

Article

Weekly Nowcasting of New COVID-19 Cases Using Past Viral Load Measurements

Athar Khalil ^{1,2,†} , Khalil Al Handawi ^{3,4,†} , Zeina Mohsen ⁵, Afif Abdel Nour ⁶, Rita Feghali ⁵, Ibrahim Chamseddine ^{7,*}  and Michael Kokkolaras ^{3,4,‡} 

¹ Department of Genetics and Genome Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH 44106, USA; axk1472@case.edu

² Clinical Research Unit, Rafik Hariri University Hospital, Beirut 2010, Lebanon

³ Group for Research in Decision Analysis (GERAD), Montréal, QC H3T 1J4, Canada; khalil.alhandawi@mail.mcgill.ca (K.A.H.); michael.kokkolaras@mcgill.ca (M.K.)

⁴ Systems Optimization Lab, Department of Mechanical Engineering, McGill University, Montréal, QC H3A 0G4, Canada

⁵ Department of Laboratory Medicine, Rafik Hariri University Hospital, Beirut 2010, Lebanon; zeinamohsen@outlook.com (Z.M.); rfeghali@ul.edu.lb (R.F.)

⁶ School of Engineering, The Holy Spirit University of Kaslik, Jounieh 446, Lebanon; afif.abdelnour@gmail.com

⁷ Department of Radiation Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

* Correspondence: ichamseddine@mgh.harvard.edu

† These authors contributed equally to this work.

‡ These authors contributed equally to this work.



Citation: Khalil, A.; Al Handawi, K.; Mohsen, Z.; Abdel Nour, A.; Feghali, R.; Chamseddine, I.; Kokkolaras, M. Weekly Nowcasting of New COVID-19 Cases Using Past Viral Load Measurements. *Viruses* **2022**, *14*, 1414. <https://doi.org/10.3390/v14071414>

Academic Editors: Luis Martinez-Sobrido and Fernando Almazan Toral

Received: 4 May 2022

Accepted: 20 June 2022

Published: 28 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: The rapid spread of the coronavirus disease COVID-19 has imposed clinical and financial burdens on hospitals and governments attempting to provide patients with medical care and implement disease-controlling policies. The transmissibility of the disease was shown to be correlated with the patient's viral load, which can be measured during testing using the cycle threshold (Ct). Previous models have utilized Ct to forecast the trajectory of the spread, which can provide valuable information to better allocate resources and change policies. However, these models combined other variables specific to medical institutions or came in the form of compartmental models that rely on epidemiological assumptions, all of which could impose prediction uncertainties. In this study, we overcome these limitations using data-driven modeling that utilizes Ct and previous number of cases, two institution-independent variables. We collected three groups of patients ($n = 6296$, $n = 3228$, and $n = 12,096$) from different time periods to train, validate, and independently validate the models. We used three machine learning algorithms and three deep learning algorithms that can model the temporal dynamic behavior of the number of cases. The endpoint was 7-week forward number of cases, and the prediction was evaluated using mean square error (MSE). The sequence-to-sequence model showed the best prediction during validation ($MSE = 0.025$), while polynomial regression (OLS) and support vector machine regression (SVR) had better performance during independent validation ($MSE = 0.1596$, and $MSE = 0.16754$, respectively), which exhibited better generalizability of the latter. The OLS and SVR models were used on a dataset from an external institution and showed promise in predicting COVID-19 incidences across institutions. These models may support clinical and logistic decision-making after prospective validation.

Keywords: COVID-19; deep neural networks; viral load; Ct values; predictive modeling; machine learning; now-casting; statistical analysis

1. Introduction

Coronavirus disease (COVID-19) was declared a pandemic by the World Health Organization (WHO) on March 2020 following the global spread of the underlying severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) [1,2]. The clinical outcomes of patients can range from an asymptomatic state to acute respiratory distress syndrome and

multi-organ dysfunction that can lead to death. The most identified risk factors are age, sex, ethnicity, smoking status and other comorbidities such as cardiovascular disorders, chronic kidney diseases and diabetes [2,3]. SARS-CoV-2 can be transmitted via direct contact; within a distance of one meter through coughing, talking, or sneezing; or indirectly via infectious secretions from infected patients [3]. COVID-19 put a strain on the economy and caused the general well-being of the population to diminish due to the public health and social measures (PHSMs) employed to control it [4]. Now-casting models are used to infer the epidemic trajectory and make informed decisions about its severity and necessary actions needed to bring the epidemic under control. Information regarding the origin of the pathogen, serological assays, and social behavior, among other aspects, are used to inform now-casting models and provide situational awareness to policymakers [5]. Several works in the literature have used epidemiological size indicators such as the frequency of tests, fatalities and new confirmed cases to infer the pandemic trajectory [6,7]. This paper focuses on both the epidemic size and serological assays from a cross-sectional sample of patients to develop a now-casting framework. Predictive modeling and now-casting of epidemic trajectories can alert policymakers and health institutions about an increase in incidence rates. This allows sufficient time to use other detailed scenario models to proactively test and deploy various PHSMs [8–11]. Therefore, the transmission rate depends on the patient's contagious stage, viral load, and the time of exposure between individuals [4].

RT-qPCR is a serological test that remains the gold standard for COVID-19 diagnosis [12]. It measures the first PCR cycle, denoted as the cycle threshold (Ct), at which a detectable signal of the targeted DNA appears [13]. The Ct value is inversely proportional to the viral load; a 3-point increase in Ct value equals a 10-fold decrease in the quantity of the virus' genetic material [14]. Ct values were proposed to have potential prognostic value in predicting severity, infectiousness, and mortality among patients [15]. Ct values were also used to determine the duration an infected patient needs to quarantine [16,17]. A high Ct value (indicating a low viral load) is detected at early stages of the infection before the person becomes contagious and at the late stages when the risk of transmission is low [18]. The lowest possible Ct value is usually reported within three days of the onset of symptoms and coincides with peak detection of cultivable virus and infectivity that implies an increase in transmissibility by up to 8-fold [19]. Using viral load measurements, individuals with high viral load and mild symptoms can be identified as potential super-spreaders [20]. Thus, early testing is highly recommended alongside isolation practices, to interrupt SARS-CoV-2 transmission [21].

Reverse-transcription quantitative polymerase chain reaction (RT-qPCR) is a serological test that remains the gold standard for COVID-19 diagnosis [12]. It measures the first PCR cycle, denoted as the cycle threshold (Ct), at which a detectable signal of the targeted DNA appears [13]. The Ct value is inversely related to the viral load; a 3-point increase in Ct value equals a 10-fold decrease in the quantity of the virus' genetic material [14]. Ct values were proposed to have potential prognostic value in predicting severity, infectiousness, and mortality among patients [15]. Ct values were also used to determine the duration an infected patient needs to quarantine [16,17]. A high Ct value (indicating a low viral load) is detected at the early stages of the infection before the person becomes contagious and at the late stages when the risk of transmission is low [18]. The lowest possible Ct value is usually reported within three days of the onset of symptoms and coincides with peak detection of cultivable virus and infectivity implies an increase in transmissibility by up to 8-fold [19]. Individuals with high viral load and mild symptoms can be identified as potential superspreaders using viral load measurements [20]. Thus, early testing is highly recommended alongside isolation practices to interrupt SARS-CoV-2 transmission [21].

We believe that the use of Ct for now-casting has its merits since it is a commonly available parameter irrespective of demographics and is highly correlated with transmissibility and incidence rates [22,23]. A popular approach for now-casting the pandemic trajectory is to use Bayesian inference frameworks to inform the posterior distributions for susceptible-exposed-infectious-recovered (SEIR) models and the corresponding time-varying incidence

rate [6,24]. These approaches are limited by the assumptions of the underlying SEIR models (homogeneous distribution of population traits and contacts). On the other hand, machine learning approaches make little to no assumptions about the underlying models describing the mechanics of the transmission and can potentially generalize better when the viral transmission is not completely understood and sufficient data is available.

We demonstrate the merits of this approach using a robust framework that leverages observed viral load measurements for time series now-casting of new COVID-19 cases for an upcoming 7-day time frame. The models are developed using a large cohort from a single cross-sectional virologic test center in Lebanon with a hold-out cohort for independent testing after the model is finalized. The Lebanese patient cohort used in this study is the largest and most consistent one in terms of serological assessment. This fact made the retrieved Ct values representative and reflective of the whole country.

Now-casting the pandemic trajectory can facilitate its containment and improves healthcare providers' preparedness against new SARS-CoV-2 variants and the surge in new cases caused by them. Furthermore, now-casting the pandemic trajectory can support policymakers during the decline phase of the pandemic (e.g., when vaccination rates are high and herd immunity is beginning to take hold) to suggest the best time frame for relaxing current PHSMs without the risk of the pandemic relapsing.

2. Materials and Methods

2.1. Patient Population

We retrospectively collected de-identified data for all COVID-19 patients diagnosed at Rafik Hariri University Hospital (RHUH) in Lebanon between 1 March 2020, and 31 March 2021. Rafik Hariri University Hospital (RHUH) is the country's leading institution for COVID-19 testing and treatment, and the collected cohort represents the nation's COVID-19 trajectory well [25]. Ct values were retrieved from the electronic medical database of the hospital, considering the date of the first positive RT-qPCR test for each patient while disregarding any subsequent positive tests that may have resulted during follow-up visits. RNA extraction and RT-qPCR processing protocols were consistent over time and the used PCR machines had similar calibration. The daily COVID-19 confirmed case counts in Lebanon were obtained from the Lebanese Ministry of Public Health and Worldometers website [26,27]. This study was approved by the Ethical Committee of RHUH. Written informed consent was waived since the study is retrospective and the patients' information was de-identified.

2.2. Study Design

We created 3 cohorts (discovery, testing, and independent validation) using a longitudinal data split. The discovery group (Group 1) was used for training and cross-validation [28] to tune the hyperparameters and calibrate the model weights. The testing group (Group 2) was reserved for testing the models' performance and calculating the test error. This approach complies with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) [29], which represents a classification criterion for predictive modeling. It has four types of increasing reliability. Since the data was split randomly into discovery (Group 1) and test groups (Group 2) at the beginning of the study, the model is a TRIPOD type 2b. We used a third portion of the data (Group 3) for further independent validation of the models developed using the discovery group (Group 1). This third group of data is called the 'unseen data group'. Such a longitudinal split of data enhances the model validation, making it between internal and external validation [29].

2.3. Predictive Modeling

We first identify the relevant input features needed by the models to predict epidemic trajectories in Lebanon using Spearman's correlation test. We analyzed the association between the patients' Ct values (Figure 1) and age (Figure S3a) with respect to the epidemic

trajectory and only selected the features with $p < 0.05$. Recent studies pointed out the case ascertainment rates may change over time (due to changes in PHSMs), resulting in biased Ct values [30]. The daily number of confirmed positive patients was plotted alongside the incidence rates in Lebanon to verify that this was not the case for the cohort used in this paper (see Section S1 and Figure S3b).

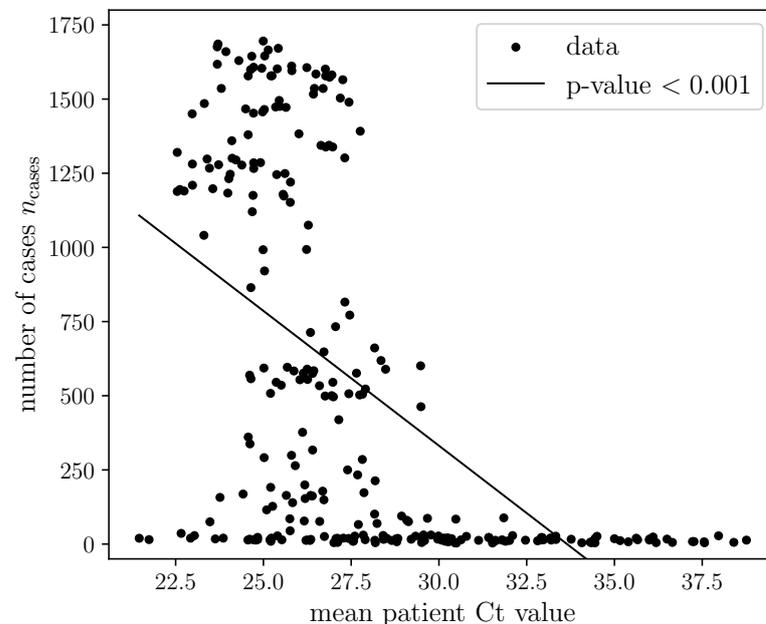


Figure 1. Scatter plot of biweekly mean Ct values and observed number of cases nationwide showing a clear negative value that is significant as given by p -value < 0.05 .

In addition to the previously mentioned features, the epidemic trajectory also depends on the past number of COVID-19 confirmed cases and is therefore aggregated with the input features during now-casting [5,7]. The period of time over which the input features and confirmed cases counts are aggregated is defined as the sliding window T_1 . The input to all models is therefore a sequence of data over the past T_1 days.

The epidemic trajectory is given by a sequence of predicted case counts in the upcoming T_2 days and is fixed to 7 days throughout the study in this paper. The window size of 7 days on the epidemiological calendar was chosen due to its clinical relevance to health providers. Furthermore, other studies based on cross-sectional virological data have used the 7-day window size for now-casting pandemic trajectories [24]. We developed 6 different machine learning algorithms for now-casting the epidemic trajectory, which are described below.

2.3.1. Recurrent Neural Network (RNN) Models

The first two models are built around recurrent neural networks (RNNs), which accommodate time-series data that are often temporarily correlated (i.e., the independent and identically distributed (i.i.d) assumption does not hold for time series data). This type of neural network can capture the temporal relationship between a decrease in Ct value and a subsequent (possibly delayed) rise in the number of cases. The RNN unit used in the models is the long short-term memory (LSTM) cell, which can capture long-term temporal effects and trends encoded by a long sequence of inputs and avoid the problem of vanishing gradients during backpropagation [31]. The LSTM has a cell for storing temporal data and gates to control data flow and capture long-term dependencies. Each gate is composed of a multilayer perceptron with n_{hidden} neurons [32]. We used stacked LSTM cells with several layers (given by n_{layers}) in the RNN models to learn high-level feature representations (the interaction of Ct values with the past number of cases) and

used a dropout probability P_{dropout} on all but the first layer to generalize better and avoid overfitting. Dropout arbitrarily excludes a number of hidden neurons from weight and bias updates during backpropagation to improve generalization performance [33]. Temporal information at time step t_i of the n -th layer LSTM cell is represented by its hidden $h_{t_i}^n$ and cell states $C_{t_i}^n$.

The first model is given by a sequence-to-sequence (S2S) model commonly used in natural language processing (NLP) translation tasks. The model consists of an encoder RNN that accepts an input sequence of features of length T_1 and yields a context vector $\mathbf{z}_n = [C_{t-1}^n \quad h_{t-1}^n]^T$, where $t - 1$ is the final time step of the input series. The context vector is fed to a decoder that outputs a predicted sequence of length T_2 corresponding to the projected number of cases n_{cases} . During training, the decoder uses its predictions $\hat{n}_{\text{cases}}^{t_i}$ at time step t_i as an input for the next the time step t_{i+1} . To speed up training, teacher forcing can be used to provide the actual value $n_{\text{cases}}^{t_i}$ at time step t_{i+1} instead of the decoder's prediction with a probability P_{teacher} [34]. The architecture of the S2S model used in this paper is shown in Figure 2.

We also developed a second RNN model based on the stacked LSTM cells alone (i.e., the size of the input sequence T_1 must be equal to the size of the output sequence T_2) (Figure S4a). This model is called the stacked LSTM (SEQ) model.

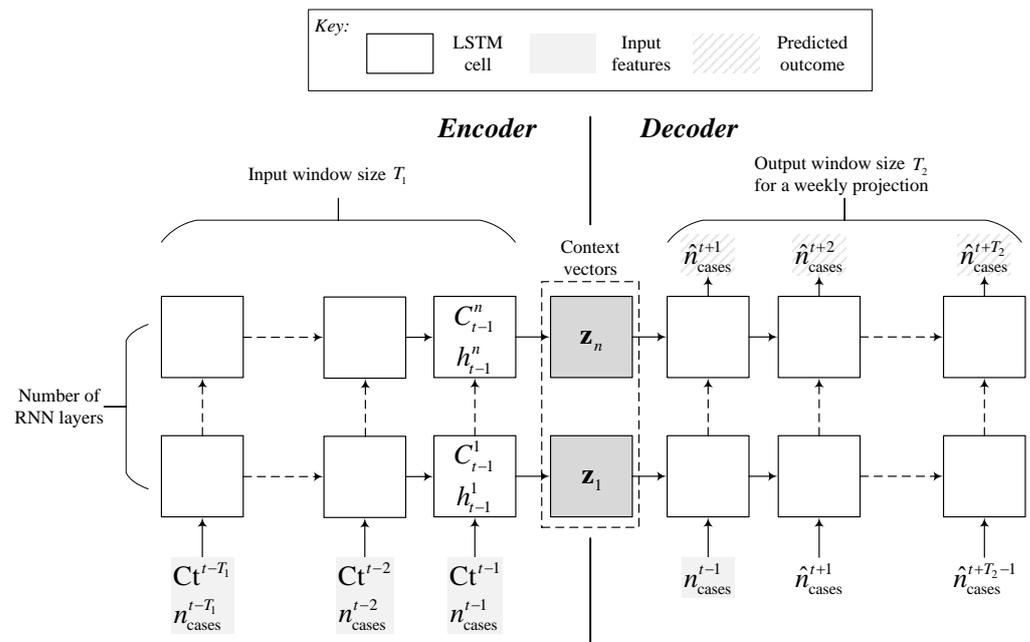


Figure 2. Structure of the sequence-to-sequence (S2S) model used for now-casting the weekly number of cases. The left side of the network is the encoder that uses past information on Ct and the number of cases to create context vectors used to initialize the hidden and cell states of the decoder LSTM cells.

The average number of predicted and actual number of cases for the next T_2 days is given by Equations (1) and (2), respectively

$$\hat{n}_{\text{cases}} = \frac{1}{T_2} \sum_{t_i=1}^{T_2} \hat{n}_{\text{cases}}^{t_i} \tag{1}$$

$$\bar{n}_{\text{cases}} = \frac{1}{T_2} \sum_{t_i=1}^{T_2} n_{\text{cases}}^{t_i} \tag{2}$$

2.3.2. Feedforward Neural Network (DNN) Model

We then developed a third model based on deep learning using feedforward neural networks. The feedforward neural network (DNN) model has several hidden layers (n_{layers})

with several hidden neurons (n_{hidden}) each. All layers had a dropout probability P_{dropout} and a rectified linear unit (ReLU) activation function (Figure S4b). All deep learning models were trained using the stochastic gradient descent algorithm ADAM with a learning rate l_{rate} and batch size b_{size} [35]. Early stopping was used on all deep learning models to avoid overfitting if no improvement in the validation error occurred after a certain number of epochs (given by the patience parameter n_{patience}) [36].

2.3.3. Regression Models

We developed three additional models that are not based on deep learning, namely a support vector machine regression (SVR) model [37], a gradient boosting machine (GBM) regression model [38], and a polynomial regression (OLS) model. Unlike deep learning models, these models do not yield a sequence of predictions for the next T_2 days. Instead, they compute a single value predicting the average number of confirmed COVID-19 cases for the next T_2 days (\hat{n}_{cases}). This is because such models are primarily used for regression of univariate functions. This allows for a fair comparison with the models described in Sections 2.3.1 and 2.3.2.

2.3.4. Hyperparameter Tuning

The hyperparameters of each model (listed in and described in Table 1) were optimized using cross-validation on the discovery group (Group 1) only. The cross-validation consists of outer and inner loops (Figure 3).

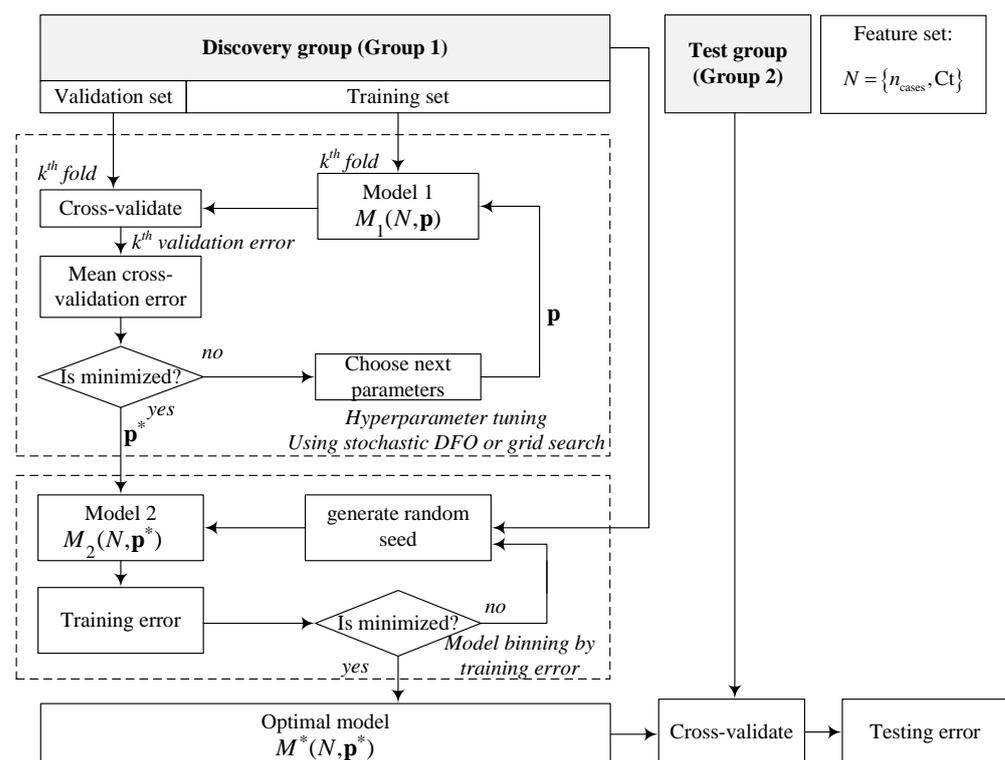


Figure 3. Cross-validation and hyperparameter determination scheme for model development. Following the discovery group (Group 1), the inner loop tuned the model's hyperparameters by minimizing the average k -fold cross-validation error using a stochastic direct search algorithm or a grid search. The second loop (following tuning) generates several models randomly and bins them by training error. The best model with the lowest training error is tested on the test group to obtain the testing error.

Table 1. Optimal hyperparameters of different models.

Hyperparameter	Symbol	Value	Possible Values
Sequence-to-sequence model (S2S)			
Sliding window size	T_1	6	1–40
Number of hidden neurons	n_{hidden}	1500	1–2500
Probability of dropout	P_{dropout}	0.8	0.0–0.9
Number of hidden layers	n_{hidden}	2	1–5
Teacher forcing probability	P_{teacher}	0.3	0.0–0.9
Learning rate	l_{rate}	1×10^{-4}	1×10^{-5} – 1×10^{-2}
batch size	b_{size}	32	4–128
best epoch	$n_{\text{epochs}}^{\text{best}}$	31	1– n_{epochs}
Sequence completion model (SEQ)			
Number of hidden neurons	n_{hidden}	2500	1–2500
Probability of dropout	P_{dropout}	0.8	0.0–0.9
Number of hidden layers	n_{hidden}	3	1–5
Learning rate	l_{rate}	1×10^{-4}	1×10^{-5} – 1×10^{-2}
batch size	b_{size}	64	4–128
best epoch	$n_{\text{epochs}}^{\text{best}}$	1	1– n_{epochs}
Deep neural network (DNN)			
Sliding window size	T_1	6	1–40
Number of hidden neurons	n_{hidden}	1000	1–2500
Probability of dropout	P_{dropout}	0.9	0.0–0.9
Number of hidden layers	n_{hidden}	1	1–5
Learning rate	l_{rate}	1×10^{-3}	1×10^{-5} – 1×10^{-2}
batch size	b_{size}	4	4–128
best epoch	$n_{\text{epochs}}^{\text{best}}$	4	1– n_{epochs}
Support vector machine regression (SVR)			
Sliding window size	T_1	6	1–40
Ridge factor	λ	1×10^{-4}	1×10^{-3} –1.0
Margin of tolerance	ϵ	1×10^{-2}	1×10^{-3} –1.0
Stopping criteria tolerance	ϵ_{tol}	0.1	1–5
Learning rate	l_{rate}	1×10^{-5}	1×10^{-5} – 1×10^{-2}
Gradient boosting machine (GBM)			
Sliding window size	T_1	36	1–40
Subsample fraction	f_{sample}	0.8	0.1–1.0
Maximum portion of features	f_{features}	0.1	0.1–1.0
Decision tree maximum depth	D	7	1–5
Learning rate	l_{rate}	0.01	1×10^{-5} – 1×10^{-2}
Maximum number of boosting stages	n_{stages}	5000	50–5000
Polynomial regression (OLS)			
Sliding window size	T_1	6	1–40
Ridge factor	λ	1.0	1×10^{-3} –1.0
Degree	n_{degree}	1	1–5
Common fixed parameters			
Output window size (all models)	T_2	7	1–40
Maximum number of epochs (all models)	n_{epochs}	5000	
Kernel (SVR)		linear	
Early stopping patience (S2S, SEQ, DNN)	n_{patience}	200	
Optimizer (S2S, SEQ, DNN)		Adam	

The tuned hyperparameters of each model are reported underneath it. The fixed hyperparameters are reported at the bottom of the table.

The outer loop split (Group 1) into five groups and sent four into the inner loop for training the models and subsequent hyperparameter optimization with respect to the average k-fold cross-validation error [39]. The cross-validation error of each fold was calculated using the mean squared error (MSE) criterion on the predicted and actual average number of cases for the following T_2 days as shown in Equation (3).

$$\text{MSE} = \frac{1}{n_{\text{days}}} \sum_{i=1}^{i=n_{\text{days}}} (n_{\text{cases}} - \hat{n}_{\text{cases}})^2, \quad (3)$$

where n_{cases} and \hat{n}_{cases} are defined by Equations (1) and (2), respectively.

Several models used in this paper (S2S, SEQ, DNN, and GBM) involve random variables associated with the training algorithm (backpropagation and gradient boosting), which are often ignored in the literature of applied machine learning. Examples of these random variables include the initial value of learnable parameters (weights, biases, and decision tree parameters), dropout, and gradient descent step sizes. Fixing the random seed of these random variables could result in model bias.

We address this issue by randomly sampling different training runs during hyperparameter optimization and optimizing the mean cross-validation errors of all the sampled runs. We apply this approach to a grid search on the hyperparameter space to discern the sensitivity of the cross-validation error to the hyperparameters. We then use a stochastic derivative-free optimization (DFO) algorithm (stochastic mesh adaptive direct search (StoMADS)) to fine-tune the hyperparameters [40]. StoMADS is an extension of the mesh adaptive direct search (MADS) algorithm that automatically updates its estimates of a stochastic objective function (in this paper, the objective function is given by the cross-validation error) depending on the level of uncertainty in the current incumbent solution.

After obtaining the optimal model in the internal loop, we scored it using the outer loop data. We then performed a random draw to obtain 30 models using the tuned hyperparameters. These models were binned by training error and the top-performing model was stored and used to make predictions for the test group (Group 2).

We note that binning and sampling of the cross-validation error are unnecessary for the OLS and SVR models since their training is deterministic and does not involve random variables.

3. Results

3.1. Patient Population

The entire dataset included 23,185 patients with a median age of 37 years. We aggregated the individuals' Ct into a sequence of daily mean Ct values. Group 1 contained 6296 patients admitted to RHUH between 2 March 2020 and 17 October 2020; Group 2 contained 3228 patients from 18 October 2020 to 30 November 2020, and the unseen group contained 12,097 patients from 1 December 2020 to 16 March 2021. All three groups have comparable median ages (34.0, 37.0, and 37.25 years, respectively). Group 1 was further split into five groups during model development for cross-validation: four training and one validation interchangeably.

Figure 4 shows the bi-weekly average Ct values observed and the corresponding number of cases in Lebanon nationwide for the period of time spanning groups 1 and 2 used in the model development phase. The entire dataset is provided in the Supplementary Material, Figure S1).

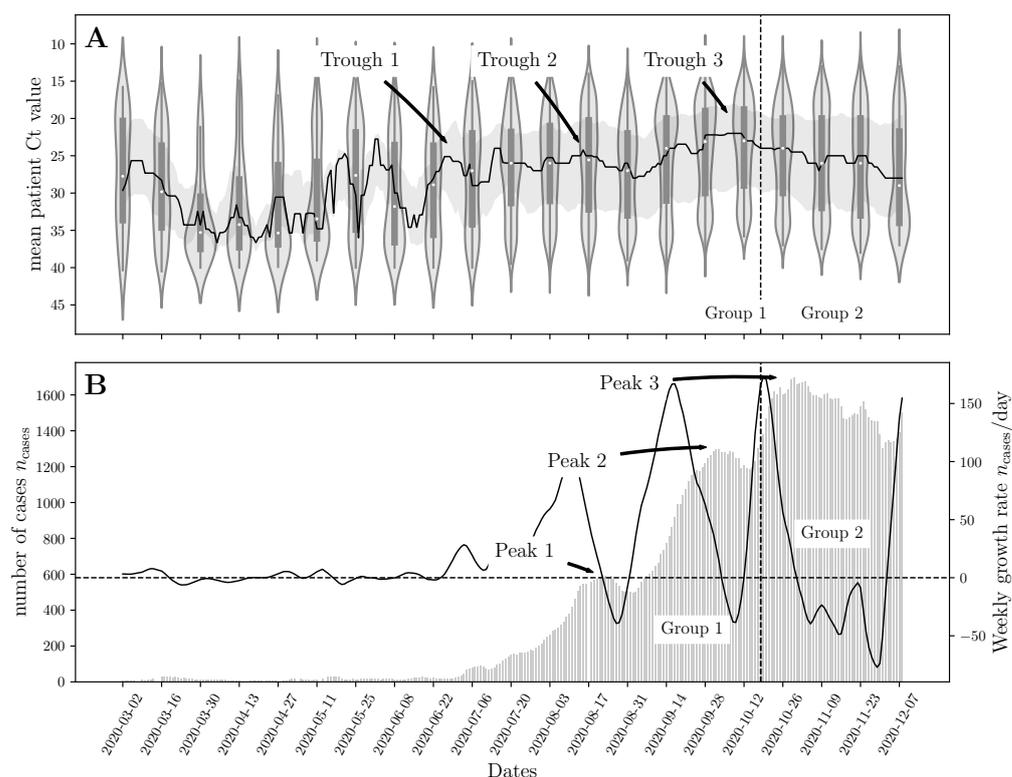


Figure 4. (A) Bi-weekly mean Ct values of RHUH patients. The solid line represents the median bi-weekly Ct values, and the gray shaded area represents the inter-quartile range (25–75 percentile) of the observed Ct values. (B) The grey bars show the weekly running average of the number of cases observed nationwide in Lebanon between 1 March 2020, and 7 December 2020 (the running average can be computed until 23 November). The solid black line represents the growth rate in the weekly number of cases.

3.2. Correlation between the National Daily Number of COVID-19 Cases and Mean Ct

We observed a temporal delay between the incidence rate and the observed Ct values. For example, the trough in mean Ct values on 8 October 2020 (Trough 3 in Figure 4A) was followed by an increase in the number of cases, on 29 October 2020, with more than 1640 cases per day (Peak 3 in Figure 4B). This delay could be due to the time needed for population dynamics of disease transmission to take hold. Low Ct values indicate nascent infections circulating in the population that need time to reach the rest of the population. This observation has been reported by Hay et al. [24] using compartmental SEIR models to show that cross-sectional Ct observations with a low median value signal the growth phase of a pandemic (when case counts are still typically low). A similar trend was observed for case count peaks 1 and 2, which were superseded by median Ct troughs 1 and 2, respectively. This visual analysis of the data indicates that the median Ct value is temporally related to incidence.

We also investigate the relationship between the median Ct value and case counts using correlation analysis. We observed a clear inverse correlation between mean Ct and number of cases ($p < 0.001$), quantified by the Spearman correlation test (Figure 1). This indicates that the mean cross-sectional Ct value is an important feature for now-casting the pandemic trajectory.

3.3. Now-Casting the Epidemic Trajectories

We developed six types of predictive models for now-casting the COVID-19 epidemic trajectory in Lebanon using the data in the discovery group (Group 1). The optimal hyperparameters for each model are listed in Table 1. Early stopping terminated the

backpropagation algorithm at 31, 1, and 4 epochs for the S2S, SEQ, and DNN models, respectively. All models except the GBM had an optimal input window size T_1 of 6 days. This implies that an aggregate measure of cross-sectional Ct values and past incidence rates over the last 6 days could be used to now-cast the expected number of positive COVID-19 cases over the following 7 days. The models developed using Group 1 were used to now-cast the trajectory from 18 October 2020 to 30 November 2020 (Group 2) (Figure 5). The models were then retrained in Groups 1 and 2 using the hyperparameters in Table 1 and used to now-cast the epidemic trajectory after 1 December 2020 (Group 3). Table 2 lists the MSE error for the predicted trajectories on Groups 2 and 3 (see Table 2 footnote).

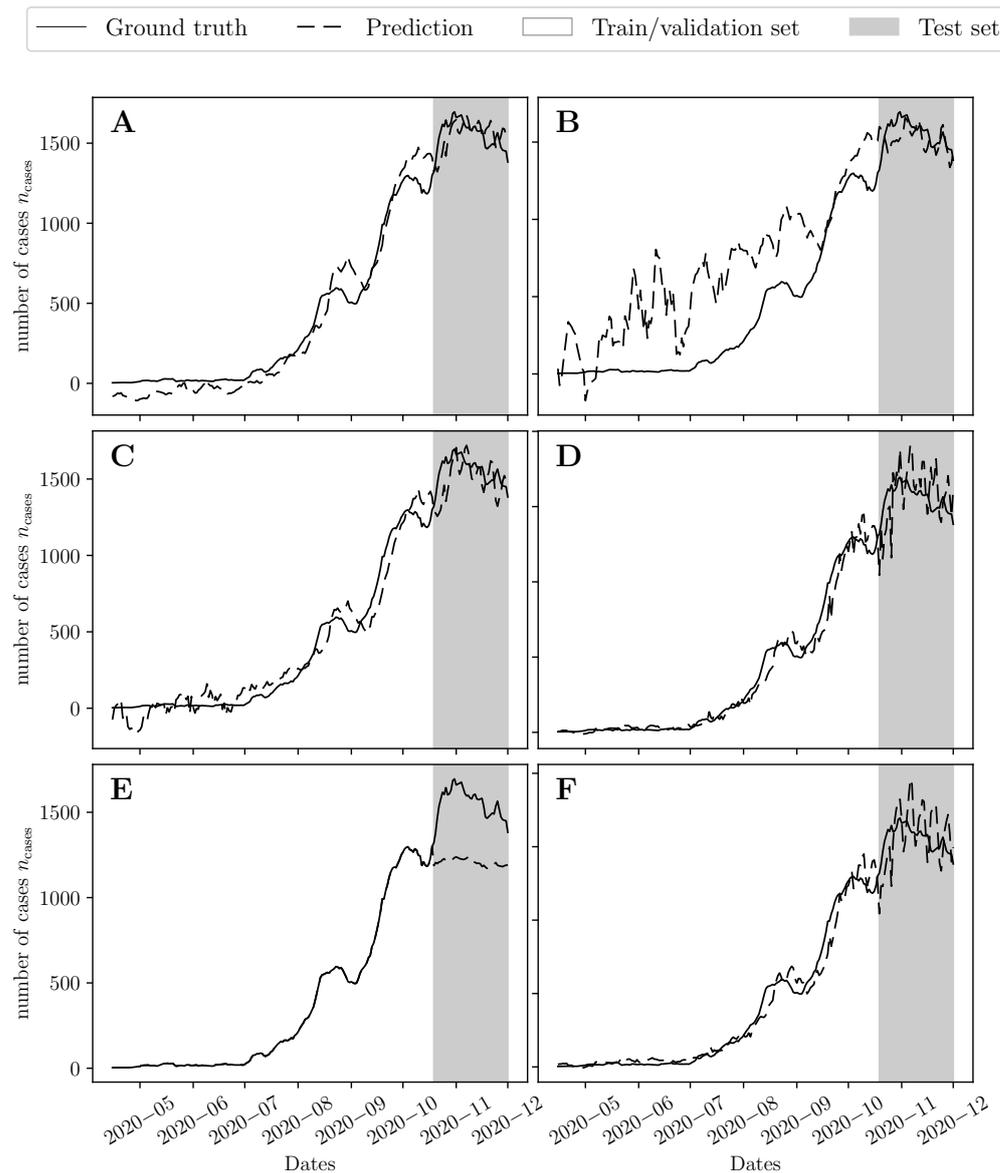


Figure 5. Predicted 7-day rolling average of daily number of cases on the unseen data set using (A) the sequence-to-sequence (S2S) model, (B) the stacked LSTM (SEQ), (C) The feedforward neural network (DNN), (D) The support vector machine regression (SVR) model, (E) The gradient boosting machine (GBM), and (F) the polynomial regression (OLS) model. All models were tuned using the cross-validation error of the discovery set. The grey shaded region represents the test data set (Group 2) used to test the models' performance.

Table 2. Training and testing errors given by mean squared error (MSE) of different models constructed using different feature sets.

Model	Figure 5		Figure 6	
	Train Error	Test Error	Train Error	Unseen Error
	Group 1	Group 2	Groups 1,2	Group 3
Sequence-to-sequence (S2S)	0.02462	0.02504	0.01309	0.57112
Stacked LSTM (SEQ)	0.38373	0.02724	0.78142	0.32584
Feedforward neural network (DNN)	0.02223	0.04179	0.00919	0.25547
Support vector machine regression (SVR)	0.01362	0.08347	0.00518	0.16754
Gradient boosting machine (GBM)	2.316×10^{-6}	0.32589	2.316×10^{-6}	1.44463
Polynomial regression (OLS)	0.01335	0.08954	0.00459	0.15954

The MSE in Equation (3) is computed using the standardized value of the predictions by normalizing them using the mean and standard deviation of all the daily number of cases (n_{cases}) given by 463.8 and 597.0, respectively.

The RNN models (S2S and SEQ) performed well on Group 2 (MSE of 0.025 and 0.027, respectively), followed by the DNN model with an MSE that is two-fold larger (0.042). The OLS and SVR had an MSE that is four-fold larger than that of the RNN models (0.090 and 0.083, respectively). The GBM was heavily biased and did not generalize well on Group 2 (MSE of 0.326). The training error for the RNN models was higher than that of the parametric models (OLS and SVR) due to the regularization performed by the early-stopping criterion to avoid overfitting. Movie S1 shows an example training run of the S2S model with arbitrary hyperparameters, where early stopping helped avoid overfitting.

The MSE error of the SVR, OLS, and DNN models was comparable on the unseen data group (MSE values of 0.168, and 0.160, and 0.255, respectively). The SEQ and S2S performed worse on the unseen group, implying that simpler models perform better on the unseen group due to the limited number of data points available for training and hyperparameter tuning. Deep learning models generally excel when a large dataset is available for model development and has been reported by several studies in the literature [41,42].

To verify this, the RNN models were re-developed using both Groups 1 and 2 for training, hyperparameter tuning and validation. The generalization performance improved significantly, bringing the MSE error down from 0.571 to 0.106 for the S2S model (Table S1). This implies that the RNN models generalize better when more training data is available (see Supplementary Material Section S3). If limited data are available (at the start of a pandemic), simpler models can provide better generalization performance. We deployed the models developed on the combined dataset (Groups 1 and 2) as a web application to facilitate prospective validation in the future [43].

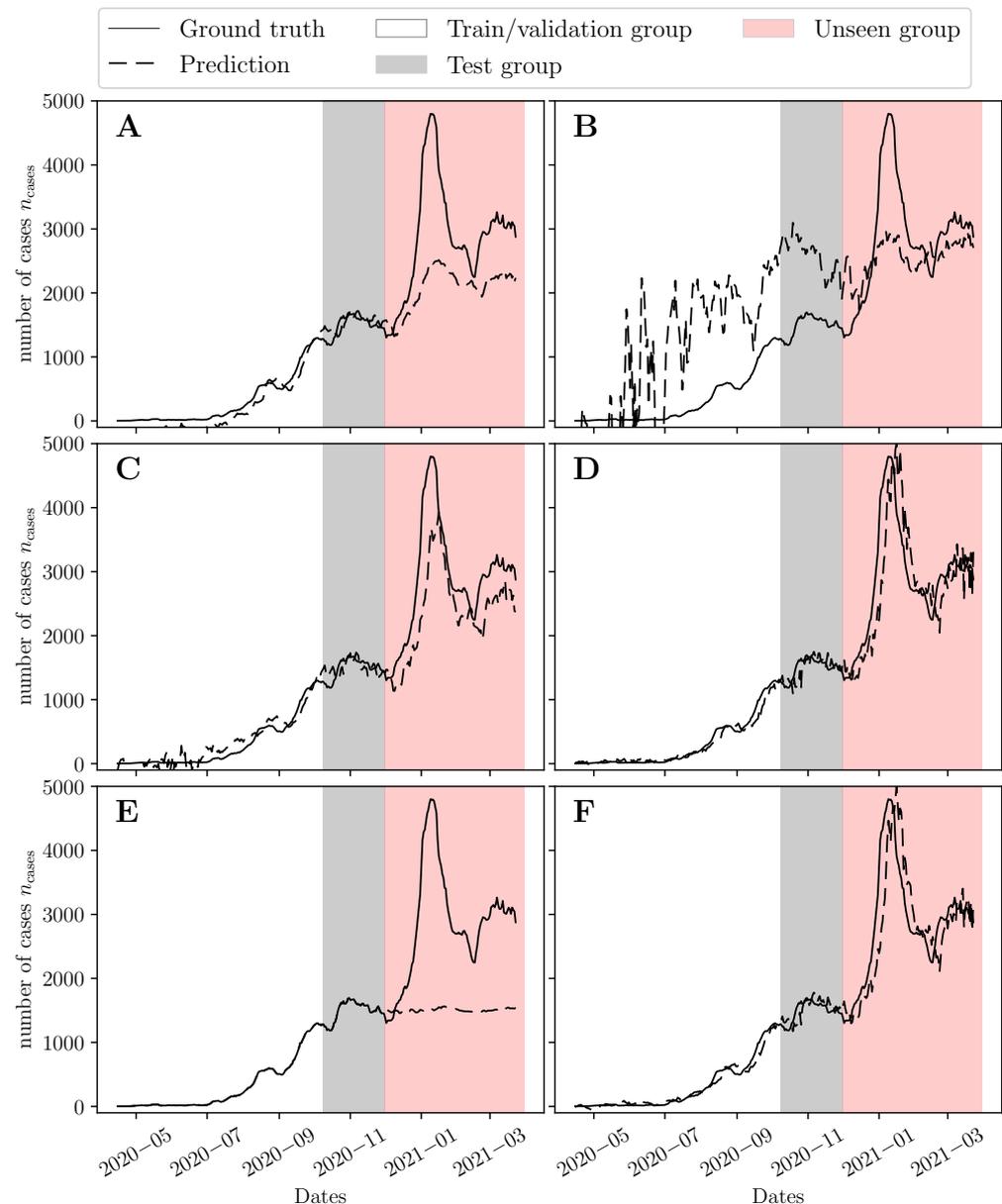


Figure 6. Predicted 7-day rolling average of daily number of cases on the unseen data set using (A) the sequence-to-sequence (S2S) model, (B) the stacked LSTM (SEQ), (C) The feedforward neural network (DNN), (D) The support vector machine regression (SVR) model, (E) The gradient boosting machine (GBM), and (F) the polynomial regression (OLS) model. All models were tuned using the validation error of the discovery set. The grey shaded region represents the test data set (Group 2) used to test the models' performance. The models were retrained using both the discovery and test data sets and subsequently used to infer the number of cases in the unseen data set (the red shaded region).

4. Discussion

Host viral load and the resultant Ct values have been widely proposed to evaluate the progression of SARS-CoV-2 infection and address patients' contagiousness [44]. Mathematical modeling has been widely used for predicting the course of the COVID-19 pandemic. These prediction models were developed based on the applied intervention measurements and the population behavioral fluctuations, including social distancing and mask-wearing [45]. The COVID-19 reproduction number (R_0), defined as the average number of naive individuals a patient can infect, has a mean estimate of 3.28 and could

range from 1.4 to 6.49 [46]. Although R_0 can widely vary by country, culture, and stage of the outbreak, it has been used to justify the need for community mitigation strategies and political interventions [47]. So far, only few advanced and more recent models have evaluated the disease spread based on viral kinetics and serological assays (such as RT-qPCR tests) [22,24]. Furthermore, these studies focused on serological assays or pandemic size indicators (such as R_0 and incidence rates) in isolation without combining the two. This paper utilizes both past incidence rates and serological viral load measurements to now-cast the pandemic trajectory.

Hay et al. [24] used Bayesian inference to predict the growth rate in the daily number of COVID-19 cases as a function of Ct values [24]. They showed that the population-level Ct distribution is strongly correlated with growth rate estimates of new infections in MA, USA. They estimated R_0 and growth rate by using observations of Ct values to inform priors on key viral kinetics parameters (such as the viral load wane rate, and Ct at peak viral load and the pandemic trajectory (daily probability of infection is used as a proxy for the trajectory)). The prior on the pandemic trajectory is assumed to come from a Gaussian process that makes no assumptions regarding the evolution of the trajectory as more Ct observations are made. We have used the Gaussian process regression model to predict the pandemic trajectory using the RHUH cross-sectional patient cohort (see Supplementary Material Section S4). The advantage of such models is that they are highly interpretable as they estimate the viral kinetics model parameters that are most likely to give rise to the observed Ct values [48]. This provides useful information about the virulence and severity of the pathogen. However, such models make assumptions about the likelihood used to update the priors. These assumptions limit the predictive capability of the model if any of these assumptions (such as the viral kinetics models) do not hold in reality, potentially resulting in poor generalization performance. This is the case when a different clade of viruses takes hold. Another dataset from Bahrain demonstrated the effectiveness of Ct in predicting the epidemiological dynamics of COVID-19 [23]. However, the study did not consider the interaction between different features (i.e., number of positive cases and Ct), nor does it consider temporal effects observed in epidemics.

In comparison, the presented data-driven approach of inferring the epidemic trajectory using past case counts and Ct values makes very little assumptions about the pandemic trajectory and viral kinetics models that gave rise to the observed Ct values. This has the benefit of potentially generalizing to a wide range of scenarios. To prove this, we used all the models developed in this paper using Group 1 (Figure 5) to infer the case counts in the state of Massachusetts using the patient cohort of Brigham and Women's Hospital (BWH) provided by Hay et al. [24] (Figure S9A). Most models captured the underlying trend except for GBM and the stacked LSTM (SEQ) models (Figure S10). SVR performed the best on this dataset (Figure S9B). However, further prospective validation is needed in the future to ensure that these models can generalize to different testing centers and reject disturbances in Ct values due to sample collection and handling methods.

The inferred trajectories for the state of Massachusetts (15 April 2020–15 December 2020) and Lebanon (1 December 2020–31 March 2021) show that simple machine learning models (such as OLS and SVR) perform well with limited training data (when developing the models using data from Group 1 only). Deep learning models begin to outperform such models when including more data in the development set (Groups 1 and 2) to infer the trajectory in Lebanon (Figure S5). Although the outcomes of this study favored simpler regression models, their simplicity provides an advantage in terms of interpretability [48].

The dataset used in this study contained fluctuations that allowed us to extract the Ct temporal effect on the trajectory of the pandemic. Since the data came from a single institution, the fluctuations are likely to be signals in the data rather than noise. The significant changes in the Ct values mirrored the well-recognized political, economic, and social turning points that happened in Lebanon during the pandemic. These incidences impacted the population's behavior towards COVID-19 in a consistent and well-defined manner, allowing us to track and correlate these changes with the variation in the mean Ct values

and subsequently the disease spread. The early reported high mean Ct values and the low number of COVID-19 cases in Lebanon between March 2020 and June 2020 co-occurred with a strictly imposed lockdown and a harsh awareness campaign executed by local media platforms [25]. In comparison, the sharp rise in COVID-19 cases and the decrease in mean Ct values upon diagnosis were detected after releasing the first national lockdown in July, which occurred with a significant shifting of local media attention towards the economic crisis peaking in the country. Yet, the highest jump in the number of national COVID-19 patients and the sharpest drop in Ct values were reported after the explosion of Beirut's port in August 2020, which was classified among the most significant chemical explosions in history [49]. The devastating effects of the explosion amplified the country's pre-existing social, economic, and health challenges, causing a significant increase in the COVID-19 positivity rate in September and November 2020, which had reached 13.9% [49,50]. The consequences of this explosion shifted the residents' attention away from proper precautions. This was reflected by the sharp decrease in the mean Ct values indicating less responsible behavior and a delay in diagnosis time among suspected patients, which resulted in a subsequent increase in SAR-CoV-2 spread among individuals. These events caused three significant peaks in the number of cases and three drops in mean Ct. We trained the models on two of these peaks and tested its ability to detect the third peak using the unseen data. Thus, the training and validation errors reflect the models' robustness against unexpected events.

The detected inverse relationship between Ct values and the number of national COVID-19 positive cases reflects population dynamics of transmission and demonstrates the temporal significance of Ct values. The results emphasized the importance of early testing when the patient's viral load and infectivity are low to prompt isolation practices and thus, suppress the national spread of the virus. The models were able to predict the upcoming one-week expected number of national COVID-19 cases based on a commonly available diagnostic measurement, the Ct value. This shows that viral load measurements are a rigid input that can enhance the outcomes of disease forecasting models. Interestingly, this model is still valuable among vaccinated patients as these patients were shown to have a similar viral load pattern as unvaccinated patients and thus, can efficiently transmit the disease in the same manner upon infection [51]. Ultimately, the data promoted incorporating Ct values with other epidemiological variables and patient demographics to predict new COVID-19 waves and to study epidemic behaviors. The models in this paper could be extended to now-cast other contagious viral diseases that are diagnosed by qPCR, provided that sufficient training data is available (at least one wave of the viral disease has been observed).

This study is limited to a single-institution cohort. Although the cohort represents the national number of cases, and the model's variable (Ct) is country-independent, a prospective validation on multi-institutional data is needed before translation. To facilitate this process, we have hosted the models on a web interface to be used in future studies that compare the predicted and observed number of cases [43]. Another limitation is the inability of the model to compare the effect of preventative policies such as lockdowns and quarantining. The model provides an alert when the number of cases is about to rise significantly, allowing more informed triage decisions and better allocation of medical resources during the pandemic. However, it does not provide guidance on what measures could best control an upcoming peak. Mechanistic models, on the other hand, such as individual-based models (IBMs), can provide such insights but their application is limited to a much smaller population size due to computational cost [52–55]. A future study could focus on combining IBMs with viral load models such as those developed by Hay et al. [24] to estimate Ct values for a cross-section of the population and use them to retrain the models developed in this paper to now-cast the trajectory under different intervention policies [24].

5. Conclusions

This study is a first attempt at combining serologic assays from a representative cross-sectional patient cohort with epidemiological indicators such as incidence rates and infection size to now-cast the number of nationwide positive COVID-19 cases in a specific region [5]. This was motivated the premise that SARS-CoV-2 spread is highly dependent on the individual viral dynamics. The models used in this paper showed the merits of this approach using observations of Ct values and historical infection data from Lebanon in a now-casting framework. The modeling framework relied on multiple machine learning algorithms that make little assumptions about population and transmission dynamics. The patient cohort revealed that the evolution of the viral load mirrored the growth of positive national cases in the country. Low mean Ct values were followed by a large number of national positive COVID-19 cases and vice versa, in line with similar observations in the literature [22,24]. This finding is also supported by applying the machine learning models in this paper to the BWH dataset provided by Hay et al. [24]. To account for the effect of social interactions that could occur a few days before and after testing, we used a sequence of daily mean Ct values across multiple machine learning algorithms. We trained the models on a training dataset and independently validated them on unseen data forming TRIPOD type 2b models [29]. The training process utilized a cross-validation approach combined with a state-of-the-art stochastic direct search for hyperparameter tuning to prevent model over-fitting [56]. The sequence-to-sequence (S2S) model had the best accuracy when a large amount of data was used for its development, while the support vector machine regression (SVR) model provided better accuracy with limited development data as given by the MSE criterion. Since the models were trained and validated on datasets from different time periods, they have the potential to extend to future data. In addition, since the variables used for prediction (Ct values) are not specific to the institution from which the data were acquired, the models are ready to undergo a prospective and external validation in the future. This will form a TRIPOD 4 study, which is recommended to translate the model to practice to now-cast a 7-day forward number of cases based on recently reported Ct values.

Supplementary Materials: The following are available at <https://www.mdpi.com/article/10.3390/v14071414/s1>, Figure S1: (A) Bi-weekly mean Ct values of RHUH patients. The solid line represents the median bi-weekly Ct values, and the gray shaded area represents the inter-quartile range (25–75 percentile) of the observed Ct values. (B) The grey bars show the weekly running average of the number of cases observed nationwide in Lebanon between 1 March 2020, and 31 March 2021 (the running average can be computed until March 16). The solid black line represents the growth rate in the weekly number of cases. Figure S2: (A) Bi-weekly mean age of RHUH patients. The solid line represents the median patient age, and the gray shaded area represents the inter-quartile range (25–75 percentile) of the observed patient ages. (B) Bi-weekly mean number of confirmed positive RHUH patients. (C) The grey bars show the weekly running average of the number of cases observed nationwide in Lebanon between 1 March 2020, and 30 December 2020 (the running average can be computed until December 07). Figure S3: (A) Scatter plot of biweekly mean patient age and observed number of cases nationwide showing no significant relationship as given by p -value > 0.05 . (B) Scatter plot of biweekly number of confirmed positive RHUH patients and observed number of cases nationwide showing a clear positive value that is significant as given by p -value < 0.05 . Figure S4: Model architecture of the (a) stacked LSTM (SEQ) and the (b) feedforward neural network (DNN) models. Figure S5: Predicted 7-day rolling average of daily number of cases on the unseen data group using (A) the sequence-to-sequence (S2S) model, (B) the stacked LSTM (SEQ), (C) The feedforward neural network (DNN), (D) The support vector machine regression (SVR) model, (E) The gradient boosting machine (GBM), and (F) the polynomial regression (OLS) model. All models were tuned using the validation score of the combined discovery and test sets (Groups 1 and 2). The grey shaded region represents the unseen data group used to test the models' performance. Figure S6: Illustration of variance in (a) training errors on the discovery group (Group 1) and (b) test errors on the test group (Group 2) for different models. The errors were calculated using the MSE of the predicted and actual trajectories shown in Figure 5. The green triangles represent the mean error of 30 independent training runs for each model type. The orange lines represent the

median error. Figure S7: Illustration of variance in (a) training errors on combined discovery and test groups (Groups 1 and 2) and (b) the test errors on the unseen group (Group 3) for different models. The errors were calculated using the MSE of the predicted and actual trajectories shown in Figure S5. The green triangles represent the mean error of 30 independent training runs for each model type. The orange lines represent the median error. Figure S8: Incidence rate and pandemic trajectory predictions using the predictive framework developed by Hay et al. [24] (A) shows the cross-sectional Ct samples (violin plots) and smoothed average (solid blue line) obtained from RHUH throughout the pandemic in Lebanon. (B) Posterior distribution of relative probability of infection by date from a Gaussian process (GP) model fit to all observed Ct values (ribbons show 95% and 50% credible intervals, line shows posterior median). The y-axis shows relative rather than absolute probability of infection, as the underlying incidence curve must sum to one. The grey bars show the true case counts in Lebanon from the start of infection and have been normalized by the total number of cases observed in Lebanon throughout the observation time period shown (1 March 2020 through 30 November 2020). Figure S9: Incidence rate and pandemic trajectory predictions using the support vector machine regression (SVR) model (A) shows the cross-sectional Ct samples (violin plots) and smoothed average (solid blue line) obtained from Brigham and Women's Hospital (BWH) throughout the pandemic in Massachusetts. (B) Predicted pandemic trajectory of the SVR model fit to all observed Ct values. The grey bars show the true case counts in Massachusetts from the start of infection. Figure S10: Predicted 7-day rolling average of the daily number of cases in Massachusetts predicted using (A) the sequence-to-sequence (S2S) model, (B) the stacked LSTM (SEQ), (C) The feedforward neural network (DNN), (D) The support vector machine regression (SVR) model, (E) The gradient boosting machine (GBM), and (F) the polynomial regression (OLS) model. The Ct values used in inference were obtained from Brigham and Women's Hospital (BWH) [24]. Table S1: Training and testing errors given by mean squared error (MSE) of different models constructed using the combined discovery and test sets (Groups 1 and 2). Table S2: Optimal hyperparameters of models developed using combined discovery and test groups (Groups 1 and 2).

Author Contributions: Conceptualization of the work: A.K., K.A.H. and A.A.N.; Resources provided by: A.K., I.C. and M.K.; Data curation: Z.M., K.A.H. and A.K.; Software development: K.A.H.; Formal analysis of the data: K.A.H. and I.C.; Supervision by: I.C. and M.K.; Validation of methods: K.A.H.; Investigation of different modeling strategies: K.A.H. and I.C.; Visualization: K.A.H.; Methodology: K.A.H. and I.C.; Drafting of the manuscript: K.A.H. and A.K.; Project administration: R.F. and I.C.; Review and editing of manuscript: K.A.H., A.K., I.C. and M.K. All authors consent to be held accountable for all aspects of work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The studies involving human participants were reviewed and approved by the Ethical Committee of RHUH. The title of the approved study is "Using Ct Values to predict COVID-19 numbers" and the date of approval was 3 March 2021. The ethics committee waived the requirement of written informed consent for participation.

Acknowledgments: The authors would like to thank the general manager of Rafik Hariri University Hospital, Firass Abiad, for his continuous support.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. World Health Organization Declares Global Emergency: A Review of the 2019 Novel Coronavirus (COVID-19). *Int. J. Surg.* **2020**, *76*, 71–76. [[CrossRef](#)] [[PubMed](#)]
2. Khalil, A.; Kamar, A.; Nemer, G. Thalidomide-Revisited: Are COVID-19 Patients Going to Be the Latest Victims of Yet Another Theoretical Drug-Repurposing? *Front. Immunol.* **2020**, *11*, 1248. [[CrossRef](#)] [[PubMed](#)]
3. Rabaan, A.A.; Al-Ahmed, S.H.; Al-Malkey, M.; Alsubki, R.; Ezzikouri, S.; Hassan Al-Hababi, F.; Sah, R.; Mutair, A.A.; Alhumaid, S.; Al-Tawfiq, J.A.; et al. Airborne transmission of SARS-CoV-2 is the dominant route of transmission: Droplets and aerosols. *Infez. Med.* **2021**, *29*, 10–19. [[PubMed](#)]
4. Hoertel, N.; Blachier, M.; Blanco, C.; Olsson, M.; Massetti, M.; Rico, M.S.; Limosin, F.; Leleu, H. A stochastic agent-based model of the SARS-CoV-2 epidemic in France. *Nat. Med.* **2020**, *26*, 1417–1421. [[CrossRef](#)] [[PubMed](#)]
5. Wu, J.T.; Leung, K.; Lam, L.T.; Ni, M.Y.; Wong, C.K.; Peiris, J.S.; Leung, G.M. Nowcasting epidemics of novel pathogens: Lessons from COVID-19. *Nat. Med.* **2021**, *27*, 388–395. [[CrossRef](#)]

6. Irons, N.J.; Raftery, A.E. Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2103272118. [[CrossRef](#)]
7. Alaimo Di Loro, P.; Divino, F.; Farcomeni, A.; Jona Lasinio, G.; Lovison, G.; Maruotti, A.; Mingione, M. Nowcasting COVID-19 incidence indicators during the Italian first outbreak. *Stat. Med.* **2021**, *40*, 3843–3864. [[CrossRef](#)]
8. Kamar, A.A.; Maalouf, N.; Hitti, E.; El Eid, G.; Isma'eel, H.; Elhajj, I.H. The Challenge of Forecasting Demand of Medical Resources and Supplies during a Pandemic: A Comparative Evaluation of Three Surge Calculators for COVID-19. *Epidemiol. Infect.* **2021**, *149*, e51. [[CrossRef](#)]
9. Abrams, S.; Wambua, J.; Santermans, E.; Willem, L.; Kuylen, E.; Coletti, P.; Libin, P.; Faes, C.; Petrof, O.; Herzog, S.A.; et al. Modelling the early phase of the Belgian COVID-19 epidemic using a stochastic compartmental model and studying its implied future trajectories. *Epidemics* **2021**, *35*, 100449. [[CrossRef](#)]
10. Reiner, R.C. Modeling COVID-19 scenarios for the United States. *Nat. Med.* **2021**, *27*, 94–105. [[CrossRef](#)]
11. Pinto Neto, O.; Kennedy, D.M.; Reis, J.C.; Wang, Y.; Brizzi, A.B.; Zambrano, G.J.; de Souza, J.M.; Pedroso, W.; de Mello Pedreiro, R.C.; de Matos Brizzi, B.; et al. Mathematical model of COVID-19 intervention scenarios for São Paulo—Brazil. *Nat. Commun.* **2021**, *12*, 1–13. [[CrossRef](#)] [[PubMed](#)]
12. Péré, H.; Podglajen, I.; Wack, M.; Flamarion, E.; Mirault, T.; Goudot, G.; Hauw-Berlemont, C.; Le, L.; Caudron, E.; Carrabin, S.; et al. Nasal swab sampling for SARS-CoV-2: A convenient alternative in times of nasopharyngeal swab shortage. *J. Clin. Microbiol.* **2020**, *58*, e00721-20. [[CrossRef](#)] [[PubMed](#)]
13. Ade, C.; Pum, J.; Abele, I.; Raggub, L.; Bockmühl, D.; Zöllner, B. Analysis of cycle threshold values in SARS-CoV-2-PCR in a long-term study. *J. Clin. Virol.* **2021**, *138*, 104791. [[CrossRef](#)] [[PubMed](#)]
14. *Understanding Cycle Threshold (Ct) in SARS-CoV-2 RT-PCR A Guide for Health Protection Teams Understanding Cycle Threshold (Ct) in SARS-CoV-2 RT-PCR 2*; Technical Report; Public Health England: London, UK, 2020.
15. Rao, S.N.; Manissero, S.; Steele, V.R.; Pareja, J. A Narrative Systematic Review of the Clinical Utility of Cycle Threshold Values in the Context of COVID-19. *Infect. Dis. Ther.* **2020**, *9*, 573–586. [[CrossRef](#)] [[PubMed](#)]
16. Rodríguez-Grande, C.; Catalán, P.; Alcalá, L.; Buenestado-Serrano, S.; Adán-Jiménez, J.; Rodríguez-Maus, S.; Herranz, M.; Sicilia, J.; Acosta, F.; Pérez-Lago, L.; et al. Different dynamics of mean SARS-CoV-2 RT-PCR Ct values between the first and second COVID-19 waves in the Madrid population. *Transbound. Emerg. Dis.* **2021**, *68*, 3103–3106. [[CrossRef](#)]
17. Miller, E.H.; Zucker, J.; Castor, D.; Annajhala, M.K.; Sepulveda, J.L.; Green, D.A.; Whittier, S.; Scherer, M.; Medrano, N.; Sobieszczyk, M.E.; et al. Pretest Symptom Duration and Cycle Threshold Values for Severe Acute Respiratory Syndrome Coronavirus 2 Reverse-Transcription Polymerase Chain Reaction Predict Coronavirus Disease 2019 Mortality. *Open Forum Infect. Dis.* **2021**, *8*, ofab003. [[CrossRef](#)]
18. *COVID-19: Management of Staff and Exposed Patients or Residents in Health and Social Care Settings*; Technical Report; UK Health Security Agency: London, UK, 2022.
19. Sarkar, B.; Sinha, R.; Sarkar, K. Initial viral load of a COVID-19-infected case indicated by its cycle threshold value of polymerase chain reaction could be used as a predictor of its transmissibility—An experience from Gujarat, India. *Indian J. Community Med.* **2020**, *45*, 278. [[CrossRef](#)]
20. Avadhanula, V.; Nicholson, E.G.; Ferlic-Stark, L.; Piedra, F.A.; Blunck, B.N.; Fragoso, S.; Bond, N.L.; Santarcangelo, P.L.; Ye, X.; McBride, T.J.; et al. Viral load of Severe Acute Respiratory Syndrome Coronavirus 2 in adults during the first and second wave of Coronavirus Disease 2019 pandemic in Houston, Texas: The potential of the superspreader. *J. Infect. Dis.* **2021**, *223*, 1528–1537. [[CrossRef](#)]
21. Singanayagam, A.; Patel, M.; Charlett, A.; Bernal, J.L.; Saliba, V.; Ellis, J.; Ladhani, S.; Zambon, M.; Gopal, R. Duration of infectiousness and correlation with RT-PCR cycle threshold values in cases of COVID-19, England, January to May 2020. *Eurosurveillance* **2020**, *25*, 2001483. [[CrossRef](#)]
22. Walker, A.S.; Pritchard, E.; House, T.; Robotham, J.V.; Birrell, P.J.; Bell, I.; Bell, J.I.; Newton, J.N.; Farrar, J.; Diamond, I.; et al. CT threshold values, a proxy for viral load in community sars-cov-2 cases, demonstrate wide variation across populations and over time. *eLife* **2021**, *10*, e64683. [[CrossRef](#)]
23. Abdulrahman, A.; Mallah, S.I.; Alawadhi, A.; Perna, S.; Janahi, E.M.; AlQahtani, M.M. Association between RT-PCR Ct values and COVID-19 new daily cases: A multicenter cross-sectional study. *Le Infez. Med.* **2021**, *29*, 416. [[CrossRef](#)]
24. Hay, J.A.; Kennedy-Shaffer, L.; Kanjilal, S.; Lennon, H.J.; Gabriel, S.B.; Lipsitch, M.; Mina, M.J. Estimating epidemiologic dynamics from cross-sectional viral load distributions. *Science* **2021**, *373*, eabh0635. [[CrossRef](#)] [[PubMed](#)]
25. Khalil, A.; Feghali, R.; Hassoun, M. The Lebanese COVID-19 Cohort; A Challenge for the ABO Blood Group System. *Front. Med.* **2020**, *7*, 585341. [[CrossRef](#)] [[PubMed](#)]
26. *Epidemiological Surveillance*; Technical Report; Ministry of Public Health: Baabda, Lebanon, 2022.
27. Worldometer. Daily New Cases in Lebanon. Available online: <https://www.worldometers.info/coronavirus/country/lebanon/> (accessed on 31 March 2022).
28. Allen, D.M. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **1974**, *16*, 125–127. [[CrossRef](#)]
29. Moons, K.G.M.; Altman, D.G.; Reitsma, J.B.; Ioannidis, J.P.A.; Macaskill, P.; Steyerberg, E.W.; Vickers, A.J.; Ransohoff, D.F.; Collins, G.S. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann. Intern. Med.* **2015**, *162*, W1–W73. [[CrossRef](#)] [[PubMed](#)]

30. Smith, M.R.; Trofimova, M.; Weber, A.; Duport, Y.; Kühnert, D.; von Kleist, M. Rapid incidence estimation from SARS-CoV-2 genomes reveals decreased case detection in Europe during summer 2020. *Nat. Commun.* **2021**, *12*, 1–13. [[CrossRef](#)] [[PubMed](#)]
31. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
32. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386–408. [[CrossRef](#)]
33. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958. [[CrossRef](#)]
34. Goyal, A.; Lamb, A.; Zhang, Y.; Zhang, S.; Courville, A.; Bengio, Y. Professor forcing: A new algorithm for training recurrent networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 4608–4616.
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015.
36. Prechelt, L. Automatic early stopping using cross validation: Quantifying the criteria. *Neural Netw.* **1998**, *11*, 761–767. [[CrossRef](#)]
37. Drucker, H.; Surges, C.J.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 2–5 December 1997; pp. 155–161.
38. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
39. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. In *Encyclopedia of Database Systems*; Springer: Boston, MA, USA, 2009; pp. 532–538. [[CrossRef](#)]
40. Audet, C.; Dzahini, K.J.; Kokkolaras, M.; Le Digabel, S. Stochastic mesh adaptive direct search for blackbox optimization using probabilistic estimates. *Comput. Optim. Appl.* **2021**, *79*, 1–34. [[CrossRef](#)]
41. Boulmaiz, T.; Guermoui, M.; Boutaghane, H. Impact of training data size on the LSTM performances for rainfall-runoff modeling. *Model. Earth Syst. Environ.* **2020**, *6*, 2153–2164. [[CrossRef](#)]
42. Emmert-Streib, F.; Yang, Z.; Feng, H.; Tripathi, S.; Dehmer, M. An introductory review of deep learning for prediction models with big data. *Front. Artif. Intell.* **2020**, *3*, 4. [[CrossRef](#)]
43. Heroku. COVID-19 Weekly Forecaster. Available online: <https://covid-forecaster-lebanon.herokuapp.com> (accessed on 31 March 2022).
44. Clinical importance of reporting SARS-CoV-2 viral loads across the different stages of the COVID-19 pandemic. *medRxiv* **2020**. [[CrossRef](#)]
45. Booton, R.D.; Macgregor, L.; Vass, L.; Looker, K.J.; Hyams, C.; Bright, P.D.; Harding, I.; Lazarus, R.; Hamilton, F.; Lawson, D.; et al. Estimating the COVID-19 epidemic trajectory and hospital capacity requirements in South West England: A mathematical modelling framework. *BMJ Open* **2021**, *11*, 41536. [[CrossRef](#)]
46. Liu, Y.; Gayle, A.A.; Wilder-Smith, A.; Rocklöv, J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J. Travel Med.* **2020**, *27*, 1–4. [[CrossRef](#)]
47. Linka, K.; Peirlinck, M.; Kuhl, E. The reproduction number of COVID-19 and its correlation with public health interventions. *Comput. Mech.* **2020**, *66*, 1035–1050. [[CrossRef](#)]
48. Fuhrman, J.D.; Gorre, N.; Hu, Q.; Li, H.; El Naqa, I.; Giger, M.L. A review of explainable and interpretable AI with applications in COVID-19 imaging. *Med. Phys.* **2022**, *49*, 1–14. [[CrossRef](#)]
49. Al-Hajj, S.; Mokdad, A.H.; Kazzi, A. Beirut explosion aftermath: Lessons and guidelines. *Emerg. Med. J.* **2021**, *38*, 938–939. [[CrossRef](#)]
50. Koweyes, J.; Salloum, T.; Haidar, S.; Merhi, G.; Tokajian, S. COVID-19 Pandemic in Lebanon: One Year Later, What Have We Learnt? *mSystems* **2021**, *6*, e00351-21. [[CrossRef](#)] [[PubMed](#)]
51. Singanayagam, A.; Hakki, S.; Dunning, J.; Madon, K.J.; Crone, M.A.; Koycheva, A.; Derqui-Fernandez, N.; Barnett, J.L.; Whitfield, M.G.; Varro, R.; et al. Community transmission and viral load kinetics of the SARS-CoV-2 delta (B.1.617.2) variant in vaccinated and unvaccinated individuals in the UK: A prospective, longitudinal, cohort study. *Lancet Infect. Dis.* **2022**, *22*, 183–195. [[CrossRef](#)]
52. Al Handawi, K.; Kokkolaras, M. Optimization of Infectious Disease Prevention and Control Policies Using Artificial Life. *IEEE Trans. Emerg. Top. Comput. Intell.* **2021**, *6*, 26–40. [[CrossRef](#)]
53. Hinch, R.; Probert, W.J.; Nurtay, A.; Kendall, M.; Wymant, C.; Hall, M.; Lythgoe, K.; Bulas Cruz, A.; Zhao, L.; Stewart, A.; et al. OpenABM-COVID19—An agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. *PLoS Comput. Biol.* **2021**, *17*, e1009146. [[CrossRef](#)]
54. Willem, L.; Abrams, S.; Libin, P.J.; Coletti, P.; Kuylen, E.; Petrof, O.; Møgelmoose, S.; Wambua, J.; Herzog, S.A.; Faes, C.; et al. The impact of contact tracing and household bubbles on deconfinement strategies for COVID-19. *Nat. Commun.* **2021**, *12*, 1–9. [[CrossRef](#)] [[PubMed](#)]
55. Kerr, C.C.; Stuart, R.M.; Mistry, D.; Abeysuriya, R.G.; Rosenfeld, K.; Hart, G.R.; Núñez, R.C.; Cohen, J.A.; Selvaraj, P.; Hagedorn, B.; et al. Covasim: An agent-based model of COVID-19 dynamics and interventions. *PLoS Comput. Biol.* **2021**, *17*, e1009149. [[CrossRef](#)] [[PubMed](#)]
56. Lakhmiri, D.; Le Digabel, S.; Tribes, C. HyperNOMAD: Hyperparameter optimization of deep neural networks using mesh adaptive direct search. *ACM Trans. Math. Softw.* **2021**, *47*, 1–27. [[CrossRef](#)]