Improving epidemic testing and containment strategies using machine learning

MACHINE
LEARNING
Science and Technology

CrossMark

OPEN ACCESS

RECEIVED
4 December 2020

REVISED
10 March 2021

ACCEPTED FOR PUBLICATION
22 March 2021

PUBLISHED
12 May 2021

PAPER

Improving epidemic testing and containment strategies using machine learning

Laura Natali[1], Saga Helgadottir[1], Onofrio M Maragò[2] and Giovanni Volpe[1,*]

[1] Department of Physics, University of Gothenburg, Gothenburg SE-41296, Sweden
[2] CNR-IPCF, Istituto per i Processi Chimico-Fisici, I-98158 Messina, Italy
[*] Author to whom any correspondence should be addressed.

E-mail: giovanni.volpe@physics.gu.se

Keywords: neural networks, SIR model, disease spreading

Supplementary material for this article is available online

Abstract

Containment of epidemic outbreaks entails great societal and economic costs. Cost-effective containment strategies rely on efficiently identifying infected individuals, making the best possible use of the available testing resources. Therefore, quickly identifying the optimal testing strategy is of critical importance. Here, we demonstrate that machine learning can be used to identify which individuals are most beneficial to test, automatically and dynamically adapting the testing strategy to the characteristics of the disease outbreak. Specifically, we simulate an outbreak using the archetypal susceptible-infectious-recovered (SIR) model and we use data about the first confirmed cases to train a neural network that learns to make predictions about the rest of the population. Using these predictions, we manage to contain the outbreak more effectively and more quickly than with standard approaches. Furthermore, we demonstrate how this method can be used also when there is a possibility of reinfection (SIRS model) to efficiently eradicate an endemic disease.

Compartmental epidemiological models provide a simple and powerful mathematical framework to capture the main features of a disease outbreak in a population [1, 2]. They consider how a disease spreads in a finite population of individuals over a time interval. The individuals are compartmentalized into categories based on their epidemiological condition. The first such model, known as the susceptible-infectious-recovered (SIR) model, was proposed in 1927 by Kermack and McKendrick [3] and is still widely employed today [4]. In the SIR model, there are three categories: susceptible individuals that have never been infected; infectious individuals that are currently infected; and recovered individuals that have previously been infected and are now immunized against the disease. Initially, all individuals are susceptible except for a limited group of infectious individuals, who seed the disease.

In the event of a disease outbreak, it is often desirable to attempt to contain or eradicate it. Different factors influence how effective a containment strategy is, including the characteristics of the disease and of the population [5, 6]. However, these characteristics are often difficult to measure or model precisely, especially for novel diseases during their first outbreaks [6–13]. The World Health Organization provides some general guidelines for strategies to prevent disease spread [14], which include travel restrictions, social distancing, and enforced quarantine. In particular, the isolation of potentially infected individuals is often the most effective measure to limit the spread of the infection. The safest approach would be to isolate and quarantine all individuals regardless of their epidemiological condition. However, this cannot be implemented and maintained on a large scale for a prolonged period because of its societal and economic deleterious effects [15].

In order to implement efficient, cost-effective strategies to contain an outbreak, it is therefore critical to promptly identify infectious individuals. The most straightforward approach would be to test all the individuals and immediately identify and isolate/treat the infectious ones [16]. In a real-life large-scale epidemic, however, extensive testing is not usually feasible because of economic and logistic constraints

[17–19]. Therefore, the containment of the disease requires interventions also on individuals who have not been tested yet, which again entails societal and economic costs [20].

Here, we demonstrate that machine learning can be used to identify an optimized test strategy, i.e. which are the individuals that is most beneficial to test. Specifically, we introduce a neural-network-powered strategy [21, 22] for testing and isolating individuals, even though the parameters of the model are not known and infectious individuals can be asymptomatic. The neural network (NN) informs the decision on which individuals should be tested and isolated. Modelling a disease outbreak using the SIR model [3, 4], we demonstrate that, for an equal number of quarantined individuals, the neural-network-informed strategy manages to contain the disease outbreak more effectively than alternative standard contact-tracing strategies, while autonomously and dynamically adapting to the specifics of the outbreak using only the information about the first confirmed cases. Furthermore, since for many diseases immunity is not lasting, we also demonstrate how the neural-network-informed approach can be used to efficiently prevent a new disease from becoming endemic when there is a possibility of reinfection (SIRS model). We envision that similar methods can be employed in public health to control epidemic outbreaks and to eradicate endemic diseases.

1. Results

1.1. Epidemic outbreak model and containment strategies

We model an epidemic outbreak using an agent-based SIR model [1, 23] (see details in section 3.2), where the population consists of $N = 10^5$ individuals distributed uniformly on a square lattice with 320 × 320 cells, resulting in an average density of 0.98. The individuals move as random walkers on the lattice [24, 25] being each confined to a region with an average radius of $r = 10$ cells [26]. All their positions are updated simultaneously at each time step. Each individual always belongs to one of the SIR categories (figure 1(a)). At the beginning of the simulation, 50 individuals (0.05% of total population) are randomly selected and made infectious (I). The rest of the population, instead, is initialized as susceptible (S). The disease is transmitted with probability β when susceptible and infected individuals are occupying the same cell, to mimic the short-range interactions necessary for disease spreading. An infected individual has a probability γ of recovering in each time step, after which it becomes immunized against the disease. We choose the values of β and γ to have a stochastic evolution with basic reproductive number in the range of those observed for typical viral diseases such as influenza [27, 28] or Covid-19 [29, 30] (see supplementary note 1 and supplementary figure S1 (available online at stacks.iop.org/MLST/2/035007/mmedia)). Each individual is also characterized by a 'temperature', which slightly increases as the disease develops; the temperature is normally distributed and corresponds to 36.8 ± 1.0 for healthy (i.e. susceptible and recovered) individuals, and to 37.4 ± 1.2 for infectious individuals (figure 1(b)), so that there is a significant overlap between the two distributions and, thus, some individuals can be 'asymptomatic'. We let the model evolve for 150 time steps, which can be thought of as the days of an epidemic outbreak that lasts approximately six months, but can easily be rescaled to fit another time scale.

Figure 1(c) provides an example of the *free evolution* of the outbreak in the absence of any containment measures. By $t = 20$, the disease has spread from the initial infectious individuals creating a few hotspots. These hotspots steadily grow ($t = 50$) until most of the population has been infected ($t = 100$) and the outbreak starts to subside. Figure 1(d) shows how the fraction of individuals in each category varies over time: as the disease spreads, the number of susceptible individuals steadily decreases and the number of recovered ones increases, while the number of infectious individuals initially grows and then slowly decreases until the outbreak ends because essentially the whole population is immunized.

The spread of the disease can be controlled by enacting containment measures. For example, figures 1(e) and (f) show the evolution of the outbreak when potentially infectious individuals are isolated based on standard *contact tracing* [18, 19, 31, 32] (see details in section 3.3). At each time step, a fixed number of tests ($N_{\mathrm{test}} = 100 \ll N$) are performed to assess whether individuals are infectious. The value of N_{test} is set low enough to simulate a limited access to testing so that only a small portion of the population can be tested (15% in 150 time steps). The individuals to be tested are selected randomly from the susceptible individuals with the highest temperature, i.e. those that show more clear symptoms. Selecting the individuals to be tested in this way presents two advantages compared to a purely random testing strategy: it avoids a slow start (with an initial probability of success around 1/2000), and it is more representative of reality (where symptomatic cases first indicate an outbreak). For simplicity, we assume that the test never fails and that there is no delay between performing the test and receiving the result. However, we remark that the task of identifying the infectious individuals is made harder by the fact that some of their temperatures are in the healthy range (figure 1(b)), making them asymptomatic. The individuals who test positive are quarantined: from that time step on, they neither move nor interact with the rest of the population. For the tested individuals, the

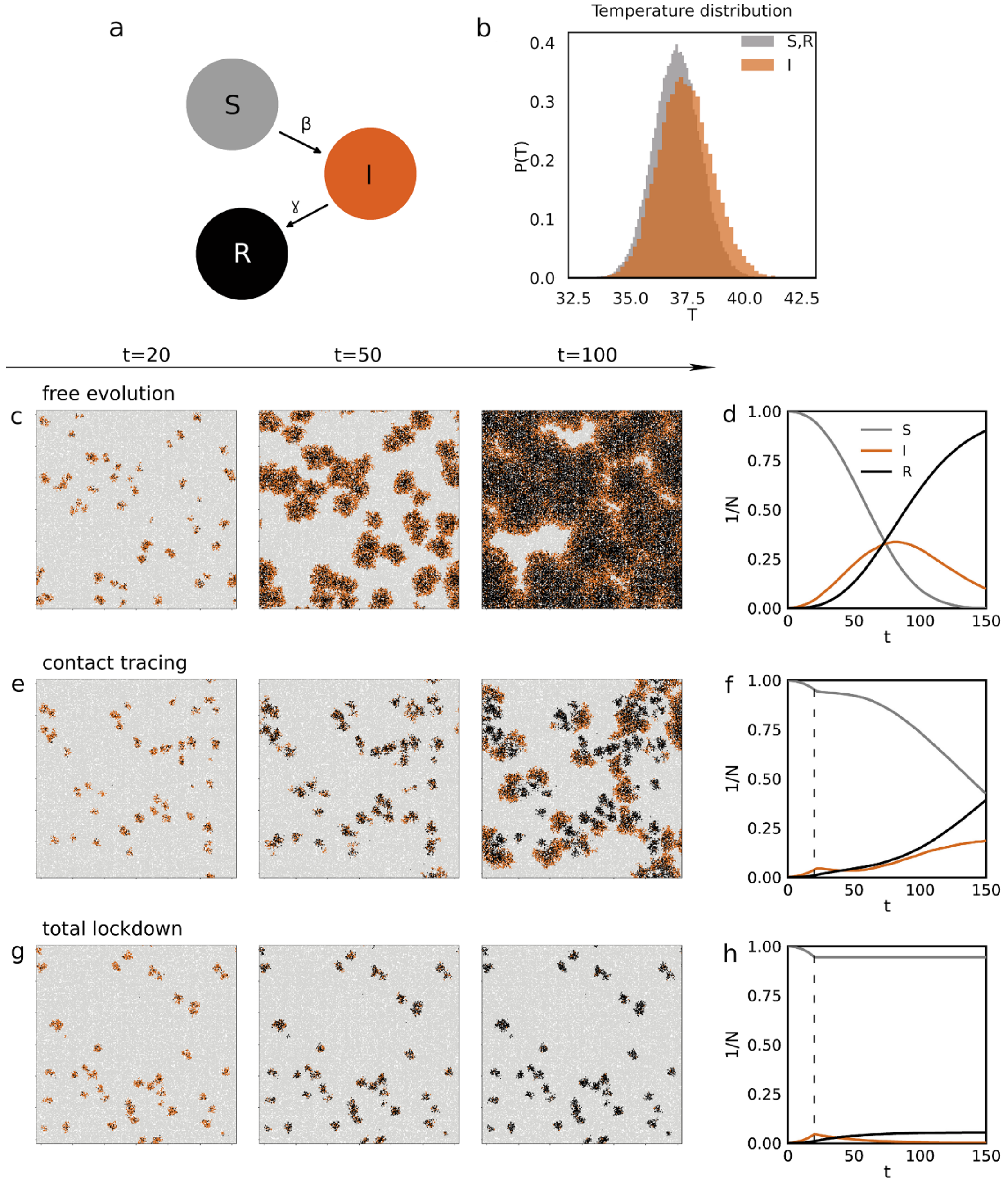


Figure 1. SIR model and containment strategies. (a) We consider a population of 10^5 individuals moving on a square lattice (320 × 320 cells). Each individual can be either susceptible (S, grey), infectious (I, orange), or recovered (R, black). At each time step, the susceptible individuals become infectious with probability β when they occupy the same cell as an infectious individual, and the infectious individuals recover with probability γ becoming immunized against the disease. (b) Temperature of individuals that are healthy (S and R, 36.8 ± 1.0) and infectious (I, 37.4 ± 1.2); note the range of asymptomatic individuals, i.e. infectious individuals with temperature in a healthy range. (c) Disease spread at times $t = 20, 50, 100$ in the absence of any containment measures and (d) corresponding fraction of the population in each category; unchecked, the disease spreads to almost all the population. (e) Disease spread using standard contact tracing to isolate potentially infectious individuals starting at $t = 20$ (dashed vertical line) and (f) corresponding fraction of the population in each category; the disease spreads more slowly than in (a), but is not contained. (g) Disease spread when a total lockdown is implemented at $t = 20$ (dashed vertical line) preventing any further spread of the disease and (h) corresponding fraction of the population in each category. See also supplementary video 1.

isolation is temporary, so the system knows when they stop being infectious and can safely return to interact with the rest of the population.

Due to the limited number of tests, quarantining only the individuals that test positive is not enough to contain the outbreak. It is therefore necessary to use contact tracing to isolate also individuals who have not been tested. (While testing starts from the first time step, contact tracing and isolation of individuals starts only at $t = 20$.) For all detected infectious individuals, we trace back their previous contacts up to 50 time steps in the past. Within this group of individuals that interacted with confirmed cases, we test those with the highest temperature. We rank the other individuals according to their number of contacts with infectious individuals, and, given the same number of contacts, according to their current temperature. We isolate a number of individuals until reaching a predetermined fraction of the population (here, 25%) (see details in section 3.3).

It is interesting to compare the free evolution of the outbreak (figures 1(c) and (d)) with the case with isolation based on contact tracing (figures 1(e) and (f)). While at $t = 20$ both outbreaks are similar, the containment measures take hold almost immediately, significantly reducing the size of the outbreaks and the fraction of individuals that are infectious at the same time. The epidemic outbreak remains confined to a few areas reaching only a part of the population (figure 1(e)) and the curve of infected individuals is flatter (figure 1(f)). We remark that, despite its success in slowing down the spread rate of the disease, also the strategy relying on isolation of potentially infectious individuals identified by contact tracing does not lead to a complete suppression of the outbreak, as can be seen from the fact that nearly 20% of the population is infectious still at $t = 150$.

Complete eradication of the disease is in principle possible by adopting an unrealistic *total lockdown*, where the whole population is quarantined simultaneously (figures 1(g) and (h)). From $t = 20$, all individuals are isolated so that they cannot move or interact. Figure 1(g) shows how this leads to an almost immediate containment of the disease hotspots. More interestingly, figure 1(h) shows how the fraction of infectious individuals quickly drops down and, unlike for the free evolution (figures 1(c) and (d)) and the contact-tracing isolation (figures 1(e) and (f)), reaches zero by $t = 120$, so that the disease is extinguished by the end of the simulation.

Different containment strategies lead to different evaluations of the outbreak. The free evolution (figures 1(c) and (d)) and the total lockdown (figures 1(g) and (h)) approach represent the two limiting policies, leading to the least and the most effective containment. The contact-tracing isolation (figures 1(e) and (f)) achieves an intermediate level of containment, but does not achieve eradication of the disease, despite isolating up to about 25% of the population.

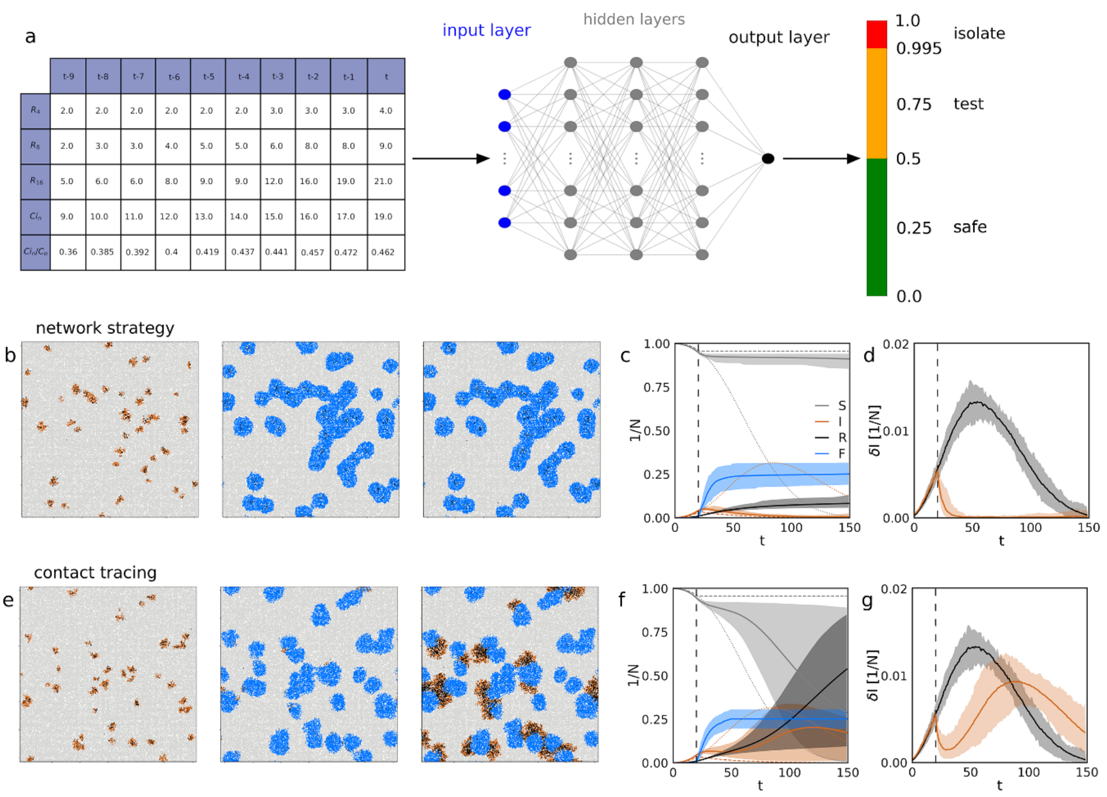
1.2. Neural-network-informed testing

It would be desirable to achieve disease eradication as in the total-lockdown strategy (see figures 1(g) and (h)), but isolating only part of the population as in the contact-tracing strategy (see figures 1(e) and (f)). To achieve this, we propose a strategy that employs a NN to inform which individuals to test and isolate.

The schematic of the NN we employ is shown in figure 2(a) (see details in section 3.4). In general, a NN receives some inputs, elaborates them through of a series of hidden layers of artificial neurons, and returns an output [33]. In our case, the input consists of contact-tracing information for a given individual n for the last 10 time steps. Specifically, we provide the NN with five time series: $R_{4,n}(t)$, $R_{8,n}(t)$, $R_{16,n}(t)$, $C_n^{\mathrm{i}}(t)$, and $C_n^{\mathrm{i}}/C_n^{\mathrm{tot}}(t)$. The first three indicate the number of tested infectious individuals within a distance $r = 4$, 8, and 16 cells from the considered one. $C_n^{\mathrm{tot}}(t)$ is the total number of contacts (i.e. defined as individuals occupying the same cell at the same time) and $C_n^{\mathrm{i}}(t)$ is the number of contacts with confirmed infectious individuals. Then, the NN elaborates this information through three dense layers of artificial neurons. Finally, the NN outputs a value p, representing the risk of being infectious at the current time step, between 0 for a putatively healthy individual and 1 for a putatively infectious individual. Individuals with $p > 0.995$ are immediately isolated, while individuals with $p \in [0.5, 0.995]$ are slated to be tested, starting from the individuals with the highest temperatures until the depletion of all available tests. In this way, we manage to freeze the infectious individuals that are easy to identify, while optimizing the deployment of the available tests: we use the tests principally to achieve a better understanding of the extent and distribution of the disease.

NNs are supervised machine learning methods and, therefore, require training [33]. In general, the training of a NN is performed by providing the NNs with a series of inputs and corresponding known outputs [33]. In our case, we can only use for training individuals that have already been tested within each run of the simulation (see details in section 3.4). Therefore, we start training at $t = 20$, when we have tested 2000 individuals. In subsequent time steps, the size and accuracy of the training data set increases with the number of performed tests, so we repeatedly retrain the NN to improve its performance. This leads to a positive feedback loop, where a better-trained NN selects more efficiently individuals for testing, which in turn provides better insights into the disease distribution, which finally improves the training data set available to further improve the performance of the NN.

Figure 2(b) depicts the snapshots of the system at $t = 20, 50, 100$. The colour code is the same as that used in figure 1, with the addition of frozen individuals (F) indicated in light blue. Until $t = 20$, the outbreak evolves freely, analogously to figure 1(c), while enough data are accumulated to train the NN. From $t = 20$ and onward, the neural-network predictions are used to inform which individuals to isolate and test. By $t = 50$, all outbreaks have been identified and surrounded by frozen individuals. Subsequently ($t = 100$), the outbreaks remain under control and are prevented from spreading, in stark contrast with the wide spread of the disease in free evolution ($t = 100$ in figure 1(c)).



	t-9	t-8	t-7	t-6	t-5	t-4	t-3	t-2	t-1	t
R_4	2.0	2.0	2.0	2.0	2.0	2.0	3.0	3.0	3.0	4.0
R_8	2.0	3.0	3.0	4.0	5.0	5.0	6.0	8.0	8.0	9.0
R_{16}	5.0	6.0	6.0	8.0	9.0	9.0	12.0	16.0	19.0	21.0
Ci_n	9.0	10.0	11.0	12.0	13.0	14.0	15.0	16.0	17.0	19.0
Ci_n/C_n	0.36	0.385	0.392	0.4	0.419	0.437	0.441	0.457	0.472	0.462

Figure 2. Improved outbreak containment using neural-network-informed testing. (a) Structure of the NN. The inputs (table on the left) are $R_{4,n}(t)$, $R_{8,n}(t)$, $R_{16,n}(t)$, $C_n^{\mathrm{i}}(t)$, and $C_n^{\mathrm{i}}(t)/C_n^{\mathrm{tot}}(t)$, where $R_{r,n}$ is the number of confirmed infectious individuals in a radius r from the individual n, C_n^{i} is the number of contacts that individual n has had with confirmed infectious individuals, and C_n^{tot} the total number of contacts for individual n; for each parameter, the input includes the history during 10 time steps ($[t-9,t]$). The NN analyses these inputs through three dense layers and outputs a value p from 0 (individual predicted to be healthy) to 1 (individual predicted to be infectious): individuals with $p > 0.995$ are directly quarantined, and individuals with $p \in [0.5, 0.995]$ are tested starting from individuals with the highest temperatures until the depletion of the available tests. (b)–(d) Disease evolution when the testing and isolation strategy is determined based on the output from a NN: (b) Snapshots of susceptibles (S, grey), infectious (I, orange), recovered (R, black), and frozen (F, blue) individuals at time steps $t = 20, 50, 100$. (c) Corresponding fraction of the population in each category compared with the two limiting cases of free evolution (dotted orange line, see also figures 1(c) and (d)) and full lockdown (dashed orange line, see also figures 1(g) and (h)). The isolation of individuals starts at $t = 20$ (dashed black line). The solid lines indicate the average over $N_{\mathrm{runs}} = 100$, while the shaded areas correspond to 90% confidence interval. (d) The number of new cases δI for the neural-network-informed testing (orange line) compared to the free evolution (black line). (e)–(g) Comparison to a standard contact-tracing strategy (see figures 2(e) and (f)) where the same number of individuals are quarantined as in (b)–(d) where we employ the neural-network-informed strategy: the number of infectious individuals and the spread of the disease are greatly reduced when employing the neural-network-informed strategy. See also supplementary video 2.

The orange solid line in figure 2(c) shows the fraction of the population that is infectious as a function of time. Shortly after we switch on the NN ($t = 20$), the infectious fraction reaches its maximum (5.1% at $t = 26$) and subsequently rapidly decreases to zero. Correspondingly, the number of recovered (black solid line) and susceptible (grey solid line) individuals reach a plateau. In particular, the fraction of individuals that are infected and eventually recover is $8 \pm 4\%$.

The number of frozen individuals is initially zero and quickly increases in the first stages of neural-network-informed testing, eventually reaching the set value of 25% of the total population. We can compare the curve of the infectious individuals using the neural-network-informed testing and isolation (orange solid line) with the limiting cases of free evolution (orange dotted line, see figure 1(c)) and of total lockdown (orange dashed line, see figure 1(g)). By isolating only 25% of the population, the neural-network-informed strategy achieves a containment of the epidemic similar to that achieved by the full lockdown.

Figure 2(d) represents the fraction of new infectious individuals per time step for the neural-network-informed strategy (orange line) and for the free evolution of the epidemics (black line). The free-evolution curve reaches a maximum at $t = 59$ corresponding to $\delta I(59) = 1.4 \pm 0.2\%$. The curve for the neural-network-informed strategy starts decreasing immediately after isolation starts at $t = 20$, corresponding to a peak value $\delta I(20) = 0.55 \pm 0.08\%$, and stably reaches zero around $t = 50$.

Figures 2(e)–(g) provide comparisons with a standard contact-tracing strategy, where the same number of individuals are tested and isolated as described in detail in the previous section. Figure 2(e) shows

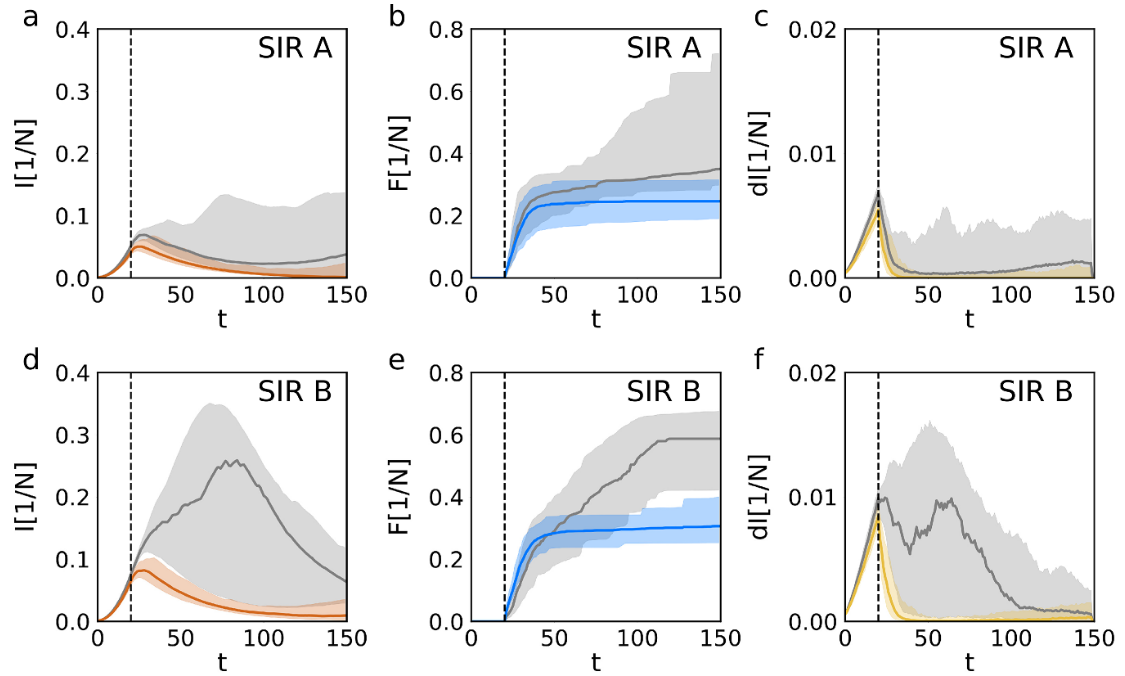


Figure 3. Automatic and dynamic adaptation of the neural network to the underlying SIR parameters. Two independent NNs (NN_A and NN_B) are trained for two SIR models (SIR_A and SIR_B, respectively) with different parameters ($\beta_A = 0.6$ and $\gamma_A = 0.03$ for SIR_A, and $\beta_B = 0.8$ and $\gamma_B = 0.03$ for SIR_B). SIR_A and NN_A are the same as those employed in figures 2(b)–(d). (a) Infectious individuals (orange line), (b) frozen individuals (blue line), and (c) new infections (yellow line) when NN_A is used on SIR_A. The grey lines are the corresponding curves when NN_B is used instead, showing a clear decrease in performance. (d)–(f) Corresponding plots where NN_B (coloured lines) and NN_A (grey lines) are used on SIR_B. Overall, these results show that the NN gets automatically optimized for the parameters of the underlying outbreak. In all cases, the shaded areas represent the 90% confidence intervals obtained from 100 simulations. See also supplementary video 3.

snapshots of the system at $t = 20, 50, 100$: starting from the same number of hotspots ($t = 20$), contact tracing manages to identify all regions reached by the disease ($t = 50$), but the disease can still spread due to the limited number of individuals that can be isolated ($t = 100$). Figure 2(f) shows that, differently from the case of the neural-network-informed strategy (figure 2(c)), the increase of the fraction of infected individuals slows down for some time steps, but then starts again to grow reaching a peak at $t = 120$ corresponding to about 20% of the total population. The total number that have been infected at the end of the simulation (i.e. all infectious and recovered individuals at $t = 150$) is strikingly lower for the neural-network-informed strategy (6%–14%) than for the contact-tracing-based strategy (30%–89%). The wide shaded area in figure 2(f) is nearly 7 times larger than in figure 2(c), showing that the contact tracing is less stable against different evolution patterns of an epidemic with same underlying SIR parameters. The orange line in figure 2(g) shows the fraction of new infectious individuals δI as a function of time, which is non-zero at the end of the simulation, unlike for the neural-network-informed strategy (orange line in figure 2(d)). We present a quantitative comparison of the performance of the neural-network-informed strategy and this contact-tracing strategy in supplementary note 2 and supplementary figure S2. Additionally, we have compared the neural-network-informed strategy with alternative contact-tracing strategies in supplementary note 3 and supplementary figure S3. We can therefore conclude that contact tracing is less effective than the NN for the same number of frozen individuals.

1.3. Automatic and dynamic adaptation to the outbreak characteristics

An important characteristic of the neural-network-informed strategy is that it can automatically and dynamically adapt itself to the underlying characteristics of the outbreak. In our model, this means that the NN does not need to have explicit knowledge of the underlying SIR model. More generally, the NN can adapt to other kinds of outbreaks and also take into account the effects of the containment measure put in place.

Figure 3 demonstrates the ability of the neural-network-informed strategy to automatically and dynamically adapt itself to the underlying characteristics of the outbreak. The coloured solid lines in figures 3(a)–(c) reproduce the performance of the strategy presented in figures 2(b)–(d), which is informed by NN_A trained on the data obtained from an outbreak (SIR_A, $\beta_A = 0.6$ and $\gamma_A = 0.03$), in terms of the evolution of infectious individuals (orange line, figure 3(a)), frozen individuals (blue line, figure 3(b)), and new infections in each timestep (yellow line, figure 3(c)). We then apply NN_B, i.e. another NN trained on a

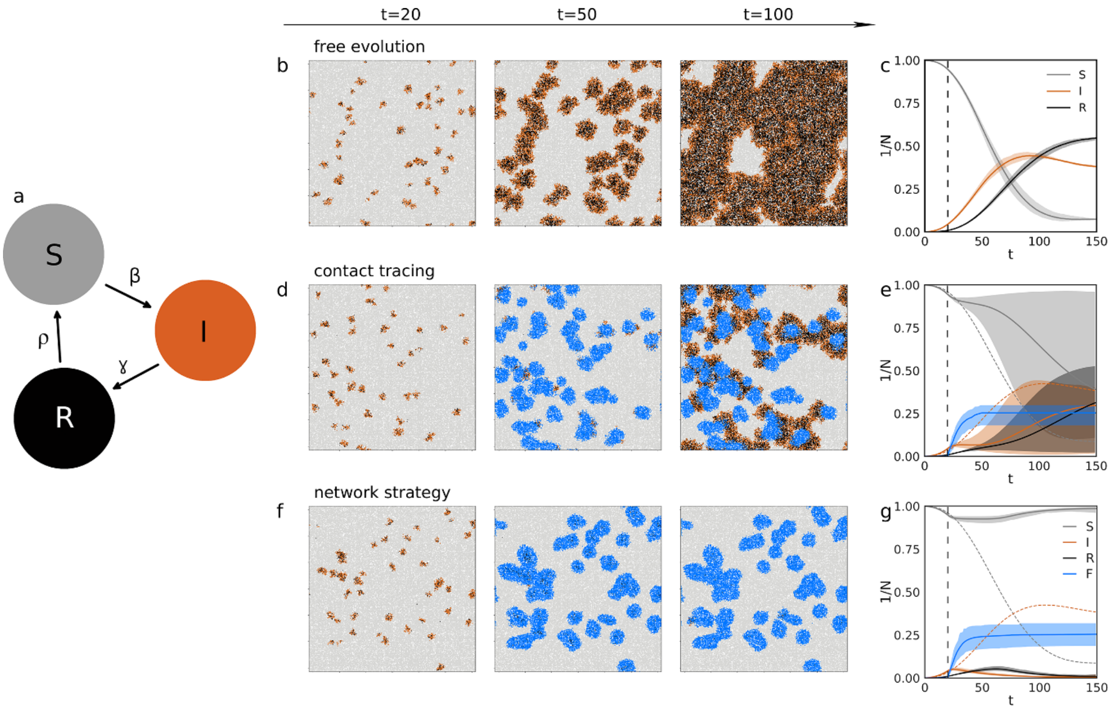


Figure 4. SIR model with temporary immunity (SIRS). (a) We consider a model with possibility of reinfection (SIRS model), where at each time step, recovered individuals have a probability ρ of becoming susceptible again (specifically, we use $\beta = 0.06$, $\gamma = 0.03$, and $\rho = 0.02$). Without countermeasures, this leads to an endemization of the disease. (b) Disease spread at $t = 20, 50, 100$ in the absence of any containment measures and (c) corresponding fraction of the population in the susceptible (S, gray line), infectious (I, orange line) and recovered (R, black line) categories; the endemization can be inferred from the stabilization of the fraction of infectious individuals towards the end of the simulation. (d) Disease spread using standard contact tracing to isolate potentially infectious individuals starting at $t = 20$ (dashed vertical line in (f)) and (e) corresponding fraction of the population in each category, including frozen individuals (F, blue line); while the disease spreads less, it still becomes endemic. (f) Disease spread when employing a NN to inform testing starting at $t = 20$ (dashed vertical line in (g)) and (g) corresponding fraction of the population in each category; in this case, the disease is completely eradicated. The dashed lines in (e) and (g) correspond to the free evolution of the disease and are reported from (c) for comparison. In all cases, the solid lines indicate the average over multiple runs ($N_{\text{runs}} = 100$), while the shaded areas correspond to 90% confidence intervals. See also supplementary video 4.

different outbreak whose underlying SIR model has a slightly different transmission rate (SIR$_B$, $\beta_B = 0.8$ and $\gamma_B = 0.03$). The resulting performance can be seen in the grey lines in figures 3(a)–(c). While overall NN$_B$ manages to improve the outbreak with underlying SIR$_A$ model compared to its free evolution, it performs much worse that NN$_A$. At the end of the simulation in figure 3(a), the fraction of infectious individuals is still in the range (0.12%–13.7%) of the population for the grey confidence bands, while the overall fraction of individuals in isolation is in the range (30%–72%), as shown in figures 3(b). This suggests that, thanks to its training using the information acquired by the testing during the first 20 time steps, the neural-network-informed strategy gets fine-tuned to the specific characteristics of the underlying outbreak.

We further validate the fine-tuning of the NN by training NN$_B$ on the testing data obtained from the outbreak with underlying model SIR$_B$. The coloured lines in figures 3(d)–(f) show the results of applying NN$_B$ on the SIR$_B$ outbreak, which demonstrate a good containment of the outbreak. Instead, the grey lines show what happens when using NN$_A$, which leads to a much worse outcome. In this scenario, the peak for the curve of infected is around $t = 84$ and 25.7% against 8.1% of the population for the training performed on SIR$_B$. Figures 3(f) shows that δI oscillates between 540 and 995 new cases per time step in the interval $t \in [20, 73]$ before decreasing.

1.4. Disease eradication with possibility of reinfection

We now consider the case when the immunity against the disease is not permanent [34–36]. Thus, we consider a SIRS model (figure 4(a)), which is an extension of the SIR model where recovered individuals have a probability ρ at each time step to become again susceptible [34, 35] (see details in section 3.2). In the absence of any containment measures, the possibility of reinfection leads to an endemization of the disease. Figure 4(b) shows such free evolution of the disease: from the initial hotspots ($t = 20$), the disease spreads quickly to a large portion of the population ($t = 50$) until reaching a steady state. Figure 4(c) shows how the fraction of individuals in each category varies over time: during the initial spread of the disease, the number of susceptible individuals steadily decreases and the number of infectious ones increases; once the disease reaches its steady state, the fraction of infectious individuals stabilizes to a value that depends on the characteristics of the SIRS model, i.e. on the value of its parameters β, γ and ρ. Therefore, the disease becomes endemic [1].

Figures 4(d) and (e) show the development of the disease when a standard contact-tracing-based containment strategy is implemented, like that employed in figures 1(e) and (f). The solid lines represents the averages for susceptibles (S, grey), infectious (I, orange), recovered (R, black) and frozen (F, blue) individuals throughout the simulation. The colour bands, which denote the 90% confidence interval, is larger than those in figure 2(f); this implies that the performance of the contact-tracing strategy can vary significantly depending on the specific outbreak. It can be seen that this containment approach manages to reduce the number of infectious individuals in the steady state of the disease, but not to eradicate the disease itself.

Finally, figures 4(f) and (g) show the performance of the neural-network-informed strategy. We employ the same approach and NN architecture shown in figure 2(a) and the same strategy that we employed to contain the outbreaks in the SIR model shown in figures 2(b)–(d). Briefly, we start testing individuals from the beginning of the simulation accumulating data to train the NN. From $t = 20$, we start training the NN to predict infectious individuals and use this information to decide which individuals to isolate and test. The neural-network-informed strategy manages to eradicate the disease, as can be seen from the fact that the fraction of infectious individuals approaches zero by the end of the simulation (orange solid line in figure 4(g)), while the number of susceptible individuals increases as recovered individuals gradually lose their immunity. Therefore, by employing the neural-network-informed strategy, it is possible to prevent the initial outbreak from leading to the endemization of the disease.

2. Discussion

The current outbreak of the novel coronavirus disease (COVID-19) [7, 37–40] has dramatically brought to worldwide attention the crucial importance of epidemiological models for choosing the best strategies and policies to contain disease outbreaks [6, 7, 9, 13, 20]. Machine-learning approaches have been already proposed to help disease diagnosis [41] and epidemics handling [13]. In fact, in the last few years, various neural-network architectures have been employed to manage human diseases [42–45], such as malaria [46], and animal diseases, such as in swine flu [47]. In this work, we have now shown how a neural-network-informed strategy can improve the containment of an epidemic, even when only a small number of specific tests is available and some of the individuals are asymptomatic. This improvement can be seen in three key aspects. First, integrating the NN into the outbreak handling improves the performance of contact tracing, while performing the same number of tests and isolating the same fraction of individuals. Second, the NN autonomously tunes its weights to the ongoing outbreak, without needing to explicitly know its underlying model or its parameters, and therefore does not require *a priori* knowledge of the disease outbreak characteristics. Third, since the NN is regularly retrained as new data become available, it can automatically and dynamically adapt itself to the evolution of the outbreak as well as to the changes in the behaviour of the population, e.g. due to containment measures or different social habits. As a striking example, we have shown that, in the case of temporary immunization, the neural-network-informed strategy can prevent a disease outbreak from becoming endemic.

Even though we used a SIR model to describe the dynamics underlying the disease, the NN will automatically adapt itself to different underlying dynamics described by more complex epidemiological models, which might include, e.g. the disease incubation time [9], delays in the testing process [19], or even different patterns of movement of the individuals (e.g. periodic motion, and long-range travel) [10]. It is also possible to provide the NN with demographic information (e.g. individual risk factors, such as age, employment, and preexisting conditions) as well as with spatial information [48] (e.g. the location of the individuals, differentiating various places of aggregation, such us hospitals, markets, and schools), or even with simple-access medical tests (e.g. cough recordings [49]). For example, in order to construct the lattice information, one can label the individual data by the zip code of residence area to have anonymous spatially-resolved data. In this case, the structure of the lattice would be given by the zip codes. Another possibility could be to group individuals to have a coarse-graining according to family groups or neighbourhood spatial labelling. The key point is that each labelling structure will have its own specific neural-network-informed containment strategy after a first stage of training to adapt the testing strategy to the local characteristic and temporal evolution of the specific disease. Finally, the neural-network-informed approach presented in this work can be generalized to other situations, such as fire prevention [50] or econometrics [51].

3. Methods

3.1. SIR model

We divide the population of $N = 10^5$ identical individuals into three epidemiological categories: susceptible individuals S, infectious individuals I, and recovered individuals R, as in the original SIR model [3]. The individuals move on a square lattice with side $l = 320$ according to a stochastic model [24, 52]. The initial positions of the individuals on the square lattice are random and drawn from a uniform distribution. We store the latter values $\mathbf{x}_n(0) = [x_n(0), y_n(0)]$ throughout the simulation as the area of residence for each individual. The position of each individual $n \in [1, N]$ at each time step $t \in [0, 150]$ is given by its coordinates $\mathbf{x}_n(t) = [x_n(t), y_n(t)]$. Each individual is an independent random walker confined to move within a small area of the lattice centered around its initial random position $\mathbf{x}_n(0) = [x_n(0), y_n(0)]$. The position of each individual evolves as

$$\mathbf{x}_n(t+1) = \mathbf{x}_n(t) + \boldsymbol{\Delta}\mathbf{x}_n(t), \tag{1}$$

with displacements $\boldsymbol{\Delta}\mathbf{x}_n(t) = [\Delta x_n(t), \Delta y_n(t)]$ for each individual selected inside its Moore neighbourhood [53], given by

$$\Delta x_n = \begin{cases} -1 & \text{with probability } \dfrac{1}{3} + k[x_n(t) - x_n(0)] \\ 0 & \text{with probability } \dfrac{1}{3} \\ +1 & \text{with probability } \dfrac{1}{3} - k[x_n(t) - x_n(0)] \end{cases} \tag{2}$$

$$\Delta y_n = \begin{cases} -1 & \text{with probability } \dfrac{1}{3} + k[y_n(t) - y_n(0)] \\ 0 & \text{with probability } \dfrac{1}{3} \\ +1 & \text{with probability } \dfrac{1}{3} - k[y_n(t) - y_n(0)] \end{cases} \tag{3}$$

where $k = 0.04$ determines the radius $r_k \approx 10$ cells within which each individual moves. The positions of all individuals are updated synchronously and independently from each other.

The spread of the infection occurs because when a susceptible individual occupies the same cell as an infectious individual, it becomes infectious with probability β in each time step. The transmission applies only for the infectious individuals that are not frozen. Each infectious individual becomes recovered with probability γ at each time step. The parameters used are $\beta = 0.6$ and $\gamma = 0.03$, except for figure 3, where we also employ $\beta = 0.8$. The choice of the SIR parameters is motivated in order to have a stochastic evolution with basic reproductive number $R_0 \approx 3.3$ in the range of those observed for typical viral diseases such as influenza [27, 28] or Covid-19 [29, 30] (see supplementary note 1 and supplementary figure S1).

Each individual is also characterized by a 'temperature', which is normally distributed and corresponds to 36.8 ± 1.0 for healthy (i.e. susceptible and recovered) individuals, and to 37.4 ± 1.2 for infectious individuals, with a great overlap between the two distributions (figure 1(b)). The temperature characterizes the level of symptomaticity continuously, instead of adopting a binary division between asymptomatic and symptomatic cases [54–57]. The temperature T_n for each individual raises to $T_n^i = T_n + dT_n$, when the corresponding individual becomes infectious. In the case $dT_n \approx 0$, the individuals are asymptomatic, while they are symptomatic for $dT_n > 0$.

3.2. SIRS model

The SIRS is an alternative to the SIR model that assumes the immunization to the disease is temporary. Therefore, recovered individuals lose immunization and return susceptible with probability ρ in each time step. We employ $\rho = 0.02$.

3.3. Contact tracing

We present here the containment strategy based on contact tracing employed in this work, as opposed to the neural-network-informed one. In the supplementary note 3 and supplementary figure S3, we report other possible approaches, as alternative comparisons.

We keep track of individuals that occupy the same cell at a certain time step by introducing the contact matrix:

$$c_{nm}(t) = \delta(\mathbf{x}_n(t) - \mathbf{x}_m(t)), \tag{4}$$

where δ is the Kronecker delta, which has value 1 if the pair of individuals n and m occupy the same cell at time t, and 0 otherwise. Thus, the total number of contacts for individual n for the 50 time steps before time t is

$$C_n^{\mathrm{tot}}(t) = \sum_{\tau=t-50}^{t} \sum_{m\neq n} c_{nm}(\tau). \tag{5}$$

The number of contacts with confirmed infectious individuals is

$$C_n^{\mathrm{i}}(t) = \sum_{\tau=t-50}^{t} \sum_{m\neq n} c_{nm}(\tau)\delta_m^{\mathrm{i}}(t), \tag{6}$$

where $\delta_m^{\mathrm{i}}(t)$ is 1 if individual m has already been tested and found positive at time t, and 0 otherwise. When implementing the lockdown strategy based on contact tracing, we list the agents in descending order as a function of $C_n^{\mathrm{i}}(t)$, and we sort those with equal value based on their temperature. At each time step, we select for testing the first $N_{\mathrm{test}} = 100$ individuals in this list. We use the rest of such list for selecting individuals to freeze, whose number is set to match that of the neural-network-informed strategy. In this way, we can compare the two approaches using the same number of tests and the same number of frozen individuals. When the target number of individuals to isolate is larger than the individuals in the contact list (e.g. at the beginning of the simulation when the number of confirmed cases is small), we build an additional list from where to select the remaining individuals, which includes individuals that never had direct interactions with confirmed cases, but have been within a radius of 8 cells in the last 50 time steps; we sort also this additional list based on the temperature of the individuals.

3.4. Neural network

We employ a dense NN with three hidden layers with 16 neurons each and ReLU activation function [58, 59]. The output layer has one single neuron with a softmax activation function returning a value $p \in [0,1]$. Additionally, we use dropouts for the hidden layers as a way to avoid overfitting [60] (dropout rate 0.2, so that in each training epoch only 80% of the neurons is activated).

The input to the NN at time t includes $R_{4,n}(t)$, $R_{8,n}(t)$, $R_{16,n}(t)$, $C_n^{\mathrm{i}}(t)$, and $C_n^{\mathrm{i}}/C_n^{\mathrm{tot}}(t)$ for time steps $[t-9,t]$, where $C_n^{\mathrm{i}}(t)$ and $C_n^{\mathrm{tot}}(t)$ are the number of infectious and total contacts (equations (5) and 6), and $R_{r,n}(t)$ is the number of individuals that have tested positive within a radius r:

$$R_{r,n}(t) = \sum_i \delta\left(r - \parallel \mathbf{x}_n(t) - \mathbf{x}_i(t) \parallel\right), \tag{7}$$

where the summation is over all infected individuals.

The training of the NN is performed using information relative to the individuals that have already been tested (which is split between a training set and a validation set [61]). The loss function is the mean square error, we use the stochastic gradient descent method implemented in the Adam optimizer [62, 63], and the number of training epochs is fixed to 100 (see supplementary figure S1). While we use only two labels for the training (0 for susceptible individuals and 1 for infectious individuals), the trained network returns a prediction that is a continuous value $p \in [0,1]$.

Using the prediction of the network, we split the individuals that have not been tested yet into three groups: (1) $p > 0.995$: individuals with a high chance of being infectious, who are frozen without testing. (2) $0.5 < p < 0.995$: individuals with a medium chance of being infectious, amongst which the $N_{\mathrm{test}} = 100$ individuals with the highest temperature are tested. (3) $p < 0.5$: individuals with a low chance of infection.

We implement the NN using the Python libraries Tensorflow and Keras [64].

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

We thank Marj Tonini and Biagio Lucini for useful discussions. We acknowledge support from the MSCA-ITN-ETN project *ActiveMatter* sponsored by the European Commission (Horizon 2020, Project No. 812780).

ORCID iDs

Onofrio M Maragò https://orcid.org/0000-0002-7220-8527
Giovanni Volpe https://orcid.org/0000-0001-5057-1846

References

[1] Keeling M J and Rohani P 2011 *Modeling Infectious Diseases in Humans and Animals* (Princeton, NJ: Princeton University Press) (https://doi.org/10.2307/j.ctvcm4gk0)
[2] Anderson R 2013 Population and community biology series *The Population Dynamics of Infectious Diseases: Theory and Applications* (New York: Springer)
[3] Kermack W O and McKendrick A G 1927 A contribution to the mathematical theory of epidemics *Proc. R. Soc.* A **115** 115700–721
[4] Weiss H H 2013 The SIR model and the foundations of public health *Mater. Mat.* **3** 1–17
[5] Flaxman S *et al* 2020 Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe *Nature* **584** 257–61
[6] Maier B F and Brockmann D 2020 Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China *Science* **368** 742–6
[7] Bi Q *et al* 2020 Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study *Lancet Infectious Dis.* **20** 911–9
[8] Carletti T, Fanelli D and Piazza F 2020 COVID-19: The unreasonable effectiveness of simple models *Chaos Solitons Fractals* X **5** 100034
[9] Giordano G *et al* 2020 Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy *Nat. Med.* **26** 1–6
[10] Chinazzi M *et al* 2020 The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak *Science* **368** 395–400
[11] Perkins A *et al* 2020 Estimating unobserved SARS-CoV-2 infections in the United States *Proc. Natl Acad. Sci. USA* **117** 22597–602
[12] Bertozzi A L, Franco E, Mohler G, Short M B and Sledge D 2020 The challenges of modeling and forecasting the spread of COVID-19 (arXiv:2004.04741)
[13] Navascues M, Budroni C and Guryanova Y 2020 Disease control as an optimization problem (arXiv:2009.06576)
[14] Department of Communications, WHO Worldwide Strategic preparedness and response plan (available at: www.who.int/publications/i/item/strategic-preparedness-and-response-plan-for-the-new-coronavirus)
[15] Bonaccorsi G *et al* 2020 Economic and social consequences of human mobility restrictions under covid-19 *Proc. Natl Acad. Sci. USA* **117** 15530–5
[16] Lavezzo E *et al* 2020 Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo *Nature* **584** 1–5
[17] Aleta A *et al* 2020 Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19 *Nat. Human Behav.* **4** 964–71
[18] Park Y J *et al* 2020 Contact tracing during coronavirus disease outbreak, South Korea, 2020 *Emerg. Infectious Dis.* **26** 2465–8
[19] Kretzschmar M E *et al* 2020 Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study *Lancet Public Health* **5** e452–9
[20] Ferguson N *et al* 2020 Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand (available at: www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf)
[21] Goodfellow I, Bengio Y and Courville A 2016 *Deep Learn.* (Cambridge, MA: MIT Press)
[22] Cichos F, Gustavsson K, Mehlig B and Volpe G 2020 Machine learning for active matter *Nat. Mach. Intell.* **2** 94–103
[23] Black A J and McKane A J 2012 Stochastic formulation of ecological models and their applications *Trends Ecol. Evol.* **27** 337–45
[24] Codling E A, Plank M J and Benhamou S 2008 Random walk models in biology *J. R. Soc. Interface* **5** 813–34
[25] Spitzer F 2013 *Principles of Random Walk* vol 34 (New York: Springer Science and Business Media)
[26] Ichinose G, Satotani Y, Sayama H and Nagatani T 2018 Reduced mobility of infected agents suppresses but lengthens disease in biased random walk (arXiv:1807.01195)
[27] Biggerstaff M, Cauchemez S, Reed C, Gambhir M and Finelli L 2014 Estimates of the reproduction number for seasonal, pandemic and zoonotic influenza: a systematic review of the literature *BMC Infect. Dis.* **14** 1–20
[28] Peak C M, Childs L M, Grad Y H and Buckee C O 2017 Comparing nonpharmaceutical interventions for containing emerging epidemics *Proc. Natl Acad. Sci. USA* **114** 4023–8
[29] Liu Y, Gayle A A, Wilder-Smith A and Rocklöv J 2020 The reproductive number of COVID-19 is higher compared to SARS coronavirus *J. Travel Med.* **27** taaa021
[30] Cooper I, Mondal A and Antonopoulos C G 2020 A SIR model assumption for the spread of COVID-19 in different communities *Chaos Solitons Fractals* **139** 110057
[31] Ferretti L *et al* 2020 Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing *Science* **368** eabb6936
[32] Clipman S J *et al* 2020 Rapid real-time tracking of non-pharmaceutical interventions and their association with SARS-CoV-2 positivity: the COVID-19 pandemic pulse study *Clin. Infect. Dis.* ciaa1313
[33] Mehlig B 2019 Artificial neural networks (arXiv:1901.05639)
[34] Long Q-X *et al* 2020 Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections *Nat. Med.* **26** 1200–4
[35] Seow J *et al* 2020 Longitudinal evaluation and decline of antibody responses in SARS-CoV-2 infection *Nat. Microbiol.* **5** 1598–607
[36] Shaman J and Galanti M 2020 Will SARS-CoV-2 become endemic? *Science* **370** 527–9
[37] Zhu N *et al* 2020 A novel coronavirus from patients with pneumonia in China, 2019 *New Engl. J. Med.* **382** 727–33
[38] Wu F *et al* 2020 A new coronavirus associated with human respiratory disease in China *Nature* **579** 265–9
[39] World Health Organization 2020 Coronavirus disease 2019 (COVID-19): situation report, 21-09-2020 (available at: www.who.int/docs/default-source/coronaviruse/situation-reports/20200921-weekly-epi-update-6.pdf?sfvrsn=d9cf9496_6)
[40] Li X *et al* 2020 Emergence of SARS-CoV-2 through recombination and strong purifying selection *Sci. Adv.* eabb9153
[41] Pina A *et al* 2020 Virtual genetic diagnosis for familial hypercholesterolemia powered by machine learning *Eur. J. Prev. Cardiol.* **27** 1639–46

[42] Wang L and Wong A 2020 COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images *Sci. Rep.* **10** 1–12

[43] Melin P, Monica J C, Sanchez D and Castillo O 2020 Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: the case of Mexico *Healthcare* **8** 181

[44] Lalmuanawma S, Hussain J and Chhakchhuak L 2020 Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: a review *Chaos Solitons Fractals* **139** 110059

[45] Dandekar R and Barbastathis G 2020 Neural network aided quarantine control model estimation of global Covid-19 spread (arXiv:2004.02752)

[46] Kiang R *et al* 2006 Meteorological, environmental remote sensing and neural network analysis of the epidemiology of malaria transmission in Thailand *Geospat. Health* **1** 71–84

[47] Augusta C, Deardon R and Taylor G 2019 Deep learning for supervised classification of spatial epidemics *Spat. Spatio-temporal Epidemiol.* **29** 187–98

[48] Wells K *et al* 2020 COVID-19 control across urban-rural gradients *J. R. Soc. Interface* **17** 20200775

[49] Laguarta J, Hueto F and Subirana B 2020 COVID-19 artificial intelligence diagnosis using only cough recordings *IEEE Open J. Eng. Med. Biol.* **1** 275–81

[50] Tonini M *et al* 2020 A machine learning-based approach for wildfire susceptibility mapping. The case study of the Liguria region in Italy *Geosciences* **10** 105

[51] Athey S and Imbens G W 2019 Machine learning methods that economists should know about *Ann. Rev. Econ.* **11** 685–725

[52] Berg H C 1993 *Random Walks in Biology* (Princeton, NJ: Princeton University Press)

[53] Seitz M J and Köster G 2012 Natural discretization of pedestrian movement in continuous space *Phys. Rev.* E **86** 046108

[54] Leung K Y, Trapman P and Britton T 2018 Who is the infector? Epidemic models with symptomatic and asymptomatic cases *Math. Biosci.* **301** 190–8

[55] Stella L, Martínez A P, Bauso D and Colaneri P 2020 The role of asymptomatic individuals in the Covid-19 pandemic via complex networks (arXiv:2009.03649)

[56] Arcede J P, Caga-Anan R L, Mentuda C Q and Mammeri Y 2020 Accounting for symptomatic and asymptomatic in a SEIR-type model of COVID-19 *Math. Modelling Nat. Phenom.* **15** 34

[57] Chen Y-C, Lu P-E, Chang C-S and Liu T-H 2020 A time-dependent SIR model for COVID-19 with undetectable infected persons *IEEE Trans. Network Sci. Eng.* **7** 3279–94

[58] Agostinelli F, Hoffman M, Sadowski P and Baldi P 2014 Learning activation functions to improve deep neural networks (arXiv:1412.6830)

[59] Behnke S 2003 *Hierarchical Neural Networks for Image Interpretation* vol 2766 (Berlin: Springer)

[60] Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58

[61] Krogh A and Vedelsby J 1995 Neural network ensembles, cross validation and active learning *Advances in Neural Information Processing Systems* **7** 281

[62] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)

[63] Ruder S 2016 An overview of gradient descent optimization algorithms (arXiv:1609.04747)

[64] Chollet F *et al* 2015 Keras (available at: https://keras.io)