A data augmentation strategy for deep neural networks with application to epidemic modelling

Muhammad Awais,* Abu Sayfan Ali† Giacomo Dimarco‡, Federica Ferrarese§, Lorenzo Pareschi,¶

Abstract

In this work, we integrate the predictive capabilities of compartmental disease dynamics models with machine learning ability to analyze complex, high-dimensional data and uncover patterns that conventional models may overlook. Specifically, we present a proof of concept demonstrating the application of data-driven methods and deep neural networks to a recently introduced SIR-type model with social features, including a saturated incidence rate, to improve epidemic prediction and forecasting. Our results show that a robust data augmentation strategy trough suitable data-driven models can improve the reliability of Feed-Forward Neural Networks (FNNs) and Nonlinear Autoregressive Networks (NARs), making them viable alternatives to Physics-Informed Neural Networks (PINNs). This approach enhances the ability to handle nonlinear dynamics and offers scalable, data-driven solutions for epidemic forecasting, prioritizing predictive accuracy over the constraints of physics-based models. Numerical simulations of the post-lockdown phase of the COVID-19 epidemic in Italy and Spain validate our methodology.

Key words. Data Driven models in epidemiology, Deep learning, Data augmentation, Feed-Forward Neural Networks, Nonlinear Autoregressive Networks, COVID 19 data

1 Introduction

The COVID-19 pandemic has underscored the need for accurate and computationally efficient predictive models that not only describe the spread of infectious diseases but also provide fast and reliable forecasts to guide public health interventions. A key challenge in epidemic modeling is balancing mechanistic realism with computational tractability. Among the most widely used approaches are compartmental models, such as the classical Susceptible-Infectious-Recovered (SIR) models, which describe disease dynamics by partitioning the population into epidemiologically relevant compartments [16, 29, 32]. Specifically, these models categorize individuals as susceptible (capable of contracting the disease), infected (currently infected and capable of transmitting the disease), and recovered (immune or isolated from the population). Despite

*Department of Mathematics and Computer Science, University of Ferrara, Italy, e-mail: wsammm@unife.it

†Department of Mathematics and Computer Science, University of Ferrara, Italy, e-mail: abusafyan.ali@unife.it

‡Department of Mathematics and Computer Science, University of Ferrara, Italy, e-mail: giacomo.dimarco@unife.it

§Department of Mathematics and Computer Science, University of Ferrara, Italy, e-mail: federica.ferrarese@unife.it

¶Department of Mathematics and Computer Science, University of Ferrara, Italy, and Maxwell Institute for Mathematical Sciences and Department of Mathematics Heriot-Watt University, Edinburgh, UK, e-mail: lorenzo.pareschi@unife.it

their simplicity and effectiveness in capturing some fundamental epidemic trends, classical SIR models assume constant transmission and recovery rates, limiting their ability to represent the complex, nonlinear, and time-varying nature of real-world epidemics.

To address these limitations, extended versions of the SIR model have been proposed in the recent past, incorporating time- and state-dependent transmission rates, saturated incidence functions, and heterogeneous population behaviors [1, 12, 21]. For example, models incorporating saturated or nonlinear incidence rates [4, 22, 23], as well as kinetic approaches to contact dynamics [7], have proven effective in capturing the impact of social distancing measures during the pandemic. These refinements enhance realism and allow for better representation of the diverse factors influencing epidemic trajectories.

In parallel, machine learning techniques, particularly neural networks, have emerged as powerful tools for studying epidemic spread [3, 13, 33]. By leveraging their ability to approximate complex, nonlinear relationships, neural networks complement traditional modeling approaches and extend their applicability. Recent advancements, such as Physics-Informed Neural Networks (PINNs), integrate mechanistic knowledge of underlying physical or biological processes directly into the training phase of neural networks [20]. This integration ensures model robustness and consistency with known dynamics, even when data are scarce or uncertain. For instance, PINNs have been successfully applied to solve forward and inverse problems in epidemic modeling, ensuring that predictions adhere to underlying physical laws [6, 9, 28, 31, 41].

Feed-Forward Neural Networks (FNNs) and Nonlinear Autoregressive Networks (NARs) also offer promising approaches for epidemiological predictions [2, 27, 34]. FNNs map inputs directly to outputs through layers of nonlinear transformations, making them effective for tasks such as interpolation and short-term forecasting. This approach has been applied across various domains, including early recolonization detection in Parkinson disease, biological parameter estimation in disease models, COVID-19 data analysis, musculoskeletal disorder risk assessment, and statistical modeling for selecting optimal FNN architectures [17, 24, 36, 38, 42, 44]. However, FNNs do not account for temporal dependencies, limiting their ability to model the evolution of epidemics over time.

Conversely, NARs are specifically designed to process sequential data, making them well-suited for time-series forecasting in epidemic modeling. By incorporating previous values of the time series, NARs retain information about past states, enabling them to capture temporal dependencies in epidemic dynamics. These networks have been successfully applied in the recent past to forecast infection peaks, estimate recovery rates, and simulate intervention scenarios [10, 25, 30, 34, 35, 40, 43].

In this work, we introduce a hybrid epidemic forecasting approach that integrates mechanistic epidemiological models with data-driven FNNs and NARs, creating a surrogate model capable of generalizing across different epidemic scenarios. This approach enables fast, data-driven epidemic forecasting while preserving the interpretability of traditional mechanistic models. The resulting methodology is applied to analyze the post-lockdown phase of the COVID-19 epidemic in Italy and Spain, leveraging real-world data and enhanced computational strategies to improve predictive accuracy. This framework enables us to augment real-world epidemic data with synthetic simulations generated from a generalized SIR model, which includes a saturated incidence rate to better reflect intervention effects and behavioral changes. The main idea consists in addressing the limitations posed by sparse real-world data by generating additional synthetic information using a generalized SIR data driven model. Specifically, we focus on an extension known as the social-SIR model with a saturated incidence rate, referred to here as the generalized SIR model in the following. This model modifies the transmission dynamics [23] by allowing the infection rate to vary with time and the number of infected individuals. Such modifications enable the incorporation of external interventions, such as lockdowns or vaccination campaigns, as well as psychological and behavioral adaptations [14].

2

The enriched dataset, combining real and synthetic data, is then used to train the neural networks, allowing them to effectively learn epidemic dynamics and improve predictive accuracy. By leveraging the flexibility of FNNs and the sequential modeling capabilities of NARs, our approach captures both short-term and long-term epidemic trends. Unlike PINNs, which explicitly encode governing equations into the network architecture, our method enables networks to infer dynamics implicitly through data-driven learning. This eliminates the need for embedding differential equations directly into the training process, enhancing adaptability to different scenarios while reducing computational complexity.

The remainder of this paper is structured as follows. In Section 2, we recall some standard features about the classical SIR model and we introduce the social-SIR model with a saturated incidence rate used to produce synthetic data. Section 3 defines the optimization problems used to estimate parameters and the incidence function during the pre- and post-lockdown phases in Italy and Spain. In Section 4, we present the neural network structures, distinguishing between FNNs and NARs and we perform some detailed numerical experiments to validate our strategy. Finally, Section 5 summarizes our findings and outlines potential future research directions.

2 Problem setting

The governing equations of the standard SIR model, which describes the spread of an infectious disease in a population of N individuals divided into three compartments, Susceptible (S), Infected (I), and Recovered (R), are given by

$$
\begin{aligned}
\frac{dS(t)}{dt} &= -\beta \frac{S(t)I(t)}{N}, \\
\frac{dI(t)}{dt} &= \beta \frac{S(t)I(t)}{N} - \gamma I(t), \\
\frac{dR(t)}{dt} &= \gamma I(t),
\end{aligned}
\tag{2.1}
$$

where the total population corresponds to $N = S(t) + I(t) + R(t)$, and the system is complemented with suitable initial conditions. The parameters $\beta > 0$ and $\gamma > 0$ represent the disease transmission rate and the recovery rate, respectively, and are assumed to be constant in (2.1). For simplicity, we also absorb N into the coefficient β in the following.

The above model despite of its simplicity it has been shown to be able to capture some features related to the spread of an epidemic. However, as expected, it also suffers of some drawbacks due to the inability to describe more complex structured solutions which has been observed for instance during the last COVID-19 epidemic [26]. In particular, recent studies have highlighted the crucial role of transmission coefficients in relation to social behavior, which significantly influences pathogen spread [5]. In this research direction, the works in [5, 23] explored how controlling the number of daily social contacts can guide partial lockdown strategies. This is type of phenomena we address in this work and which cannot be captured with the system (2.1). Building on this perspective, some of the authors of the present work have developed in the recent past a new mathematical framework linking social contact distributions with disease transmission in multi-agent systems in [23]. This is achieved by integrating a classical compartmental epidemiological model with kinetic equations governing social contact

formation. The resulting model is given by

$$\begin{aligned}
\frac{dS(t)}{dt} &= -\beta S(t)I(t)H(t, I(t)), \\
\frac{dI(t)}{dt} &= \beta S(t)I(t)H(t, I(t)) - \gamma I(t), \\
\frac{dR(t)}{dt} &= \gamma I(t),
\end{aligned} \tag{2.2}$$

where $H(t, I(t))$ is a macroscopic incidence rate [23], representing a time-dependent adjustment to infection rates based on behavioral and policy changes (e.g., lockdowns). In the sequel we consider the following bounded function that moderates infection spread

$$H(t, I(t)) = \frac{1}{1 + \frac{\nu}{\beta} \int_0^t I(s)ds}, \tag{2.3}$$

where $\nu > 0$ is a constant that it has been shown to reflect the variance in contact rates within the population. The system (2.2) even if built on a different basis, shares conceptual similarities with the models introduced in [16], where a nonlinear, bounded function $g(I)$ modulates disease transmission. This approach accounts for the fact that, as the number of infected individuals increases, the effective infection force may decrease due to factors such as behavioral adaptations or resource limitations. On the same direction in [37] it has been introduced a nonlinear, saturated incidence rate of the form

$$g(I)S = \frac{\beta I^m S}{1 + \alpha I^k}, \quad \text{with} \quad \frac{dS(t)}{dt} = -g(I(t))S(t) \tag{2.4}$$

where $\beta I^m S$ represents the baseline infection force, and $\frac{1}{1+\alpha I^k}$ accounts for behavioral responses, such as increased caution among susceptible individuals. Here, β, m, k are positive constants, and α is a non-negative constant [19, 45, 46]. A specific case, with $m = 1, k = 1$, yields the form proposed in [16] to model the crowding effect observed during the 1973 cholera outbreak in Bari:

$$g(I) = \frac{\beta I}{1 + \alpha I}. \tag{2.5}$$

Alternatively, setting $m = 1, k = 2$ results in the Monod‚ÄìHaldane (M‚ÄìH) function,

$$g(I) = \frac{\beta I}{1 + \alpha I^2}, \tag{2.6}$$

which is commonly used to model physiological constraints in infection dynamics.

The so-called social-SIR model (2.2) exhibits a similar incidence rate behavior, extending the form in (2.5) to account for the impact of protective measures such as lockdown strategies and infection control. Unlike classical models, this generalized framework incorporates a time-dependent, saturated incidence rate to capture the cumulative effects of infection dynamics and behavioral adaptations in response to an epidemic. This model has previously been successfully employed to analyze the effects of partial lockdown strategies in various European countries. Given that we choose in the following, for demonstrating the proposed data augmentation strategy, a dataset corresponding to a period marked by COVID-19 lockdowns, we opted to use this model in our study. It is worth to mention that, the complexity introduced by lockdown measures, makes the forecasting process more challenging compared to standard scenarios. Therefore, in the following, we employ system (2.2) to develop our data-driven strategy and the associated deep learning framework, enabling epidemic forecasting.

4

3 On the the construction of a data driven social-SIR model

In the context of epidemic prediction, mathematical models such as those presented in the previous section provide valuable insights into the spread of infectious diseases. However, these models often rely on predefined assumptions and parameters that may not fully capture the complexities of real-world epidemics. Neural networks offer a powerful alternative by learning directly from data, enabling them to model complex, nonlinear relationships that traditional epidemic models may struggle to represent. However, one major limitation of training neural networks exclusively on available data is that, while they can effectively interpolate within a given time interval, they struggle to make accurate predictions beyond the observed data points [2, 9, 33]. To address this issue, we propose an approach that combines a mathematical epidemiological model,Äîthe social-SIR model (2.2),Äîwith real outbreak data. The key idea is to generate additional synthetic information by solving this generalized SIR model, thereby filling the gaps between available real or measured data points. With this enriched dataset, one can train a neural network more effectively to predict future infection trends in a given population. As we demonstrate later, this approach leverages both real and synthetic data, significantly improving predictive accuracy and enhancing the model‚Äôs ability to forecast epidemic dynamics. However, to implement this strategy, the model must be capable of interpolating through the experimental data points. Therefore, in the following section, we discuss how the parameters in (2.2) are defined.

3.1 Parameters estimation

To develop our approach, we focus on the COVID-19 outbreak in two European countries, Italy and Spain, during the 2020 pandemic. Given that both countries implemented restrictive measures during this period, we proceed as follows. We define two optimization phases: in the first phase, we estimate the parameters β and γ for the pre-lockdown period using the model (2.2) with $H(t, I(t)) = 1$; in the second phase, we estimate the function $H(t, I(t))$ from post-lockdown data.

The COVID-19 data for Italy are available from the GitHub repository of the Italian Civil Protection Department, while data for Spain can be accessed here. In practice, to estimate the parameters β and γ, we formulate the following minimization problem:

$$\min_{\beta, \gamma} \mathcal{J}(\hat{I}, I), \tag{3.1}$$

subject to the constraint that the equations of system (2.2) have to be satisfied when $H(t, I(t)) = 1$. The objective function to be minimized is defined as

$$\mathcal{J}(\hat{I}, I) = \|\hat{I}(t) - I(t)\|^2, \tag{3.2}$$

where $\hat{I}(t)$ represents the reported infected cases from the data set at disposal, while $I(t)$ is the model output, and $\| \cdot \|$ denotes the L_2 norm.

We consider data over the time interval $[t_0, t_l)$, where t_0 (set to February 24, 2020) marks the initial time, and t_l represents the lockdown start date (March 9 for Italy and March 14 for Spain). We search for a minimum under the constraints

$$0 \leq \beta \leq 0.6, \qquad 0 \leq \gamma \leq 0.06, \tag{3.3}$$

where the bounds are chosen accordingly to the available estimates [26]. To solve the optimization problems in (3.1), we rely on the use of the Matlab functions `fmincon` in combination with

a RK4 integration method of the system of ODEs. The result of the optimization procedure on the pre-lockdown phase gives, for Italy, the following values

$$\beta_{est}^{It} = 0.3107, \qquad \gamma_{est}^{It} = 0.0328, \tag{3.4}$$

while for Spain

$$\beta_{est}^{Sp} = 0.2340, \qquad \gamma_{est}^{Sp} = 0.0224. \tag{3.5}$$

In Figure 1 on the left the evolution of the dynamics in the pre-lockdown phase in Italy obtained by solving the social-SIR model (2.2) with $H(t, I(t)) = 1$, using the estimated parameters $\beta = \beta_{est}^{It}$ and $\gamma = \gamma_{est}^{It}$ in (3.4). This is compared, in the same image, to the experimental data from the pre-lockdown phase, spanning up to time $t_l = 14$ with a time step $\Delta t = 1$ corresponding to 14 days with 1 day interval between the measurements. The estimated dynamics aligns quite well with the available data, as confirmed by the embedded plot which shows the pointwise error over time between the solution of the generalized SIR model and the data at disposal, i.e.

$$\mathcal{E}(t) = |\hat{I}(t) - I(t)|. \tag{3.6}$$

In the same Figure on the right the pre-lockdown dynamics in Spain computed by solving the social-SIR model (2.2) with $H(t, I(t)) = 1$, using the estimated parameters $\beta = \beta_{est}^{Sp}$ and $\gamma = \gamma_{est}^{Sp}$ in (3.5) compared with the experimental data spanning up to time $t_l = 18$. The two solutions are in good agreement as confirmed by the plot showing the pointwise error computed as in (3.6).
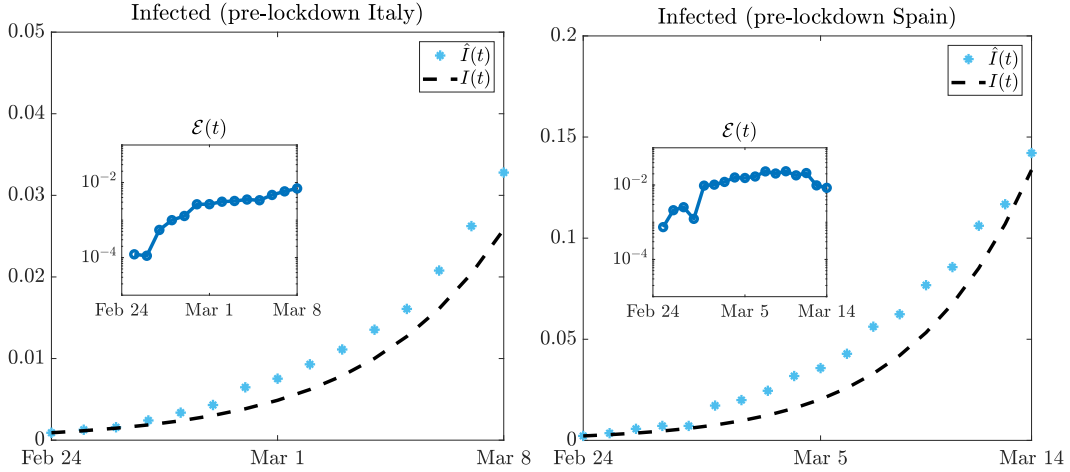

Figure 1: Pre-lockdown phase. Dynamics of the infected population computed by solving the social-SIR model (2.2) with β and γ as in (3.4) (Italy), and as in (3.5) (Spain) and $H(t, I(t)) = 1$, compared with the experimental data. In the embedded plot the pointwise error between the two quantities computed as in (3.6). On the left, Italy. On the right, Spain.

Once β and γ have been estimated, we proceed to determine the pointwise values $H(t, I(t))$ in (2.2) in such a way that the model fits as well as possible the post-lockdown data, i.e., data in the interval $[t_l, T]$, where T (set to July 1st) is the final observation time for both countries. To this end, we solve the following new optimization problem

$$\min_{H(t)} \mathcal{J}(\hat{I}, I), \tag{3.7}$$

6

subject to the constraint that the solution has to satisfy the equations of the system (2.2), with $\mathcal{J}(\cdot)$ defined as in (3.2), and that

$$0 \leq H(t) \leq 1. \tag{3.8}$$

In order to determine the optimal value of H, we split the time interval $[t_l, T]$ into N_t time intervals of length Δt, and we solve a sequence of optimal control problems

$$\min_H \mathcal{J}^{\Delta t}(\hat{I}, I) := \int_t^{t+\Delta t} \|\hat{I}(s) - I(s)\|^2 ds, \tag{3.9}$$

s.t. the explicit time discretization of the generalized SIR dynamics

$$\begin{aligned} S^{n+1} &= S^n + \Delta t \beta I^n S^n H^n, \\ I^{n+1} &= I^n + \Delta t \left(\beta I^n S^n H^n - \gamma I^n \right), \\ R^{n+1} &= R^n + \Delta t \gamma I^n, \end{aligned} \tag{3.10}$$

where we denoted by $t^n = n\Delta t$ for any $n = 1, \ldots, N_t$. We then consider a discretization of the integral in (3.9) using the rectangle rule

$$\mathcal{J}^{\Delta t}(\hat{I}, I) = \Delta t \left(\hat{I}^{n+1} - I^{n+1} \right). \tag{3.11}$$

By direct computation, it can be shown that the expression for H at each discrete time t^n is given by

$$H^n = \mathbb{P}_{[0,1]} \left(\frac{\hat{I}^n - I^n + I^n \gamma \Delta t}{I^n S^n \beta \Delta t} \right), \tag{3.12}$$

where $\mathbb{P}_{[0,1]}(\cdot)$ denotes the projection over the interval $[0, 1]$.

The results of this study are reported in Figure 2 where on the left we show the post-lockdown experimental data and the solution of the generalized SIR model with the contact function H as in (3.12). The estimated and true solution are in good agreement as shown in the embedded plot which represents the corresponding pointwise error computed as in (3.6). On the right, we show the estimated values of H^n computed as in (3.12) as a function of time. The function H exhibits oscillations over time, suggesting potential measurement errors in the data. First row, Italy. Second row, Spain.
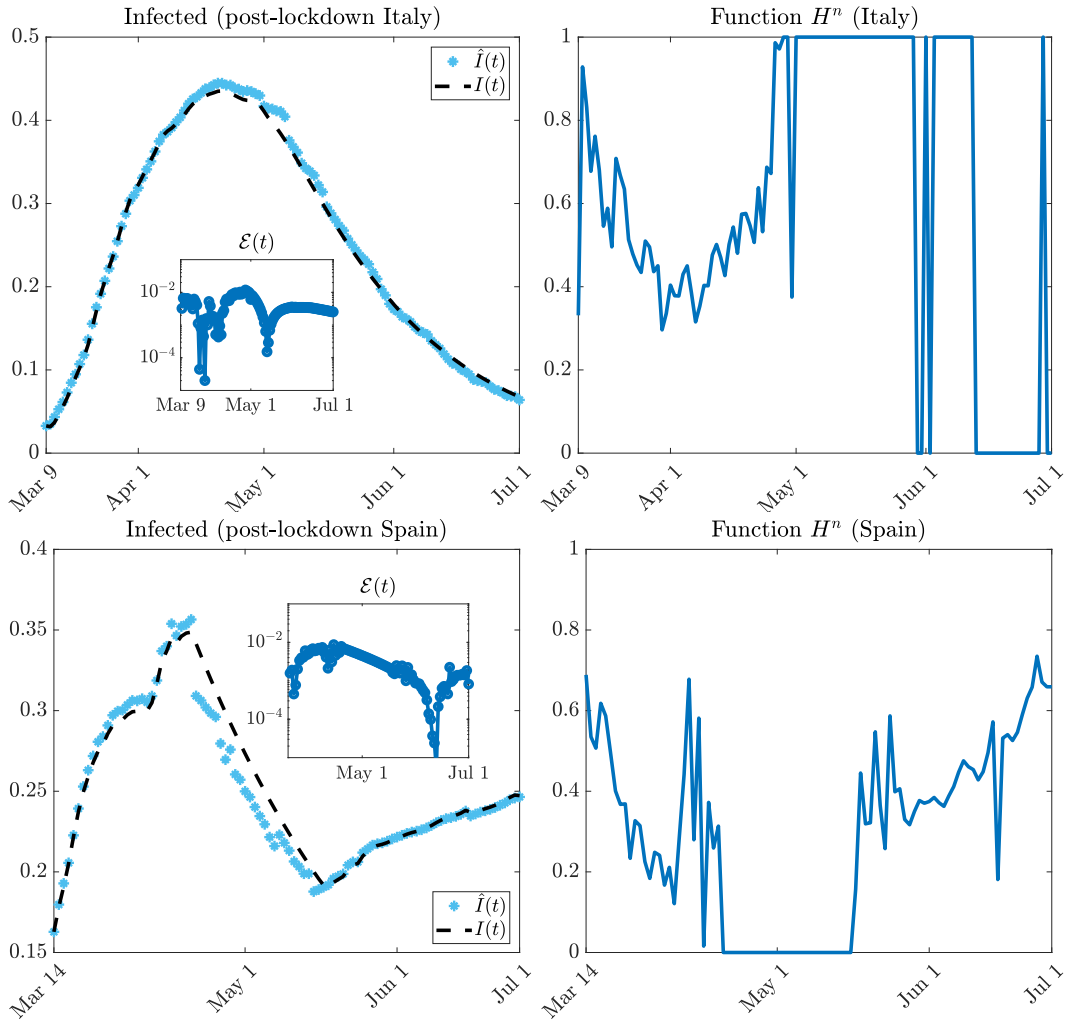
Figure 2: Post-lockdown phase. On the left the dynamics of the infected population computed by solving the generalized SIR model (2.2) with β and γ as in (3.4) (Italy), and as in (3.5) (Spain) compared with the experimental data. In the embedded plot the pointwise error between the two infected populations computed as in (3.6). On the right the contact function H computed as in (3.12). First row, Italy. Second row, Spain.

4 Estimation and prediction using real COVID-19 data and Neural Networks design for epidemic prediction

In this section, we consider two distinct types of neural network architectures: Feed-Forward Neural Networks (FFNs) and Nonlinear Autoregressive Networks (NARs). Both approaches

8

are implemented using a general neural network framework with $L+1$ layers, as follows

$$x^1 = W^1 x + b^1, \qquad x^l = \sigma \circ (W^l x^{l-1} + b^l), \quad l = 2, \ldots, L-1,$$
$$U^{NN}(x; \mathbf{b}, \mathbf{W}) = W^L z^{L-1} + b^L, \tag{4.1}$$

where $x = (x^1, \ldots, x^d) \in \mathbb{R}^d$ is the input, σ is an activation function to convert the input signals to output signals, and $\mathbf{W} = (W^1, \ldots, W^L)$ and $\mathbf{b} = (b^1, \ldots, b^L)$ represent the weights and biases. The weights and biases associated with these connections serve as the parameters of the network, adjusted iteratively through techniques like gradient descent or Adam method during the training phase to minimize prediction errors and enhance model performance. In our context, the idea is to solve the following minimization problem

$$\theta^* = \arg \min_\theta \mathcal{L}(\theta), \tag{4.2}$$

to find the optimal parameters $\theta = (W^1, b^1, \ldots, W^L, b^L)$ which minimize a certain loss function $\mathcal{L}(\theta)$ that includes errors from both real and synthetic data generated by solving the generalized SIR model. Real data provides an accurate representation of the epidemic dynamics, while synthetic data fills gaps between observed data points and provides high-resolution trajectories.

The loss function is expressed as

$$\mathcal{L}(\theta) = \omega_r \mathcal{L}_r(\theta) + \omega_s \mathcal{L}_s(\theta), \tag{4.3}$$

where $\omega_r, \omega_s > 0$ are given weights, and

$$\mathcal{L}_j(\theta) = \frac{1}{N_j} \sum_{i=1}^{N_j} \|I_j(t) - I_j^{NN}(t, \theta)\|^2, \tag{4.4}$$

for $j = \{r, s\}$, where $I_j(t)$ is the actual number of infected. These are either the real data or the data generated by solving the social SIR model (synthetic data), while $I_j^{NN}(t, \theta)$ is the predicted state. Finally, N_r, N_s refer to the number of available real and synthetic data points respectively. With this choice, it is possible to appropriately weight the contribution of real data alongside the data generated by the model.

Feed-Forward Neural Networks. FFNs are a foundational type of neural network architecture designed to map inputs directly to outputs without considering temporal dependencies. For epidemic prediction, we design a FFN that takes the current time t as input and produces the value of infected $I(t)$ as output. This approach relies on learning a direct mapping between the time and the epidemic dynamics, effectively interpolating the trajectory based on the training data. The architecture of the FFN consists of the following components

- **Input Layer:** The single input feature is the current time t.
- **Hidden Layers:** One or more fully connected layers with nonlinear activation functions (e.g., ReLU or tanh) capture the relationship between time and the epidemic states.
- **Output Layer:** The output consists of a node corresponding to the predicted value of the infected at time t, i.e. $I(t)$.

It is well known that, while FFNs are effective for interpolation tasks, their inability to model sequential dependencies limits their accuracy in predicting future epidemic dynamics. This limitation arises because FFNs lack memory mechanisms, meaning they cannot leverage information from previous states to influence current predictions. For this reason, in the following we propose an alternative strategy.

Nonlinear Autoregressive Networks. To capture the temporal dynamics of the epidemic, we then rely on Nonlinear Autoregressive Networks (NARs) capable of modeling sequential data. Unlike FFNs, NARs incorporate information about past states, making them particularly effective for time-series predictions. The NAR is designed to predict the future state of the epidemic based on the previous state of the solution. The architecture consists of the following components:

- **Input Layer:** The input features are the values of the infected at time $t - d, \ldots, t - 1$, i.e. $I(t - d), \ldots, I(t - 1)$ for a certain $d \geq 1$.

- **Hidden Layers:** One or more fully connected layers with nonlinear activation functions (e.g., ReLU or tanh) capture nonlinear relationships in the time series data.

- **Output Layer:** Produces the predicted next state, $I(t)$, representing the system's state at time t.

Once trained, NAR networks can predict the future states of the epidemic compartments by recursively feeding its outputs back as inputs for subsequent time steps (closed loop strategy). This enables the generation of full epidemic trajectories from a given initial condition.

4.1 Neural network for epidemic prediction

To evaluate the performances of FFNs and NARs in epidemic prediction, we conduct some numerical experiments using both real and synthetic data. The real data covers the period from the lockdown start date (in Italy March 9th, in Spain March 14th) to July 1st. These data are available on a daily basis, corresponding to a time step of $\Delta t_r = 1$. The synthetic data are generated by solving the generalized SIR model given in (2.2), with β and γ as in (3.4) (Italy), and as in (3.5) (Spain) over the same time interval, using a finer time step $\Delta t_s = 0.2$, and with H estimated by solving the minimization problem in (3.7). Since H is originally defined at daily intervals as in (3.12), we use linear interpolation to extend its definition to the finer grid. Both networks are trained using the Levenberg-Marquardt optimizer for 500 epochs. In the case of real data we set $\omega_r = 1$ and $\omega_s = 0$, while in the case of synthetic data we set $\omega_r = 0$ and $\omega_s = 1$ in (4.3).

Feedforward neural network. We design a FFN with three fully connected layers, each consisting of 10 nodes, and use the `tanh` activation function, relying on the use of the Matlab function `feedforwardnet`. The input to the network is the time interval $[t_l, T]$, being t_l the lockdown starting date and T the final time supposed to be July 1st. The output is the normalized number of infected $I(t)$, which corresponds to either real data or synthetic data generated as described above. We train the neural network using the Matlab function `train`. To evaluate the neural network's performance, we first compute its solution over the training set for both Italy (first row) and Spain (second row). Specifically, we consider the time interval $[t_l, T]$ with a time step $\Delta t_r = 1$. In the left panel of Figure 3 and of Figure 4, we compare the solution obtained using a neural network trained on both real and synthetic data, with the available data for respectively the Italian and the Spanish case. The neural network provides a good approximation of the solution, making it well-suited for interpolation tasks. Next, to assess its predictive capabilities, we focus on the time interval $(T, T_{test}]$, where T (July 1st) is the starting time and T_{test} (July 15th) is the final time, using the same daily discretization. The right panels of Figures 3 and 4 present a comparison between the network's predictions for the test set and the daily real data available in that period. Additionally, the figures display the pointwise error between the neural network's outputs and the real data, calculated as in (3.6). The standard FNN as expected faces challenges in forecasting future dynamics and the data augmentation does not provide a clear improvement in the results.
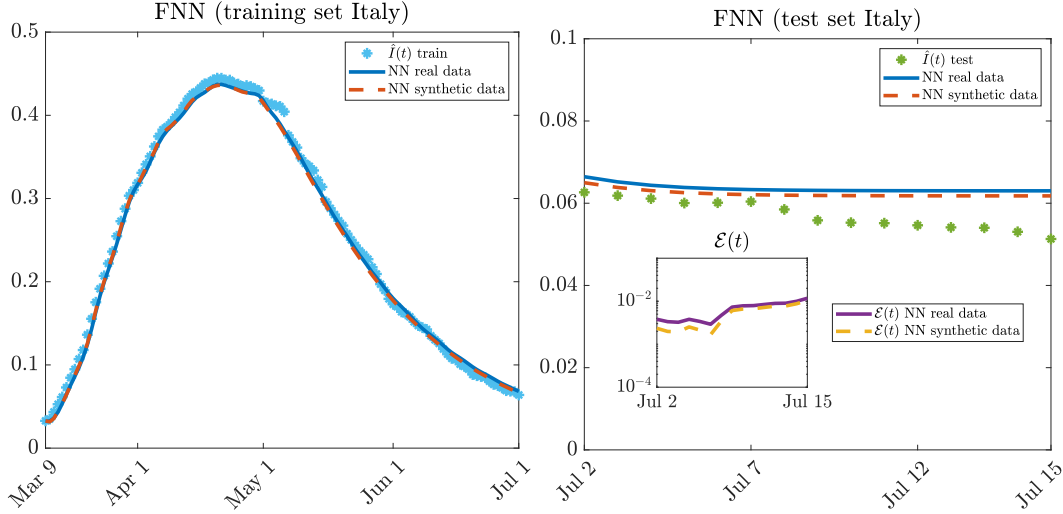
Figure 3: Feedforward neural network for Italy. On the left the dynamics of the infected population computed by training a FNN on real and synthetic data and tested over the train interval $[t_l, T]$ compared to the available data. On the right the solution computed with the same FNN tested over the time interval $(T, T_{test}]$ compared to the available data. Embedded the error between the available data test and the neural network solution computed as in (3.6).

Nonlinear autoregressive neural network.   We design now a NAR network with three fully connected layers, each consisting of 10 nodes, and use the `tanh` activation function. To implement the neural network we use the Matlab function `narnet`. We consider a time interval $[t_l, T]$, being t_l the time in which the lockdown started and T the final time supposed to be July 1st. We discretize the interval with time step $\Delta t_r = 1$ if we deal with real data, or with time step $\Delta t_s = 0.2$ if we deal with synthetic data as before. The input to the network is the number of infected at time $t - 2\Delta t, t - \Delta t$, and the output is $I(t)$ that is the number of infected at time t. Both the input and output data are generated as described above by considering a social-SIR model. As for the FNN we first evaluate the performance of the NAR network on the training set. Since the number of infected individuals during the training interval is known, we can use an open-loop prediction approach using the true training set values to forecast the next time step's solution.
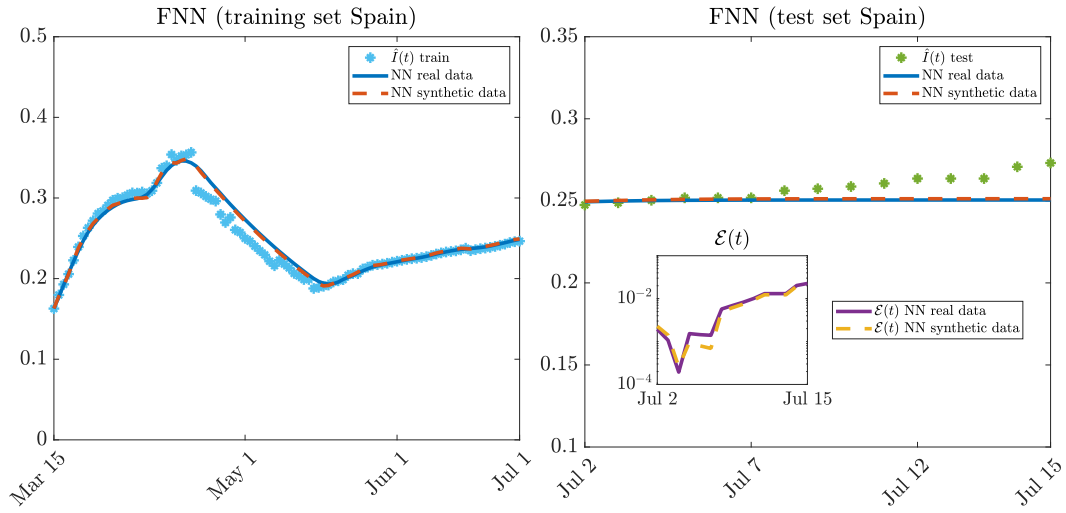
Figure 4: Feedforward neural network for the Spain case. On the left the dynamics of the infected population computed by training a FNN on real and synthetic data and tested over the train interval $[t_l, T]$ compared to the available data. On the right the solution computed with the same FNN tested over the time interval $(T, T_{test}]$ compared to the available data. Embedded the error between the available data test and the neural network solution computed as in (3.6). First row, Italy. Second row, Spain.

Figure 5 shows the solution obtained by training a NAR network on both real and synthetic data, then testing it on the same dataset and comparing it to the available data. On the left, we have the results for Italy, and on the right, for Spain. Both the NAR networks, trained on real and synthetic data, are able to accurately approximate the underlying dynamics when one considers the training interval.
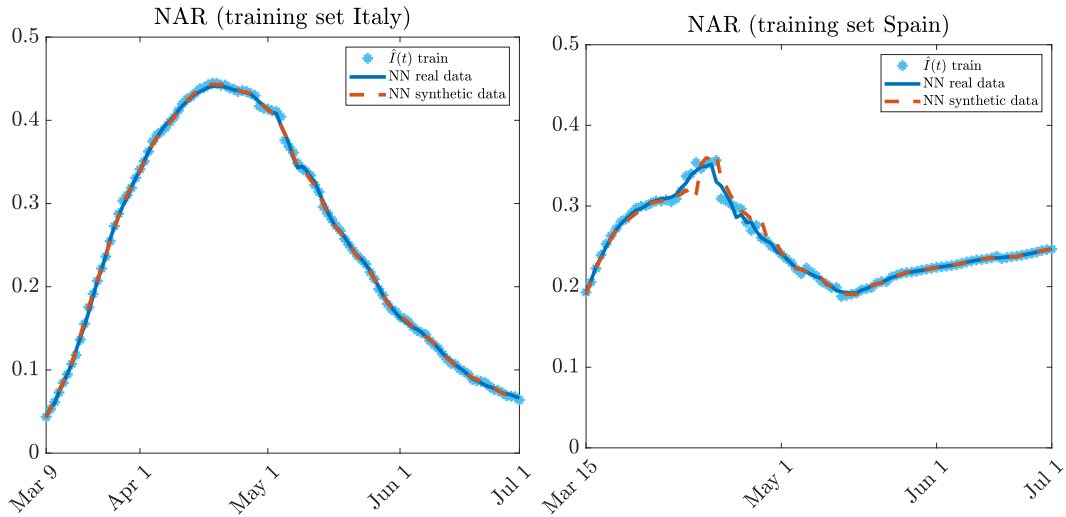
12

Figure 5: Nonlinear autoregressive network: training data. Solution obtained by training a NAR network on both real and synthetic data, then testing it on the same dataset and comparing it to the available data. On the left, the case for Italy, and on the right, the one for Spain.

We then test the NAR network for both short-term and long-term predictions. For short-term predictions, we compare the neural network solution with a test set containing daily data from July 2nd to July 8th for both Italy and Spain, while for long-term predictions, we extend the test period to July 15th. Since these data are assumed to be unknown a priori, we convert the open-loop NAR network into a closed-loop form. This allows the network to generate predictions for the desired number of steps by using its previous predictions as input, without relying on the true values for evolution. Figure 6 shows the solution produced by the NAR network trained on both real and synthetic data, compared with the available true data. Additionally, we plot the pointwise error in time computed as in (3.6). In the first row, the short-term prediction while in the second row, the long term prediction. On the left, Italy. On the right, Spain. The NAR network trained on synthetic data is able to provide a very accurate approximation of the solution and demonstrates strong predictive capabilities. Notably, it significantly outperforms the results obtained when training exclusively on real data, as it is particularly evident in the case of Italy.
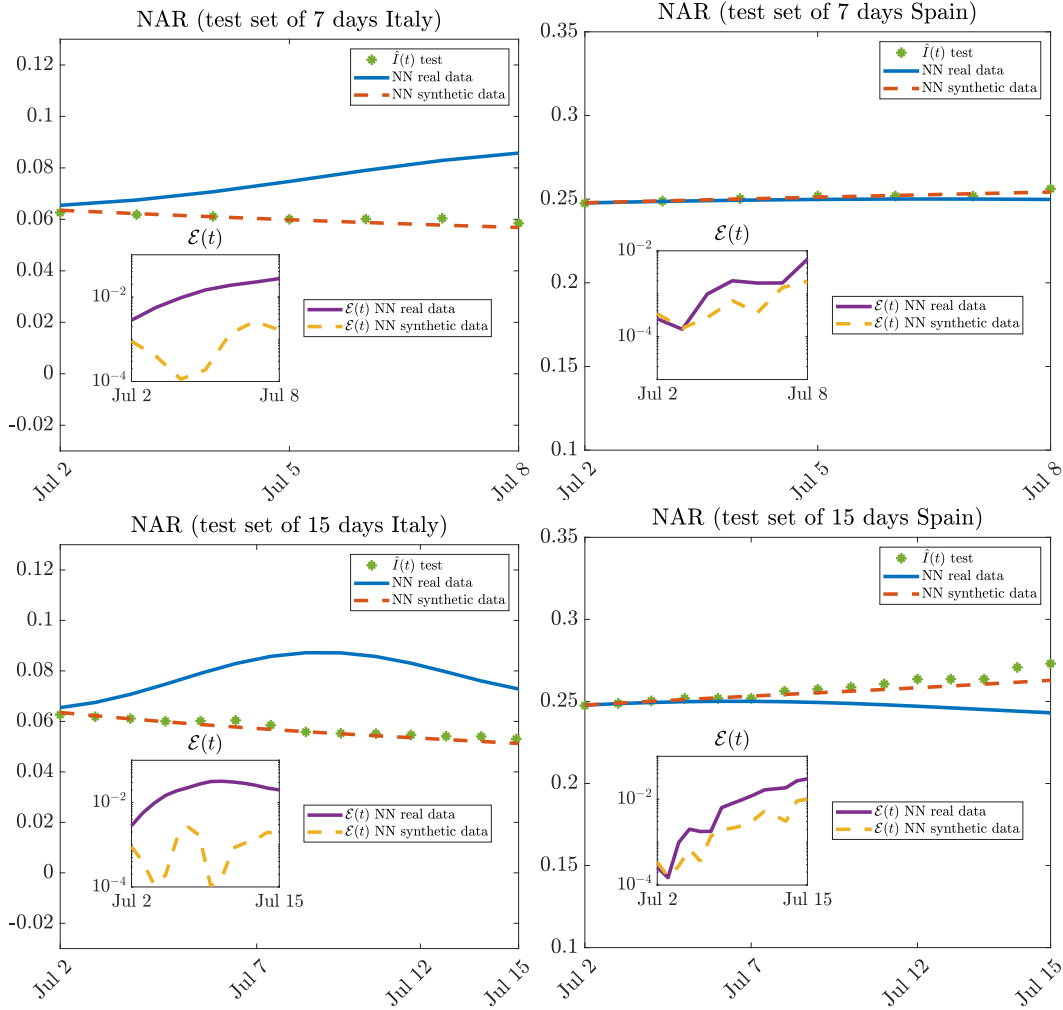
Figure 6: Nonlinear autoregressive network: short and long time predictions. Solution obtained by training a NAR network on both real and synthetic data, then testing it to produce daily data from July 2nd to July 8th (first row), and extend to July 15th (second row), compared with the available true data. On the left, the case for Italy, and on the right, the one for Spain.

5 Conclusion

In this study, we explored the impact of a data augmentation strategy on enhancing the forecasting capabilities of Feed-Forward Neural Networks (FNNs) and Nonlinear Autoregressive Networks (NARs) in analyzing and predicting the dynamics of the COVID-19 epidemic during the post-lockdown phase in Italy and Spain. By integrating real data with synthetic data generated from a generalized SIR model, we demonstrated the effectiveness of neural networks in capturing complex epidemic dynamics. FNNs performed well in interpolating within observed data, while NARs were particularly effective in capturing temporal dependencies, making them

14

suitable for long-term forecasting. The generalized SIR model provided a robust foundation for generating high-resolution synthetic data, enriching the neural network training process. The key outcome of this work is the development of a hybrid methodology that bridges the gap between traditional mathematical modeling and machine learning techniques, such as Physics-Informed Neural Networks (PINNs), offering a scalable and flexible framework for epidemic modeling and prediction in real-world scenarios.

While this study has focused on FNNs and NARs, other neural network architectures could be explored to further improve epidemic forecasting. For example, Long Short-Term Memory (LSTM) networks and Transformer-based models have been increasingly used for time-series prediction and could provide additional advantages in learning long-term dependencies [2, 11, 39]. Similarly, hybrid AI-mechanistic models, which integrate deep learning with traditional compartmental models, have shown promise in epidemic modeling [47]. Future research could investigate whether these approaches provide added value in capturing epidemic trends compared to the models used in this study, particularly in scenarios with limited or noisy data.

An important direction for future research is the incorporation of uncertainty quantification (UQ) to better account for variability and noise in real-world data [1, 8, 15, 18]. While neural networks provide powerful forecasting tools, their deterministic nature limits their ability to express confidence in predictions, which is crucial in public health decision-making. Approaches such as Bayesian deep learning, Monte Carlo Dropout, and ensemble learning can be explored to quantify the range of possible epidemic trajectories. Additionally, integrating probabilistic surrogate models with mechanistic epidemiological frameworks can improve risk assessment and early-warning systems for future outbreaks. By embedding UQ techniques within the proposed hybrid framework, epidemic forecasting can become more robust, interpretable, and reliable to better account for variability and noise in real-world data.

Acknowledgments

References

[1] G. Albi, G. Bertaglia, W. Boscheri, G. Dimarco, L. Pareschi, G. Toscani, and M. Zanella. Kinetic modelling of epidemic dynamics: social contacts, control with uncertain data, and multiscale spatial dynamics. In *Predicting Pandemics in a Globally Connected World: Toward a Multiscale, Multidisciplinary Framework through Modeling and Simulation*, volume 1, pages 43–108. Springer, 2022.

[2] A. B. Amendolara, D. Sant, H. G. Rotstein, and E. Fortune. Lstm-based recurrent neural network provides effective short term flu forecasting. *BMC Public Health*, 23(1):1788, 2023.

[3] N. Baker, F. Alexander, T. Bremer, A. Hagberg, Y. Kevrekidis, H. Najm, M. Parashar, A. Patra, J. Sethian, S. Wild, et al. Brochure on basic research needs for scientific machine learning: Core technologies for artificial intelligence. Technical report, USDOE Office of Science (SC)(United States), 2018.

[4] M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani. Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. *J. Theoret. Biol.*, 235(2):275–288, 2005.

[5] G. Beraud, S. Kazmercziak, P. Beutels, D. Levy-Bruhl, X. Lenne, N. Mielcarek, Y. Yazdanpanah, P.-Y. Boelle, N. Hens, and B. Dervaux. The french connection: The first large population-based contact survey in france relevant for the spread of infectious diseases. *PLoS ONE*, 10(7), 2015.

[6] S. Berkhahn and M. Ehrhardt. A physics-informed neural network to model COVID-19 infection and hospitalization scenarios. *Advances in continuous and discrete models*, 2022(1):61, 2022.

[7] G. Bertaglia, W. Boscheri, G. Dimarco, and L. Pareschi. Spatial spread of COVID-19 outbreak in Italy using multiscale kinetic transport equations with uncertainty. *Mathematical Biosciences and Engineering: MBE*, 18(5):7028–7059, 2021.

[8] G. Bertaglia, L. Liu, L. Pareschi, and X. Zhu. Bi-fidelity stochastic collocation methods for epidemic transport models with uncertainties. *Networks and Heterogeneous Media*, 17(3):401–425, 2022.

[9] G. Bertaglia, C. Lu, L. Pareschi, and X. Zhu. Asymptotic-preserving neural networks for multiscale hyperbolic models of epidemic spread. *Mathematical Models and Methods in Applied Sciences*, 32(10):1949–1985, 2022.

[10] G. Bontempi, S. Ben Taieb, and Y.-A. Le Borgne. Machine learning strategies for time series forecasting. *Business Intelligence: Second European Summer School, eBISS 2012, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures 2*, pages 62–77, 2013.

[11] G. Bontempi, S. Ben Taieb, and Y.-A. Le Borgne. Machine learning for time-series epidemic modeling. *Journal of AI & Healthcare*, 7(3):295–310, 2023.

[12] F. Brauer, C. Castillo-Chavez, and Z. Feng. *Mathematical Models in Epidemiology*, volume 69 of *Texts in Applied Mathematics*. Springer, New York, 2019. With a foreword by Simon Levin.

[13] S. L. Brunton and J. N. Kutz. *Data-driven science and engineering: Machine learning, dynamical systems, and control.* Cambridge University Press, 2022.

[14] B. Buonomo and R. Della Marca. Effects of information-induced behavioural changes during the COVID-19 lockdowns: The case of Italy. *Royal Society Open Science*, 7(3):201635, 2020.

[15] A. Capaldi, S. Behrend, B. Berman, J. Smith, J. Wright, and A. L. Lloyd. Parameter estimation and uncertainty quantication for an epidemic model. *Mathematical biosciences and engineering*, page 553, 2012.

[16] V. Capasso and G. Serio. A generalization of the Kermack-McKendrick deterministic epidemic model. *Mathematical Biosciences*, 42(1):43–61, 1978.

[17] C.-L. Chen, D. B. Kaber, and P. G. Dempsey. A new approach to applying feedforward neural networks to the prediction of musculoskeletal disorder risk. *Applied ergonomics*, 31(3):269–282, 2000.

[18] G. Chowell. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. *Infectious Disease Modelling*, 2(3):379–398, 2017.

[19] Q. Cui, Z. Qiu, W. Liu, and Z. Hu. Complex dynamics of an SIR epidemic model with nonlinear saturate incidence and recovery rate. *Entropy*, 19(7):305, 2017.

[20] T. De Ryck and S. Mishra. Numerical analysis of physics-informed neural networks and related models in physics-informed machine learning. *Acta Numerica*, 33:633,Äì713, 2024.

[21] O. Diekmann and J. A. P. Heesterbeek. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. Wiley, Chichester, UK, 2000.

[22] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.*, 28(4):365–382, 1990.

[23] G. Dimarco, B. Perthame, G. Toscani, and M. Zanella. Kinetic models for epidemic dynamics with social heterogeneity. *Journal of Mathematical Biology*, 83(1):4, 2021.

[24] N. Fadavi and N. Fadavi. Early recognition of parkinson's disease through acoustic analysis and machine learning. *World Journal of Advanced Research and Reviews*, 23(2).

[25] A. Fokas, N. Dikaios, and G. Kastis. Mathematical models and deep learning for predicting the number of individuals reported to be infected with SARS-CoV-2. *Journal of the Royal Society Interface*, 17(169):20200494, 2020.

[26] M. Gatto, E. Bertuzzo, L. Mari, S. Miccoli, L. Carraro, R. Casagrandi, and A. Rinaldo. Spread and dynamics of the COVID-19 epidemic in italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences of the United States of America*, 117(19):10484 ,Äì 10491, 2020.

[27] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[28] S. Han, L. Stelz, H. Stoecker, L. Wang, and K. Zhou. Approaching epidemiological dynamics of COVID-19 with physics-informed neural networks. *Journal of the Franklin Institute*, 361(6):106671, 2024.

[29] H. W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

[30] M. Jamshidi, A. Lalbakhsh, J. Talla, Z. Peroutka, F. Hadjilooei, P. Lalbakhsh, M. Jamshidi, L. La Spada, M. Mirmozafari, M. Dehghani, et al. Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment. *Ieee Access*, 8:109581–109595, 2020.

[31] S. Jin, Z. Ma, and K. Wu. Asymptotic-preserving neural networks for multiscale time-dependent linear transport equations. *Journal of Scientific Computing*, 94(3):57, 2023.

[32] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, 1927.

[33] V. La Gatta, V. Moscato, M. Postiglione, and G. Sperli. An epidemiological neural network exploiting dynamic graph structured data applied to the COVID-19 outbreak. *IEEE Transactions on Big Data*, 7(1):45–55, 2020.

[34] Z. Li, Y. Zheng, J. Xin, and G. Zhou. A recurrent neural network and differential equation based spatiotemporal infectious disease model with application to COVID-19. In *12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, volume 1, 2020.

[35] Z. Liao, P. Lan, X. Fan, B. Kelly, A. Innes, and Z. Liao. SIRVD-Dl: A COVID-19 deep learning prediction model based on time-dependent SIRVD. *Computers in Biology and Medicine*, 138:104868, 2021.

[36] D. Liu and A. Sopasakis. A combined neural ODE-bayesian optimization approach to resolve dynamics and estimate parameters for a modified sir model with immune memory. *Heliyon*, 10(19), 2024.

[37] W.-m. Liu, S. A. Levin, and Y. Iwasa. Influence of nonlinear incidence rates upon the behavior of SIRS epidemiological models. *Journal of mathematical biology*, 23:187–204, 1986.

[38] A. McInerney and K. Burke. A statistical modelling approach to feedforward neural network model selection. *Statistical Modelling*, page 1471082X241258261, 2024.

[39] K. Nabi, M. Tahmid, A. Rafi, M. Kader, and M. Haider. Comparative analysis of rnn and cnn for covid-19 forecasting. *Results in Physics*, 24:104137, 2021.

[40] K. N. Nabi, M. T. Tahmid, A. Rafi, M. E. Kader, and M. A. Haider. Forecasting COVID-19 cases: A comparative analysis between recurrent and convolutional neural networks. *Results in Physics*, 24:104137, 2021.

[41] K. I. Ouyoussef, J. El Karkri, L. M. Tine, and R. Aboulaich. Physics-informed neural networks for parameter estimation and simulation of a two-group epidemiological model. *International Journal of Modeling, Simulation & Scientific Computing*, 15(3), 2024.

[42] E. Pashaei and E. Pashaei. Training feedforward neural network using enhanced black hole algorithm: a case study on COVID-19 related ACE2 gene expression classification. *Arabian Journal for Science and Engineering*, 46:3807–3828, 2021.

[43] A. Puleio. Recurrent neural network ensemble, a new instrument for the prediction of infectious diseases. *The European Physical Journal Plus*, 136(3):319, 2021.

[44] V. S. H. Rao and M. N. Kumar. Estimation of the parameters of an infectious disease model using neural networks. *Nonlinear Analysis: Real World Applications*, 11(3):1810–1818, 2010.

[45] S. Ruan and W. Wang. Dynamical behavior of an epidemic model with a nonlinear incidence rate. *Journal of differential equations*, 188(1):135–163, 2003.

[46] Y. Tang, D. Huang, S. Ruan, and W. Zhang. Coexistence of limit cycles and homo-clinic loops in a SIRS model with a nonlinear incidence rate. *SIAM Journal on Applied Mathematics*, 69(2):621–639, 2008.

[47] L. Wang, X. Zheng, C. Sun, H. Li, and S. Zhang. Hybrid ai-epidemiology model for covid-19 forecasting: Combining machine learning and mechanistic models. *PLOS Computational Biology*, 17(5):e1009050, 2021.