomes far reaching as the socioeconomic contexts vary, a considerably accurate prediction about the dynamics of the infection spread can be crucial and useful in many ways. As PCA-GA successfully estimates the optimal parameter Θ^*, the set of parameters can also be utilized to predict the future course of the infection in that country.

To validate the capacity of the prediction strategy, the daily active cases of a country c_t is truncated to c_P keeping the first 'P' values. PCA-GA is applied on c_P to estimate the parameters Θ_P. Then Θ_P is used to predict the daily active cases $\hat{c}_t$. As shown in Fig. 8, for two countries Israel and Switzerland, the daily active case information up to 54 and 43 days respectively are considered for an attempt to predict the daily active cases up to 100 days. In the figure, the estimated curve (shown in red) is optimized using all the real data points available, whereas the predicted curve (shown in black) is optimized using the truncated real data. It can be observed that the predictive estimation closely follows the real active case data, even though only $\sim$ 50% data points are used for parameter estimation. For Israel and Switzerland, 100 days prediction of the algorithm produces (d_l, d_c, d_χ) as $(0.056, 0.008, 0.95)$ and $(0.028, 0.005, 0.43)$ respectively. As prediction of the spread of the infection is one of the most challenging tasks, the predictive ability of proposed algorithm is compared with different baseline methods to better understands its performance. As only a very few data points were available in the truncated data, fast decision tree learning algorithm [63] and Random forest regression perform poorly and give (d_l, d_c, d_χ) as $(0.43, 0.49, 243.82)$ and $(0.439, 0.51, 252.6)$ respectively for the truncated time series of Switzerland. SVM regression with RBF kernel performs satisfactorily on the same truncated data and produces (d_l, d_c, d_χ) as $(0.09, 0.02, 27.8)$. However, the proposed PCA-GA algorithm significantly outperforms the baseline algorithms and produces (d_l, d_c, d_χ) as $(0.028, 0.005, 0.43)$.

3.4. Prediction for exponentially rising active cases

As the PCA-GA methodology has been elaborately validated in Section 3.3, now, in this section, it is employed for the purpose of prediction of consistently rising real epidemic data. Though the parameter estimation works well even when the minimum information about the peak position in c_t is available, the prediction task becomes really challenging when c_t is exponential in nature. For a particular country where c_t is almost exponentially rising, proceeding with prediction, first the best set of parameters Θ^* is detected by PCA-GA with fitness f^* and error e^*. As the drop of the infection heavily depends on the transitional probabilities

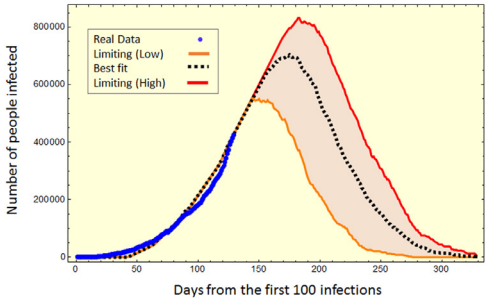


Fig. 9. Prediction of the course of the disease: Exponentially rising daily active cases for India (blue) till 20th July, 2020 are used for parameters estimation and the predictions.

p_{ir}, p_{qr} and state transitional delays τ_{ir} and τ_{qr}, this parameters are tuned to find a region of predictions bounded by the possible best case and the worst case scenarios. While estimating the best case scenario, p_{ir} and p_{qr} is chosen equal to the maximum and minimum p_{ir} and p_{qr} observed in the continent from which the country belongs. The reason behind this strategy is that the parameters related to the infection spreading are different in each continent which is also observed by [64]. In the best case scenario, transitional delays τ_{ir}^* and τ_{qr}^* are reduced to obtain best case transitional delays $\tau_{ir}^{\ominus}$ and $\tau_{qr}^{\ominus}$ respectively such that the fitness remain within 90% of f^*, where τ_{ir}^* and τ_{qr}^* are the corresponding optimized delays available in Θ^*. For the worst case scenario, we consider $\tau_{ir}^{\oplus} = \tau_{ir}^* + \alpha_{ir}$ and $\tau_{qr}^{\oplus} = \tau_{qr}^* + \alpha_{qr}$, where $\alpha_{ir} = \tau_{ir}^* - \tau_{ir}^{\ominus}$ and $\alpha_{qr} = \tau_{qr}^* - \tau_{qr}^{\ominus}$.

Fig. 9 depicts the prediction of the daily active cases using the method discussed so far. In Fig. 9, the black dotted line indicates the prediction using the optimal parameters Θ^* estimated using PCA-GA. The orange line indicates the best case scenario, where the maximum daily active cases would be minimized given the real data. The red line indicates the worst case scenario based on the specific conditions mentioned above. The best case and the worst case scenarios act as limiting cases of an area (shaded in pink colour) of probable future state. Any curve inside the pink region that contains the real data could be the evolution of the daily active cases in future given the real time series data, that is in exponentially rising state currently. This indicates that for India, which is now one of the biggest epicentres of COVID-19 in South-eastern Asia, the disease can start decline very soon if vigorous measures from government and complete support from the public could be achieved. It also shows that the maximum active cases on a day, that puts a direct burden on the health infrastructure of the country can be restricted below 750,000 if people participate to government indicated mitigation strategies,

and recovery rate remains at its current value. In that case, the peak of the disease is expected to pass during mid-September to mid-October, and the disease can be over with its first wave by March 2021. But these predictions also imply that the range of future states, that are possible for exponentially rising daily active cases, not only depend on the evolution of the epidemic so far, but also gets highly affected by the consistency and implementation efficiency of mitigation strategies.

4. Conclusion

COVID-19 outbreak has created a massive impact all across the globe. Even after nation-wide lockdowns, extensive testing strategies and medical supports, the spread of the virus has overwhelmed several countries. Thus, it is becoming more and more important to understand the nature of the infection spread and the key parameters that are controlling the spread. In this work, we proposed a probabilistic cellular automata model to understand and depict COVID-19 spread using appropriate choice of loss functions and evolutionary optimization framework. The parameters of this cellular automata model are optimized using sequential evolutionary genetic algorithm. It has been shown that this self-adapting methodology can be highly flexible and has the power to accurately estimate time trajectories of epidemics. This model works with physically interpretable parameters, which are accessible for analysis, data collection and further experiment, and can be readily identified with ground reality. This model has been successfully employed for optimizing all these parameters simultaneously for the daily active cases, total infected cases and total deaths with extreme robustness. The performance of the model has been exhibited for a large number of countries with huge diversity in population density, continents and available healthcare infrastructures. The predictive strength of the model has also been validated extensively, and demonstrated to estimate the course of the pandemic for the countries where infection peak has not been reached yet. It is important to mention that the motivation of the work was to develop a data driven, generalized, spatial framework that can be used to estimate relevant epidemiological parameters. This methodology is so powerful and flexible that physical interpretations of the results obtained from these analyses can have a wide range implications. Once the data is properly interpreted with the proposed methodology, interesting realistic features can be identified for specific countries. For example, in a pandemic situation, easily relatable factors like population clusters, variable population density, variable health facilities at different places of a country etc, can be studied to understand and predict emergence of new hotspots which can be used to design selective area containment strategies. While we propose and establish the applicability and strength of this framework in this work, we wish address these application perspectives in a study in our upcoming research studies.

With this proposed platform, the impact of individuality on contagion process can be explicitly studied, which might be directly related to the questions like lockdown behavioural differences, influence of rumours, vaccination opinion differences etc. As the effects of more complex dynamical factors like periodic lockdown or population clusters are not considered in this present model, the prediction capability of the proposed model is not satisfactory for time series data with abrupt discontinuities in the present form. The proposed framework could be enhanced with other l_p norm distances and different optimization techniques like multi-objective genetic algorithm or strength pareto evolutionary algorithm. Other swarm-based optimization techniques can also be explored for further refinement of the model. The potential of the proposed approach can be utilized to better understand the disease spreading and controlling, beyond this pandemic the world is facing currently, by keeping track of the spatial information of the dynamics, incorporating realistic behavioural aspects, and optimizing in terms of demographic as well as socioeconomic features.

CRediT authorship contribution statement

Sayantari Ghosh: Conceptualization, Methodology, Software, Validation, Writing - review & editing. **Saumik Bhattacharya:** Conceptualization, Methodology, Software, Validation, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] World Health Organization Coronavirus disease (COVID-2019) Situation Reports, 2020, Available at URL https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports. (Accessed June 2020).
[2] X. Jin, J.-S. Lian, J.-H. Hu, J. Gao, L. Zheng, Y.-M. Zhang, S.-R. Hao, H.-Y. Jia, H. Cai, X.-L. Zhang, et al., Epidemiological, clinical and virological characteristics of 74 cases of coronavirus-infected disease 2019 (COVID-19) with gastrointestinal symptoms, Gut 69 (6) (2020) 1002–1009.
[3] L. Pan, M. Mu, P. Yang, Y. Sun, R. Wang, J. Yan, P. Li, B. Hu, J. Wang, C. Hu, et al., Clinical characteristics of COVID-19 patients with digestive symptoms in Hubei, China: a descriptive, cross-sectional, multicenter study, Am. J. Gastroenterol. 115 (2020).
[4] Y. Cheng, R. Luo, K. Wang, M. Zhang, Z. Wang, L. Dong, J. Li, Y. Yao, S. Ge, G. Xu, Kidney disease is associated with in-hospital death of patients with COVID-19, Kidney Int. (2020).
[5] C. Han, C. Duan, S. Zhang, B. Spiegel, H. Shi, W. Wang, L. Zhang, R. Lin, J. Liu, Z. Ding, et al., Digestive symptoms in COVID-19 patients with mild disease severity: clinical presentation, stool viral RNA testing, and outcomes, Am. J. Gastroenterol. (2020).
[6] Y.-Y. Zheng, Y.-T. Ma, J.-Y. Zhang, X. Xie, COVID-19 and the cardiovascular system, Nature Rev. Cardiol. 17 (5) (2020) 259–260.
[7] C. Wang, R. Pan, X. Wan, Y. Tan, L. Xu, C.S. Ho, R.C. Ho, Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (COVID-19) epidemic among the general population in China, Int. J. Environ. Res. Publ. Health 17 (5) (2020) 1729.
[8] Y. Wang, Y. Wang, Y. Chen, Q. Qin, Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures, J. Med. Virol. 92 (6) (2020) 568–576.
[9] N. van Doremalen, T. Bushmaker, D.H. Morris, M.G. Holbrook, A. Gamble, B.N. Williamson, A. Tamin, J.L. Harcourt, N.J. Thornburg, S.I. Gerber, et al., Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1, New Engl. J. Med. 382 (16) (2020) 1564–1567.
[10] Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, M. Wang, Presumed asymptomatic carrier transmission of COVID-19, JAMA 323 (14) (2020) 1406–1407.
[11] H. Nishiura, T. Kobayashi, T. Miyama, A. Suzuki, S. Jung, K. Hayashi, R. Kinoshita, Y. Yang, B. Yuan, A.R. Akhmetzhanov, et al., Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19), medRxiv (2020).
[12] P. Yu, J. Zhu, Z. Zhang, Y. Han, A familial cluster of infection associated with the 2019 novel coronavirus indicating possible person-to-person transmission during the incubation period, The J. Infect. Dis. 221 (11) (2020) 1757–1761.
[13] G. Giordano, F. Blanchini, R. Bruno, P. Colaneri, A. Di Filippo, A. Di Matteo, M. Colaneri, Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy, Nature Med. (2020) 1–6.
[14] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, et al., Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions, J. Thoracic Dis. 12 (3) (2020) 165.
[15] V. Volpert, M. Banerjee, S. Petrovskii, On a quarantine model of coronavirus infection and data analysis, Math. Model. Nat. Phenom. 15 (2020) 24.
[16] C. Li, L.J. Chen, X. Chen, M. Zhang, C.P. Pang, H. Chen, Retrospective analysis of the possibility of predicting the COVID-19 outbreak from internet searches and social media data, China, 2020, Eurosurveillance 25 (10) (2020) 2000199.

[17] L. Li, Z. Yang, Z. Dang, C. Meng, J. Huang, H. Meng, D. Wang, G. Chen, J. Zhang, H. Peng, et al., Propagation analysis and prediction of the COVID-19, Infect. Dis. Model. 5 (2020) 282–292.
[18] S.J. Fong, G. Li, N. Dey, R.G. Crespo, E. Herrera-Viedma, Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction, Appl. Soft Comput. (2020) 106282.
[19] K. Chatterjee, K. Chatterjee, A. Kumar, S. Shankar, Healthcare impact of COVID-19 epidemic in India: A stochastic mathematical model, Med. J. Armed Forces India (2020).
[20] S.J. Fong, G. Li, N. Dey, R.G. Crespo, E. Herrera-Viedma, Finding an accurate early forecasting model from small dataset: A case of 2019-ncov novel coronavirus outbreak, 2020, arXiv preprint arXiv:2003.10776.
[21] G. Baltas, F.A. Prieto Rodríguez, M. Frantzi, C. García Alonso, P. Rodríguez Cortés, et al., Monte Carlo Deep Neural Network Model for Spread and Peak Prediction of COVID-19, Loyola Tech, 2020.
[22] D. Khatua, A. De, S. Kar, E. Samanta, A.A. Seikh, D. Guha, A fuzzy dynamic optimal model for COVID-19 epidemic in India based on granular differentiability, 2020, Available at SSRN 3621640.
[23] P. Liu, P. Beeler, R.K. Chakrabarty, COVID-19 Progression timeline and effectiveness of response-to-spread interventions across the United States, medRxiv (2020).
[24] M.C. Traini, C. Caponi, G.V. De Socio, Modelling the epidemic 2019-nCoV event in Italy: a preliminary note, medRxiv (2020).
[25] S. Lai, I.I. Bogoch, N.W. Ruktanonchai, A. Watts, X. Lu, W. Yang, H. Yu, K. Khan, A.J. Tatem, Assessing spread risk of Wuhan novel coronavirus within and beyond China, January-April 2020: a travel network-based modelling study, medRxiv (2020).
[26] L. Wynants, B. Van Calster, M.M. Bonten, G.S. Collins, T.P. Debray, M. De Vos, M.C. Haller, G. Heinze, K.G. Moons, R.D. Riley, et al., Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal, bmj 369 (2020).
[27] C.T. Bauch, J.O. Lloyd-Smith, M.P. Coffee, A.P. Galvani, Dynamically modeling SARS and other newly emerging respiratory illnesses: past, present, and future, Epidemiology (2005) 791–801.
[28] G.R. Shinde, A.B. Kalamkar, P.N. Mahalle, N. Dey, J. Chaki, A.E. Hassanien, Forecasting models for coronavirus disease (COVID-19): A survey of the state-of-the-art, SN Comput. Sci. 1 (4) (2020) 1–15.
[29] B.M. Althouse, J. Lessler, A.A. Sall, M. Diallo, K.A. Hanley, D.M. Watts, S.C. Weaver, D.A. Cummings, Synchrony of sylvatic dengue isolations: a multi-host, multi-vector sir model of dengue virus transmission in Senegal, PLoS Negl. Trop. Dis. 6 (11) (2012) e1928.
[30] R.M. Anderson, R.M. May, Infectious Diseases of Humans: Dynamics and Control, Oxford university press, 1992.
[31] H.W. Hethcote, Asymptotic behavior in a deterministic epidemic model, Bull. Math. Biol. 35 (1973) 607–614.
[32] H. Behncke, Optimal control of deterministic epidemics, Opt. Control Appl. Methods 21 (6) (2000) 269–285.
[33] S. Bhattacharya, K. Gaurav, S. Ghosh, Viral marketing on social networks: An epidemiological perspective, Physica A 525 (2019) 478–490.
[34] Y. Liu, A.A. Gayle, A. Wilder-Smith, J. Rocklöv, The reproductive number of COVID-19 is higher compared to SARS coronavirus, J. Travel Med. (2020).
[35] E. Shim, A. Tariq, W. Choi, Y. Lee, G. Chowell, Transmission potential and severity of COVID-19 in South Korea, Int. J. Infect. Dis. (2020).
[36] A.J. Kucharski, T.W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R.M. Eggo, F. Sun, M. Jit, J.D. Munday, et al., Early dynamics of transmission and control of COVID-19: a mathematical modelling study, Lancet Infect. Dis. (2020).
[37] L. Peng, W. Yang, D. Zhang, C. Zhuge, L. Hong, Epidemic analysis of COVID-19 in China by dynamical modeling, 2020, arXiv preprint arXiv:2002.06563.
[38] W.O. Kermack, A.G. McKendrick, A contribution to the mathematical theory of epidemics, Proc. R. Soc. Lond. Ser. A 115 (772) (1927) 700–721.
[39] A. Rachah, D.F. Torres, Analysis, simulation and optimal control of a SEIR model for ebola virus with demographic effects, 2017, arXiv preprint arXiv:1705.01079.
[40] T. Berge, J.-S. Lubuma, G. Moremedi, N. Morris, R. Kondera-Shava, A simple mathematical model for Ebola in Africa, J. Biol. Dyn. 11 (1) (2017) 42–74.
[41] T. Toffoli, N. Margolus, Cellular Automata Machines: A New Environment for Modeling, MIT press, 1987.
[42] S. Wolfram, Cellular Automata and Complexity: Collected Papers, CRC Press, 2018.
[43] N. Boccara, K. Cheong, M. Oram, A probabilistic automata network epidemic model with births and deaths exhibiting cyclic behaviour, J. Phys. A: Math. Gen. 27 (5) (1994) 1585.
[44] C. Beauchemin, J. Samuel, J. Tuszynski, A simple cellular automaton model for influenza a viral infections, J. Theoret. Biol. 232 (2) (2005) 223–234.
[45] H. Fuks, A.T. Lawniczak, Individual-based lattice model for spatial spread of epidemics, Discrete Dyn. Nat. Soc. 6 (2001).
[46] R. Willox, B. Grammaticos, A. Carstea, A. Ramani, Epidemic dynamics: discrete-time and cellular automaton models, Physica A 328 (1–2) (2003) 13–22.
[47] P. Eosina, T. Djatna, H. Khusun, A cellular automata modeling for visualizing and predicting spreading patterns of dengue fever, Telkomnika 14 (1) (2016) 228.
[48] K.S. Pokkuluri, S.U.D. Nedunuri, A novel cellular automata classifier for COVID-19 prediction, J. Health Sci. 10 (1) (2020) 34–38.
[49] M. Dascalu, M. Malita, A. Barbilian, E. Franti, G.M. Stefan, Enhanced cellular automata with autonomous agents for Covid-19 pandemic modeling, Romanian J. Inf. Sci. Technol. 23 (2020) S15–S27.
[50] S. Ghosh, S. Bhattacharya, Computational model on COVID-19 pandemic using probabilistic cellular automata, 2020, arXiv preprint arXiv:2006.11270.
[51] A.H. Wright, Genetic algorithms for real parameter optimization, in: Foundations of Genetic Algorithms, vol. 1, Elsevier, 1991, pp. 205–218.
[52] L. Yao, W.A. Sethares, Nonlinear parameter estimation via the genetic algorithm, IEEE Trans. Signal Process. 42 (4) (1994) 927–935.
[53] S. Katare, A. Bhan, J.M. Caruthers, W.N. Delgass, V. Venkatasubramanian, A hybrid genetic algorithm for efficient parameter estimation of large kinetic models, Comput. Chem. Eng. 28 (12) (2004) 2569–2581.
[54] M. Gulsen, A. Smith, D. Tate, A genetic algorithm approach to curve fitting, Int. J. Prod. Res. 33 (7) (1995) 1911–1923.
[55] C.L. Karr, B. Weck, D.-L. Massart, P. Vankeerberghen, Least median squares curve fitting using a genetic algorithm, Eng. Appl. Artif. Intell. 8 (2) (1995) 177–189.
[56] P.H. Schimit, Evolutionary aspects of spatial prisoner's dilemma in a population modeled by continuous probabilistic cellular automata and genetic algorithm, Appl. Math. Comput. 290 (2016) 178–188.
[57] J.H. Holland, et al., Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, MIT press, 1992.
[58] T. Liao, Clustering of time series data—a survey, Pattern Recognit. 38 (11) (2005) 1857–1874.
[59] J. Gao, H. Sultan, J. Hu, W.-W. Tung, Denoising nonlinear time series by adaptive filtering and wavelet shrinkage: a comparison, IEEE Signal Process. Lett. 17 (3) (2009) 237–240.
[60] O. Salem, Y. Liu, A. Mehaoua, Anomaly detection in medical wsns using enclosing ellipse and chi-square distance, in: 2014 IEEE International Conference on Communications (ICC), IEEE, 2014, pp. 3658–3663.
[61] World Health Organization Coronavirus Disease (COVID-2019) Situation Reports, 2020, Available at URL: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200402-sitrep-73-covid-19.pdf. (Accessed June 2020).
[62] D. Acharjya, A. Anitha, A comparative study of statistical and rough computing models in predictive data analysis, Int. J. Amb. Comput. Intell. (IJACI) 8 (2) (2017) 32–51.
[63] J. Su, H. Zhang, A fast decision tree learning algorithm, in: AAAI, vol. 6, 2006, pp. 500–505.
[64] A. Miller, M.J. Reandelar, K. Fasciglione, V. Roumenova, Y. Li, G.H. Otazu, Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study, MedRxiv (2020).