Les Cahiers du GERAD ISSN: 0711–2440

Weekly forecasting of new COVID-19 cases using past viral load measurements

A. Khalil, K. Al Handawi, Z. Mohsen, A. Abdel Nour, R. Feghali, I. Chamseddine, M. Kokkolaras

G–2021–48

August 2021
Revised: May 2022

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2021
– Bibliothèque et Archives Canada, 2021

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2021
– Library and Archives Canada, 2021

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Weekly forecasting of new COVID-19 cases using past viral load measurements

Athar Khalil [a, b]

Khalil Al Handawi [c, d]

Zeina Mohsen [e]

Afif Abdel Nour [f]

Rita Feghali [e]

Ibrahim Chamseddine [g]

Michael Kokkolaras [c, d]

[a] *Department of Genetics and Genome Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH, United States*

[b] *Clinical Research Unit, Rafik Hariri University Hospital, Beirut, Lebanon*

[c] *GERAD, Montréal (Qc), Canada*

[d] *Systems Optimization Lab, Department of Mechanical Engineering, McGill University, Montréal (Qc), Canada*

[e] *Department of Laboratory Medicine, Rafik Hariri University Hospital, Beirut, Lebanon*

[f] *School of Engineering, The Holy Spirit University of Kaslik, Jounieh, Lebanon*

[g] *Department of Radiation Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States*

ichamseddine@mgh.harvard.edu

August 2021
Revised: May 2022
Les Cahiers du GERAD
G–2021–48

Abstract : The transmission of the contagious Coronavirus disease (COVID-19) is highly dependent on individual viral dynamics. Reverse-transcription quantitative polymerase chain reaction (RT-qPCR) tests used for diagnosing COVID-19 provide a semi-quantitative measurement of viral load within the infected host in the form of a cycle threshold (Ct) value. We solicited Ct values sampled from a cross-sectional patient cohort at Rafik Hariri University Hospital (RHUH) to now-cast COVID-19 incidences in Lebanon. Our patient cohort of 9531 patients, retrieved from a single representative cross-sectional virology test center in Lebanon, revealed that when the mean Ct value of a daily sample of patients is low, an increase in positive COVID-19 case counts is observed in Lebanon. A subset of the data was used to train several machine learning models and tune their hyperparameters with respect to the validation error. Unseen data unused during model development is used to report the models' test error. Support vector machine regression performed well on the unseen data and was able to provide predictions for the positive case counts in Lebanon over the span of 7 days. The models are a first attempt at capturing the interaction between cross-sectional Ct values and the pandemic trajectory including temporal effects that arise from population dynamics. The model has potential applications for predicting COVID-19 incidences in other countries and in assessing post-vaccination policies. Apart from emphasizing patient responsibility in adopting early testing practices, this study proposed and validated viral load measurement as a relevant input that can enhance the predictive accuracy of viral disease now-casting models.

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

---

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

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

Conflict 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.

Ethics statement: The studies involving human participants were reviewed and approved by Ethical Committee of RHUH. The ethics committee waived the requirement of written informed consent for participation.

Data availability statement: The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding authors.

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) [3, 31]. 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 [42]. 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 [26]. 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, social behaviour among other aspects are used to inform now-casting models and provide situational awareness to policy makers [56]. 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 [11, 27]. In this paper, we focus 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 alarm policy makers and health institutions towards an increase in incidence rates. This allows sufficient time to use other detailed scenario models to test and deploy various PHSMs proactively [7, 28, 40, 45].

RT-qPCR is a serological test that remains the gold standard for COVID-19 diagnosis [39]. It measures the first PCR cycle, denoted as the cycle threshold (Ct), at which a detectable signal of the targeted DNA appears [8]. 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 [2]. Ct values were proposed to have potential prognostic value in predicting severity, infectiousness, and mortality among patients [43]. Ct values were also used to determine the duration an infected patient needs to quarantine [37, 46]. 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 [4]. 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-folds [48]. Individuals with high viral load and mild symptoms can be identified as potential superspreaders using viral load measurements [14]. Thus, early testing is highly recommended alongside isolation practices, to interrupt SARS-CoV-2 transmission [50].

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 [6, 53]. 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 [22, 27]. 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 transmission and can potentially generalize better when viral transmission is not completely understood and sufficient data is available.

We demonstrate the merits of this approach using a novel 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 the preparedness of healthcare providers against new SARS-CoV-2 variants and the surge in new cases caused by them. Furthermore, now-casting the pandemic trajectory can support policy makers 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 March 1, 2020, and March 31, 2021. Rafik Hariri University Hospital (RHUH) is the country's leading institution for COVID-19 testing and treatment, and our cohort represents the nation's COVID-19 trajectory well [30]. 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 worldmeters [5, 55]. 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 longitudinal split of the data. The discovery group (Group 1) was used for training and cross-validation [12] 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) [38], which represents a classification criterion for predictive modeling. It has four types of increasing reliability. Since we split the data 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 that were developed using the discovery group (Group 1). This third group of data is called the 'unseen data group'.

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 4) and age (Figure S.IIIa) 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 [51]. 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 S.I and Figure S.IIIb).

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 [11, 56]. 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 of the 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 virological cross-sectional data have used the 7-day window size for now-casting pandemic trajectories [22]. We developed 6 different machine learning algorithms for now-casting the epidemic trajectory which are described as follows.

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 temporaly correlated (i.e. the independent and identically distributed (i.i.d) assumption does not hold for time series data). This type of neural network is able to 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 is able to capture long-term temporal effects and trends encoded by a long sequence of inputs and avoid the problem of vanishing gradients during backpropagation [25]. 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 [47]. We used stacked LSTM cells with several layers (given by n_{layers}) in our 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 [52]. Temporal information at time step t_i of the n-th layer LSTM cell is represented by its hidden $h^n_{t_i}$ and cell states $C^n_{t_i}$.

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 $\mathbf{z}_n = \begin{bmatrix} C^n_{t-1} & h^n_{t-1} \end{bmatrix}^T$, 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 own predictions $\hat{n}^{t_i}_{\text{cases}}$ 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 $n^{t_i}_{\text{cases}}$ at time step t_{i+1} instead of the decoder's prediction with a probability P_{teacher} [21]. The architecture of the S2S model used in this paper is shown in Figure 1.

We also developed a second RNN model that is 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 S.IVa). This model is called the stacked LSTM (SEQ) model.

2.3.2 Feedforward neural network (DNN) model

We then developed a third model based on deep learning using feedforward neural networks. The 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 S.IVb). All deep learning models were trained using the stochastic gradient descent algorithm ADAM with a learning rate l_{rate} and batch size b_{size} [32]. 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}) [41].

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 [17], a gradient boosting machine (GBM) regression model [19], 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. This is because such models are primarily used for regression of univariate functions.

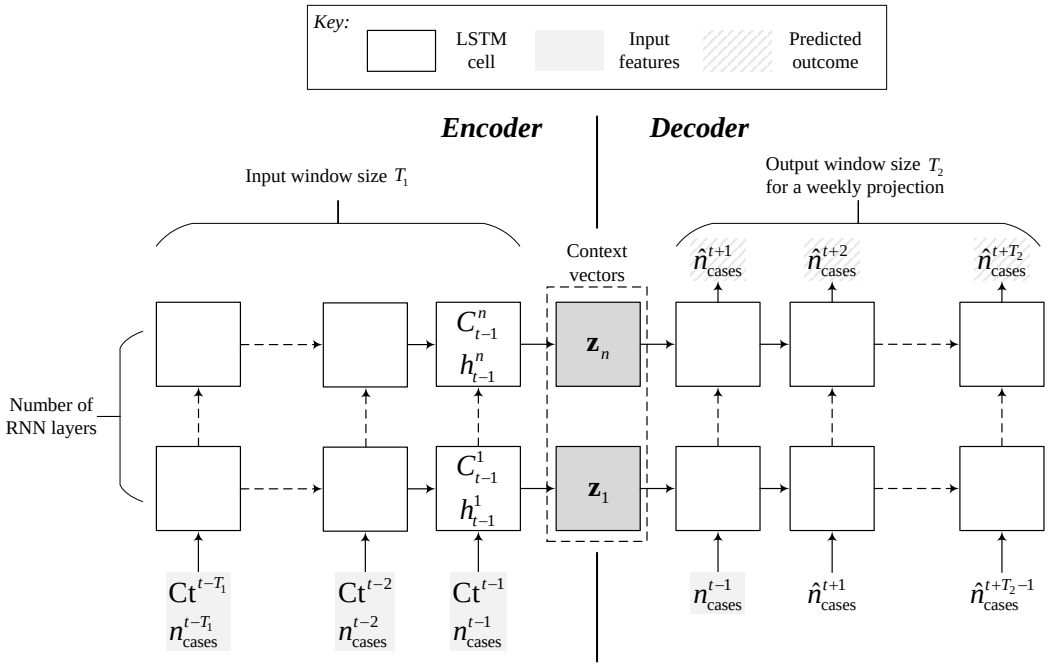


Figure 1: Structure of the sequence-to-sequence (S2S) model used for now-casting the weekly number of cases. The left side of the network is the encoder that uses past information on Ct and the number of cases to create context vectors used to initialize the hidden and cell states of the decoder LSTM cells.

2.3.4 Hyperparameter tuning

The hyperparameters of each model (listed in and described in Table 2) were optimized using cross-validation on the discovery group (Group 1) only. The cross-validation consists of outer and inner loops (Figure 2).

The outer loop split (Group 1) into five groups and sent four of them into the inner loop for training the models and subsequent hyperparameter optimization with respect to the average k-fold cross-validation error [44]. 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. 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 [13]. 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).

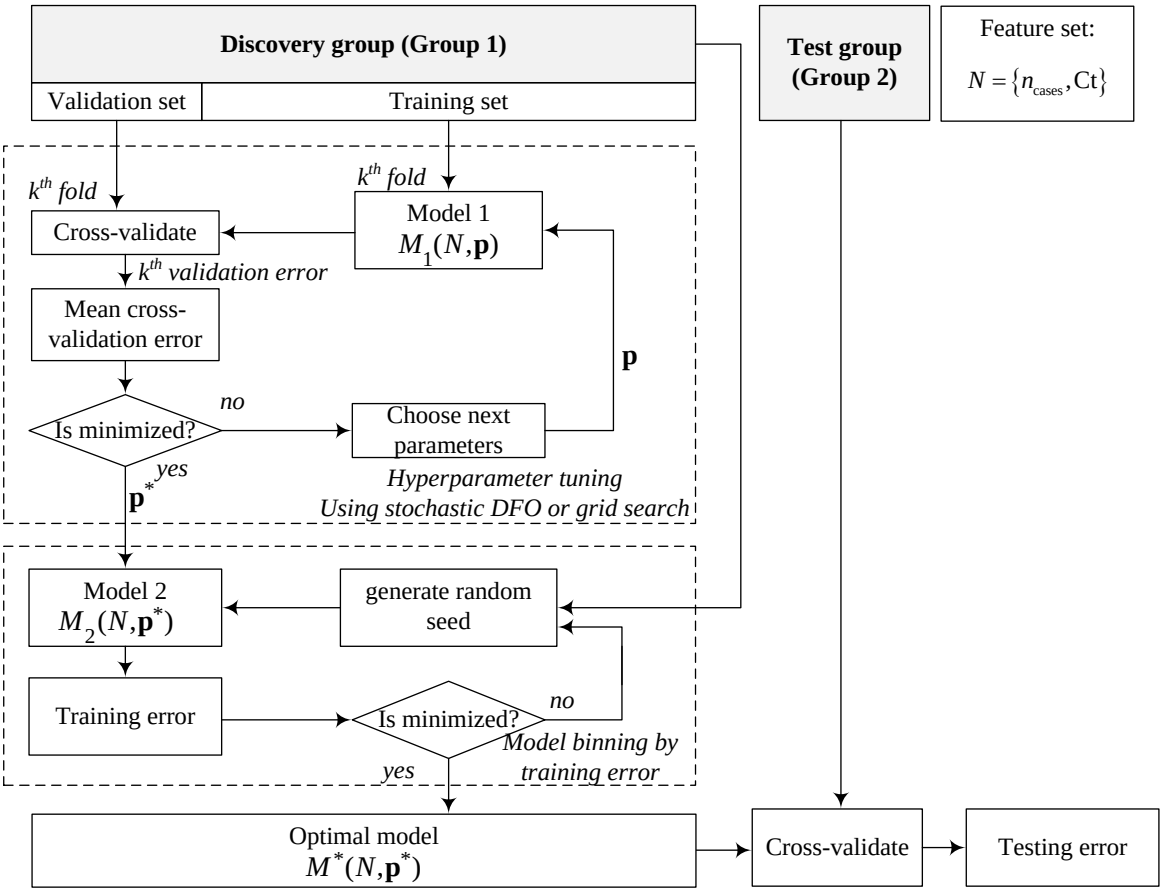


Figure 2: Cross-validation and hyperparameter determination scheme for model development. Following the discovery group (Group 1), the inner loop tuned the model's hyperparameters by minimizing the average k-fold cross-validation error using a stochastic direct search algorithm or a grid search. The 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.

We note that binning and sampling of the cross-validation error is not necessary 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 March 2, 2020 and October 17, 2020, Group 2 contained 3228 patients from October 18, 2020 to November 30, 2020, and the unseen group contained 12097 patients from December 01, 2020 to March 16, 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 3 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 including group 3 is provided in the supplementary material, Figure S.I).

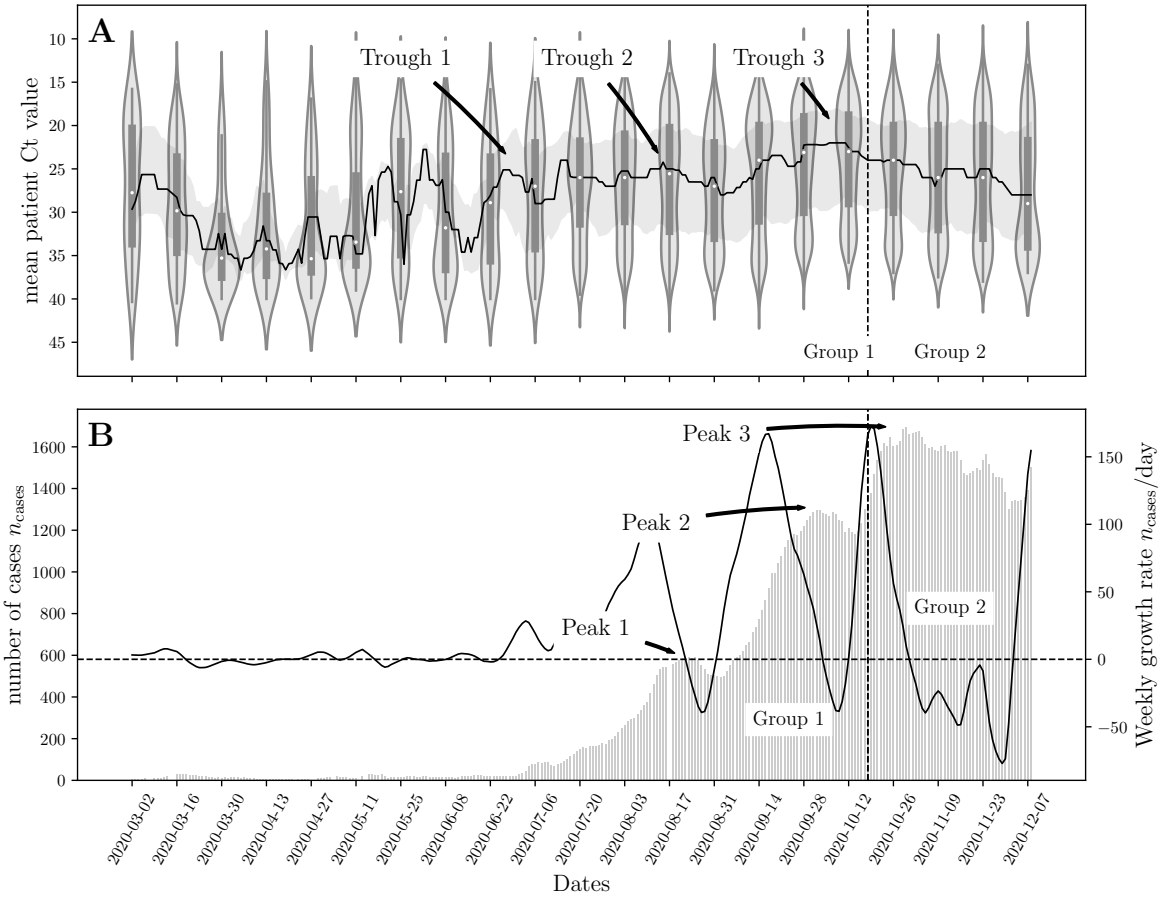


Figure 3: (A) Bi-weekly mean Ct values of RHUH patients. The solid line represents the median bi-weekly Ct values, and the gray shaded area represents the inter-quartile range (25-75 percentile) of the observed Ct values. (B) The grey bars show the weekly running average of the number of cases observed nationwide in Lebanon between March 01, 2020, and December 07, 2020 (the running average can be computed until November 23rd). 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 October 8, 2020 (Trough 3 in Figure 3A) was followed by an increase in the number of cases, on October 29, 2020, with more than 1640 cases per day (Peak 3 in Figure 3B). 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. 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 also investigate the relationship between the median Ct value and case counts using a 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 4). This indicates that the mean cross-sectional Ct value is an important feature for now-casting the pandemic trajectory.

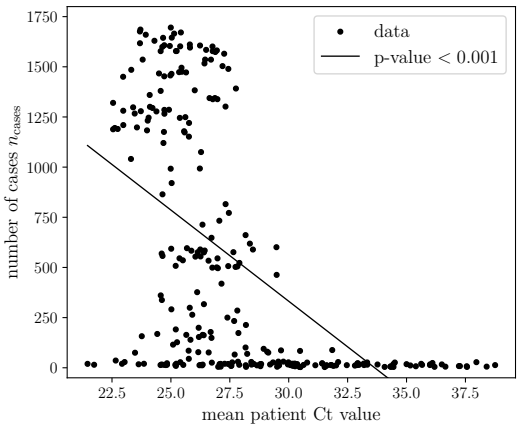


Figure 4: Scatter plot of biweekly mean Ct values and observed number of cases nationwide showing a clear negative value that is significant as given by p-value < 0.05.

3.3 Now-casting the epidemic trajectories

We developed 6 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 2. 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 October 18, 2020 to November 30, 2020 (Group 2) (Figure 5). The models were then retrained on Groups 1 and 2 using the hyperparameters in Table 2 and used to now-cast the epidemic trajectory after December 01, 2020 (Group 3). Table 1 lists the MSE error for the predicted trajectories on Groups 2 and 3 (see Table 1 footnote).

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-folds larger (0.042). The OLS and SVR had an MSE that is 4 folds 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 datapoints 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 [16, 18].

To verify this, the RNN models where 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 S.I). This implies that the RNN models generalize better when more training data is available (see supplementary material Section S.III). If limited data is 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 for the purpose of further prospective validation in the future [23].

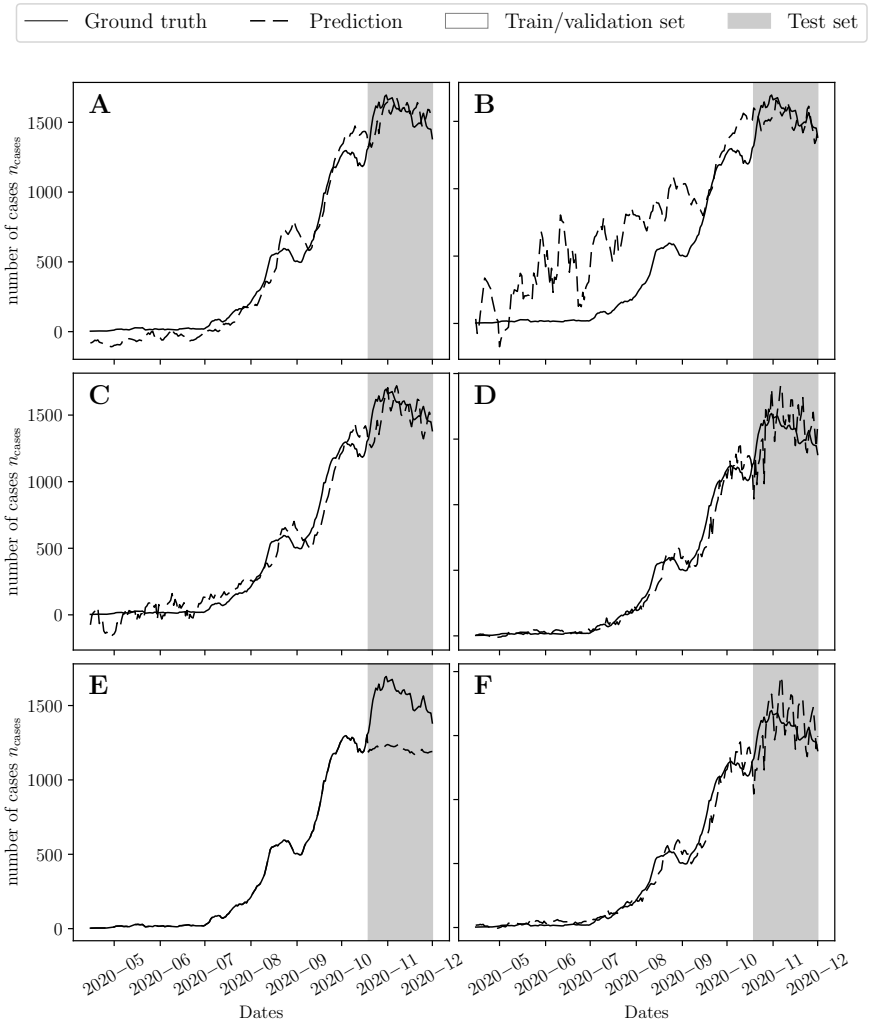


Figure 5: Predicted 7-day rolling average of daily number of cases on the unseen data set using (A) the sequence-to-sequence (S2S) model, (B) the stacked LSTM (SEQ), (C) The feedforward neural network (DNN), (D) The support vector machine regression (SVR) model, (E) The gradient boosting machine (GBM), 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 1: 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)	2.316e-6	0.32589	2.316e-6	1.44463
Polynomial regression (OLS)	0.01335	0.08954	0.00459	0.15954

The MSE 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 given by 463.8 and 597.0, respectively.

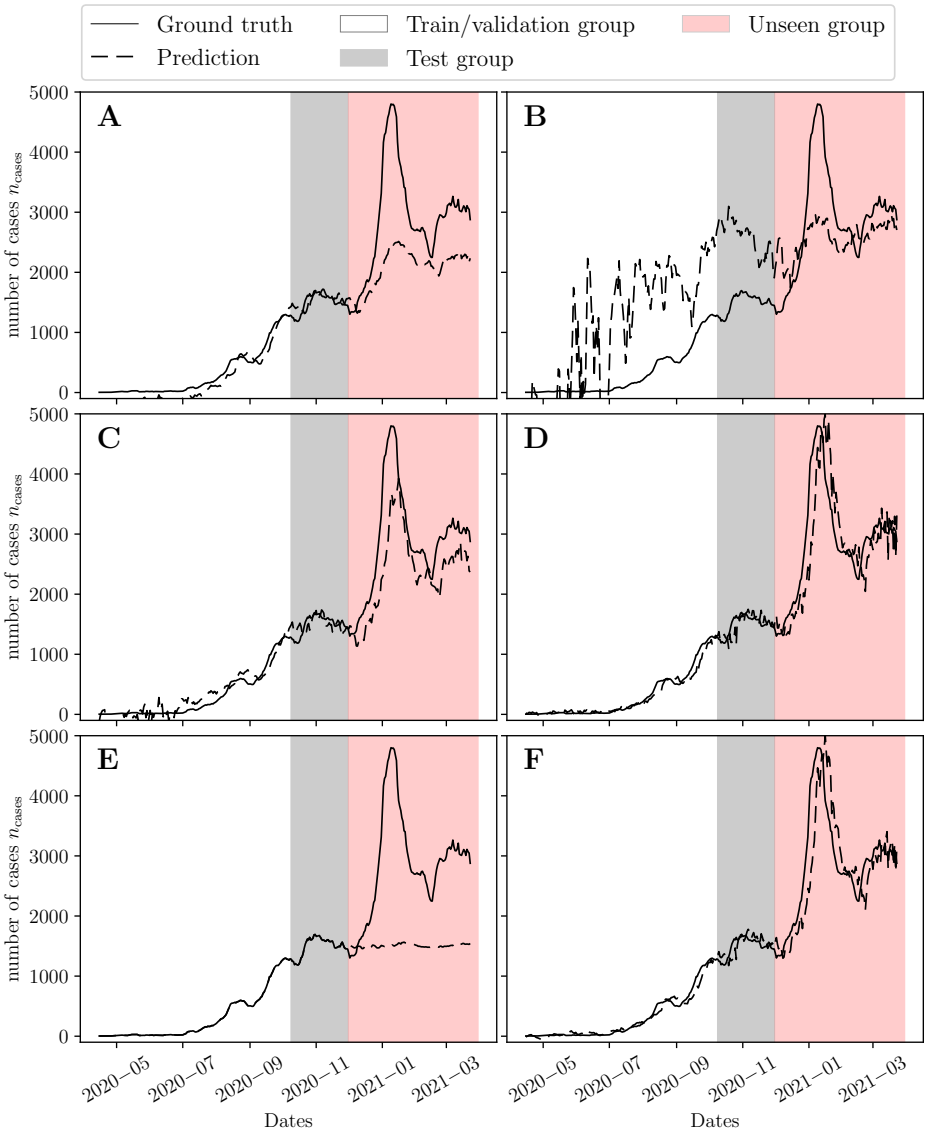


Figure 6: Predicted 7-day rolling average of daily number of cases on the unseen data set using (A) the sequence-to-sequence (S2S) model, (B) the stacked LSTM (SEQ), (C) The feedforward neural network (DNN), (D) The support vector machine regression (SVR) model, (E) The gradient boosting machine (GBM), 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 the both the discrovery 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 [1]. Mathematical modeling has been widely used for predicting the course of COVID-19 pandemic. These predicting models were developed based on the applied intervention measurements and the population behavioral fluctuations, including social distancing and mask-wearing [15]. The COVID-19 reproduction number (R0), 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 [36]. Although R0 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 [35]. So far, only few advanced and more recent models have evaluated the disease spread based on viral kinetics and

Table 2: 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	$n^{\text{best}}_{\text{epochs}}$	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	$n^{\text{best}}_{\text{epochs}}$	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	$n^{\text{best}}_{\text{epochs}}$	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 of 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.

serological assays (such as RT-qPCR tests) [22, 53]. Furthermore, these studies focused on serological assays or pandemic size indicators (such as R0 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. used Bayesian inference to predict the growth rate in the daily number of COVID-19 cases as a function of Ct values [22]. They showed that the population-level Ct distribution is strongly correlated with growth rate estimates of new infections in Massachusetts, USA. They estimated R0 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 evolution of the trajectory as more Ct observations are made. We have used the Gaussian process regression model to predict the pandemic trajectory using our cross-sectional patient cohort (see supplementary material Section S.IV). 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 [20]. 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 virus takes hold. Another dataset from Bahrain demonstrated the effectiveness of Ct in predicting the epidemiological dynamics of COVID-19 [6]. 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, our data-driven approach of inferring the epidemic trajectory using past cases counts and Ct observations using machine learning models 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. (Figure S.IXA). Most models captured the underlying trend with the exception of GBM and the stacked LSTM (SEQ) models (Figure S.X). SVR performed the best on this dataset (Figure S.IXB). 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 (April 15, 2020 - December 15, 2020) and Lebanon (December 01, 2020 - March 31, 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 S.V). Although the outcomes of this study favored simpler regression models, their simplicity provides an advantage in terms of interpretablity [20].

Our dataset 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 in our cohort mirrored the well-recognized political, economic, and social turning points that happened in Lebanon during the pandemic. These incidences impacted the population behaviour 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 in our cohort 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 [30]. 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 [9]. 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% [9, 33]. 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 a 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, our developed comprehensive training and validation errors reflect the models' robustness against unexpected events.

The detected inversely proportional relationship between Ct values and number of national COVID-19 positive cases reflects population dynamics of transmission and demonstrates the temporal significance of Ct values. Our results emphasized the importance of early testing when patient's viral load and infectivity is low to prompt isolation practises and thus, suppress national spread of the virus. Our established 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 a the same manner upon infection [49]. Ultimately, our data promotes 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).

Our 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 [23]. Another limitation is the inability of the model to compare the effect of preventative policies such as lock-downs and quarantining. The model does provide an alert when the number of cases are about the 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 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 [10, 24, 29, 54]. A future study could focus on combining IBMs with viral load models such as those developed by Hay et al. 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 [22].

5 Conclusions

Based on the premise that SARS-CoV-2 spread is highly dependent on the individual viral dynamics, we developed models that predict the national COVID-19 incidence rate based on mean Ct values retrieved from a single representative cross-sectional survey in Lebanon. The modeling framework relied on multiple machine learning algorithms that make little assumptions about population and transmission dynamics and is a first attempt at combining serological assays with epidemiological indicators to now-cast the pandemic trajectory [56]. Our COVID-19 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, 53]. To account for the effect of social interactions that could occur 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 [38]. 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 [34]. 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. The models showed that past Ct values obtained from a representative cross-sectional patient cohort could be used to now-cast the number of nationwide positive COVID-19 cases in a specific region.

References

[1] Clinical importance of reporting SARS-CoV-2 viral loads across the different stages of the COVID-19 pandemic, jul 2020.

[2] 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, 2020.

[3] World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19), apr 2020.

[4] COVID-19: management of staff and exposed patients or residents in health and social care settings. Technical report, UK Health Security Agency, 2022.

[5] Epidemiological Surveillance. Technical report, Ministry of Public Health - Lebanon, 2022.

[6] A. Abdulrahman, S. I. Mallah, A. Alawadhi, S. Perna, E. M. Janahi, and M. M. AlQahtani. Association between RT-PCR Ct values and COVID-19 new daily cases: A multicenter cross-sectional study, dec 2020.

[7] S. Abrams, J. Wambua, E. Santermans, L. Willem, E. Kuylen, P. Coletti, P. Libin, C. Faes, O. Petrof, S. A. Herzog, P. Beutels, and N. Hens. Modelling the early phase of the Belgian COVID-19 epidemic using a stochastic compartmental model and studying its implied future trajectories. Epidemics, 35:100449, jun 2021.

[8] C. Ade, J. Pum, I. Abele, L. Raggub, D. Bockmühl, and B. Zöllner. Analysis of cycle threshold values in SARS-CoV-2-PCR in a long-term study. Journal of Clinical Virology, 138, may 2021.

[9] S. Al-Hajj, A. H. Mokdad, and A. Kazzi. Beirut explosion aftermath: Lessons and guidelines, mar 2021.

[10] K. Al Handawi and M. Kokkolaras. Optimization of Infectious Disease Prevention and Control Policies Using Artificial Life. IEEE Transactions on Emerging Topics in Computational Intelligence, pages 1–15, 2021.

[11] P. Alaimo Di Loro, F. Divino, A. Farcomeni, G. Jona Lasinio, G. Lovison, A. Maruotti, and M. Mingione. Nowcasting COVID-19 incidence indicators during the Italian first outbreak. Statistics in Medicine, 40(16):3843–3864, jul 2021.

[12] D. M. Allen. The relationship between variable selection and data agumentation and a method for prediction. Technometrics, 16(1):125–127, jan 1974.

[13] C. Audet, K. J. Dzahini, M. Kokkolaras, and S. Le Digabel. Stochastic mesh adaptive direct search for blackbox optimization using probabilistic estimates. Computational Optimization and Applications, page 35, 2020.

[14] V. Avadhanula, E. G. Nicholson, L. Ferlic-Stark, F. A. Piedra, B. N. Blunck, S. Fragoso, N. L. Bond, P. L. Santarcangelo, X. Ye, T. J. McBride, L. O. Aideyan, K. D. Patel, L. Maurer, L. S. Angelo, and P. A. Piedra. 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. Journal of Infectious Diseases, 223(9):1528–1537, may 2021.

[15] R. D. Booton, L. Macgregor, L. Vass, K. J. Looker, C. Hyams, P. D. Bright, I. Harding, R. Lazarus, F. Hamilton, D. Lawson, L. Danon, A. Pratt, R. Wood, E. Brooks-Pollock, and K. M.E. Turner. Estimating the COVID-19 epidemic trajectory and hospital capacity requirements in South West England: A mathematical modelling framework. BMJ Open, 11(1):41536, jan 2021.

[16] T. Boulmaiz, M. Guermoui, and H. Boutaghane. Impact of training data size on the LSTM performances for rainfall-runoff modeling. Modeling Earth Systems and Environment, 6(4):2153–2164, dec 2020.

[17] H. Drucker, C. J.C. Surges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. In Advances in Neural Information Processing Systems, pages 155–161, 1997.

[18] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, and M. Dehmer. An introductory review of deep learning for prediction models with big data, feb 2020.

[19] J. H. Friedman. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5):1189–1232, 2001.

[20] J. D. Fuhrman, N. Gorre, Q. Hu, H. Li, I. El Naqa, and M. L. Giger. A review of explainable and interpretable AI with applications in COVID-19 imaging. Medical Physics, 49(1):1–14, jan 2022.

[21] A. Goyal, A. Lamb, Y. Zhang, S. Zhang, A. Courville, and Y. Bengio. Professor forcing: A new algorithm for training recurrent networks. In Advances in Neural Information Processing Systems, pages 4608–4616, 2016.

[22] J. A. Hay, L. Kennedy-Shaffer, S. Kanjilal, H. J. Lennon, S. B. Gabriel, M. Lipsitch, and M. J. Mina. Estimating epidemiologic dynamics from cross-sectional viral load distributions. Science, page eabh0635, jun 2021.

[23] Heroku. Covid-19 weekly forecaster. https://covid-forecaster-lebanon.herokuapp.com, June. [Online; accessed 31-March-2022].

[24] R. Hinch, W. J.M. Probert, A. Nurtay, M. Kendall, C. Wymant, M. Hall, K. Lythgoe, A. Bulas Cruz, L. Zhao, A. Stewart, L. Ferretti, D. Montero, J. Warren, N. Mather, M. Abueg, N. Wu, O. Legat, K. Bentley, T. Mead, K. Van-Vuuren, D. Feldner-Busztin, T. Ristori, A. Finkelstein, D. G. Bonsall, L. Abeler-Dörner, and C. Fraser. OpenABM-Covid19—An agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. PLOS Computational Biology, 17(7):e1009146, jul 2021.

[25] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. Neural Computation, 9(8):1735–1780, nov 1997.

[26] N. Hoertel, M. Blachier, C. Blanco, M. Olfson, M. Massetti, M. S. Rico, F. Limosin, and H. Leleu. A stochastic agent-based model of the SARS-CoV-2 epidemic in France. Nature Medicine, 26(9):1417–1421, 2020.

[27] N. J. Irons and A. E. Raftery. Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys. Proceedings of the National Academy of Sciences of the United States of America, 118(31), aug 2021.

[28] A. A. Kamar, N. Maalouf, E. Hitti, G. El Eid, H. Isma'eel, and I. H. Elhajj. The Challenge of Forecasting Demand of Medical Resources and Supplies during a Pandemic: A Comparative Evaluation of Three Surge Calculators for COVID-19. Epidemiology and Infection, feb 2021.

[29] C. C. Kerr, R. M. Stuart, D. Mistry, R. G. Abeysuriya, K. Rosenfeld, G. R. Hart, R. C. Núñez, J. A. Cohen, P. Selvaraj, B. Hagedorn, L. George, M. Jastrzebski, A. S. Izzo, G. Fowler, A. Palmer, D. Delport, N. Scott, S. L. Kelly, C. S. Bennette, B. G. Wagner, S. T. Chang, A. P. Oron, E. A. Wenger, J. Panovska-Griffiths, M. Famulare, and D. J. Klein. Covasim: An agent-based model of COVID-19 dynamics and interventions. PLOS Computational Biology, 17(7):e1009149, jul 2021.

[30] A. Khalil, R. Feghali, and M. Hassoun. The Lebanese COVID-19 Cohort; A Challenge for the ABO Blood Group System. Frontiers in Medicine, 7:585341, nov 2020.

[31] A. Khalil, A. Kamar, and G. Nemer. Thalidomide-Revisited: Are COVID-19 Patients Going to Be the Latest Victims of Yet Another Theoretical Drug-Repurposing? Frontiers in Immunology, 11:1248, may 2020.

[32] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR, dec 2015.

[33] J. Koweyes, T. Salloum, S. Haidar, G. Merhi, and S. Tokajian. COVID-19 Pandemic in Lebanon: One Year Later, What Have We Learnt? mSystems, 6(2), apr 2021.

[34] D. Lakhmiri, S. Le Digabel, and C. Tribes. HyperNOMAD: Hyperparameter optimization of deep neural networks using mesh adaptive direct search. ACM Transactions on Mathematical Software, 47(3), 2021.

[35] K. Linka, M. Peirlinck, and E. Kuhl. The reproduction number of COVID-19 and its correlation with public health interventions, jul 2020.

[36] Y. Liu, A. A. Gayle, A. Wilder-Smith, and J. Rocklöv. The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of Travel Medicine, 27(2), mar 2020.

[37] E. H. Miller, J. Zucker, D. Castor, M. K. Annavajhala, J. L. Sepulveda, D. A. Green, S. Whittier, M. Scherer, N. Medrano, M. E. Sobieszczyk, M. T. Yin, L. Kuhn, and A. Uhlemann. 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 Infectious Diseases, 8(2), feb 2021.

[38] K. G. M. Moons, D. G. Altman, J. B. Reitsma, J. P. A. Ioannidis, P. Macaskill, E. W. Steyerberg, A. J. Vickers, D. F. Ransohoff, and G. S. Collins. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Annals of Internal Medicine, 162(1):W1–W73, 2015.

[39] H. Péré, I. Podglajen, M. Wack, E. Flamarion, T. Mirault, G. Goudot, C. Hauw-Berlemont, L. Le, E. Caudron, S. Carrabin, J. Rodary, T. Ribeyre, L. Bélec, and D. Veyer. Nasal swab sampling for SARS-CoV-2: A convenient alternative in times of nasopharyngeal swab shortage. Journal of Clinical Microbiology, 58(6), jun 2020.

[40] O. Pinto Neto, D. M. Kennedy, J. C. Reis, Y. Wang, A. B. Brizzi, G. J. Zambrano, J. M. de Souza, W. Pedroso, R. C. de Mello Pedreiro, B. de Matos Brizzi, E. O. Abinader, and R. A. Zângaro. Mathematical model of COVID-19 intervention scenarios for São Paulo—Brazil. Nature Communications, 12(1):1–13, jan 2021.

[41] L. Prechelt. Automatic early stopping using cross validation: Quantifying the criteria. Neural Networks, 11(4):761–767, jun 1998.

[42] A. A. Rabaan and et al. Airborne transmission of SARS-CoV-2 is the dominant route of transmission: Droplets and aerosols. Infezioni in Medicina, 29(1):10–19, 2021.

[43] S. N. Rao, S. Manissero, Victoria R. Steele, and J. Pareja. A Narrative Systematic Review of the Clinical Utility of Cycle Threshold Values in the Context of COVID-19. Infectious Diseases and Therapy, 9(3):573–586, sep 2020.

[44] P. Refaeilzadeh, L. Tang, and H. Liu. Cross-Validation. In Encyclopedia of Database Systems, pages 532–538. Springer, Boston, MA, 2009.

[45] R. C. Reiner and et al. Modeling COVID-19 scenarios for the United States. Nature Medicine, 27(1):94–105, oct 2021.

[46] C. Rodríguez-Grande, P. Catalán, L. Alcalá, S. Buenestado-Serrano, J. Adán-Jiménez, S. Rodríguez-Maus, M. Herranz, J. Sicilia, F. Acosta, L. Pérez-Lago, P. Muñoz, and D. García de Viedma. Different dynamics of mean SARS-CoV-2 RT-PCR Ct values between the first and second COVID-19 waves in the Madrid population. Transboundary and Emerging Diseases, page tbed.14045, mar 2021.

[47] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6):386–408, nov 1958.

[48] B. Sarkar, R. Sinha, and K. Sarkar. 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 Journal of Community Medicine, 45(3):278, 2020.

[49] A. Singanayagam and 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. The Lancet Infectious Diseases, 22(2):183–195, feb 2022.

[50] A. Singanayagam, M. Patel, A. Charlett, J. L. Bernal, V. Saliba, J. Ellis, S. Ladhani, M. Zambon, and R. Gopal. Duration of infectiousness and correlation with RT-PCR cycle threshold values in cases of COVID-19, England, January to May 2020. Eurosurveillance, 25(32), 2020.

[51] M. R. Smith, M. Trofimova, A. Weber, Y. Duport, D. Kühnert, and M. von Kleist. Rapid incidence estimation from SARS-CoV-2 genomes reveals decreased case detection in Europe during summer 2020. Nature Communications, 12(1):1–13, oct 2021.

[52] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15:1929–1958, 2014.

[53] A. Walker and 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, 10, jul 2021.

[54] L. Willem, S. Abrams, P. J.K. Libin, P. Coletti, E. Kuylen, O. Petrof, S. Møgelmose, J. Wambua, S. A. Herzog, C. Faes, P. Beutels, and N. Hens. The impact of contact tracing and household bubbles on deconfinement strategies for COVID-19. Nature Communications 2021 12:1, 12(1):1–9, mar 2021.

[55] Worldometer. Daily New Cases in Lebanon.

[56] J. T. Wu, K. Leung, L. T.Y. Lam, M. Y. Ni, C. K.H. Wong, J. S. Peiris, and G. M. Leung. Nowcasting epidemics of novel pathogens: lessons from COVID-19, mar 2021.

Supplemental Material: Weekly forecasting of new COVID-19 cases using past viral load measurements

S.I Other patient cohort data

In this section, we graphically depict the unseen data (Group 3) used for independent validation of the models that were developed in Section 3.3. Figure S.I shows the unseen data from December 01, 2020 to March 16, 2021. The observation that low Ct values coincide with high case counts as was observed in Section 3.2 still holds for Group 3. Other features such as mean patient age and daily number of confirmed positive patients that were not used in model development are shown in Figure S.II. Visual inspection of Figures S.IIA and C shows that there is no significant relationship between the mean age of a cross-sectional patient cohort and the incidence rates. This is confirmed by the correlation test done in Figure S.IIIa. On the other hand, visual inspection of Figures S.IIB and C, shows a significant correlation between the daily number of confirmed positive patients and the incidence rates. This is confirmed in Figure S.IIIb. This significant correlation indicates that the patient cohort is indeed representative of the population from which it is derived and can be used to derive valid epidemiological indicators for this study. This ascertains the fact that consistent testing practices at RHUH were followed throughout the study period (March 01, 2020 through March 16, 2021).

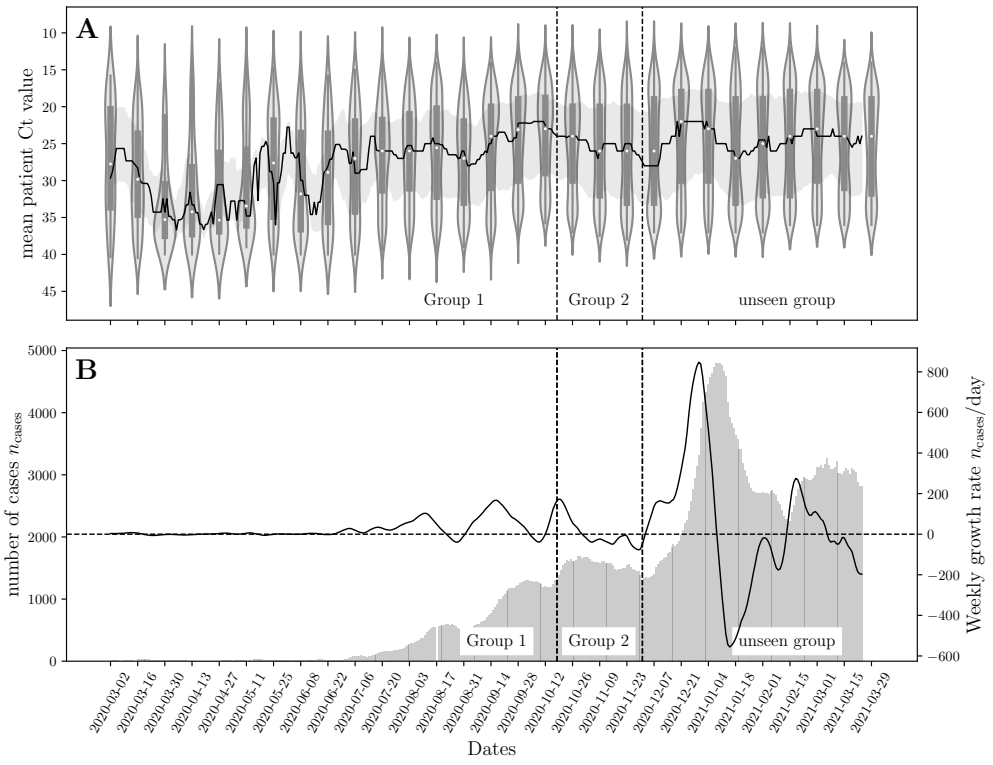


Figure S.I: (A) Bi-weekly mean Ct values of RHUH patients. The solid line represents the median bi-weekly Ct values, and the gray shaded area represents the inter-quartile range (25-75 percentile) of the observed Ct values. (B) The grey bars show the weekly running average of the number of cases observed nationwide in Lebanon between March 1, 2020, and March 31, 2021 (the running average can be computed until March 16). The solid black line represents the growth rate in the weekly number of cases.

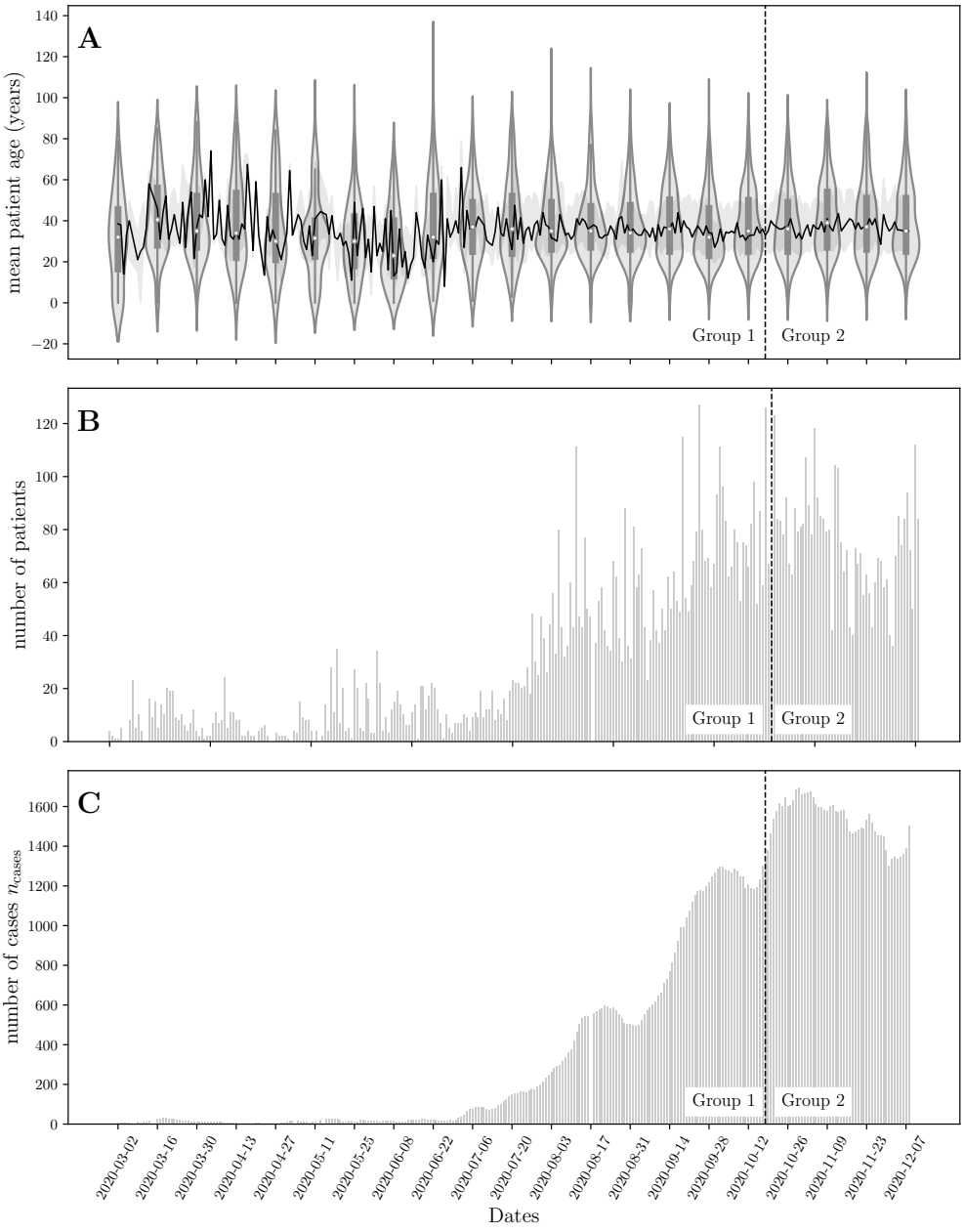
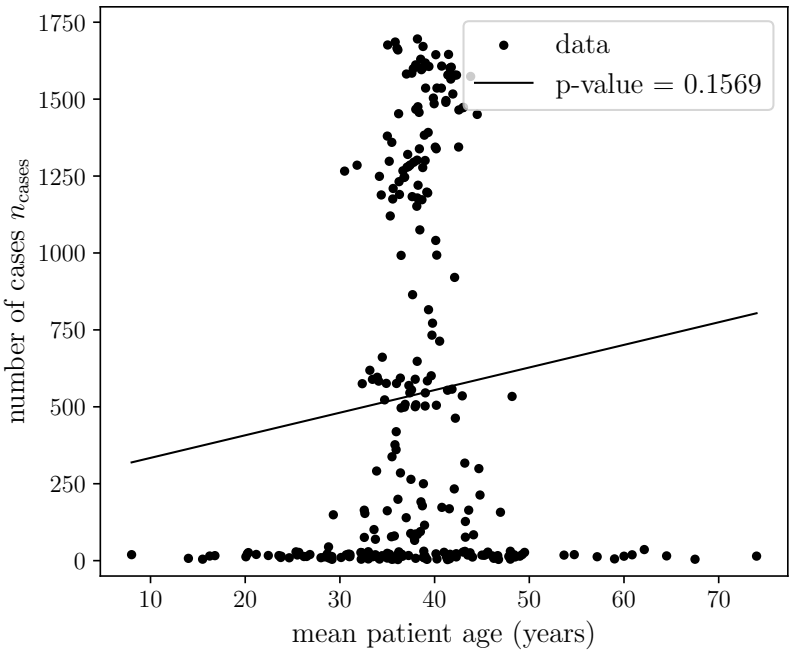
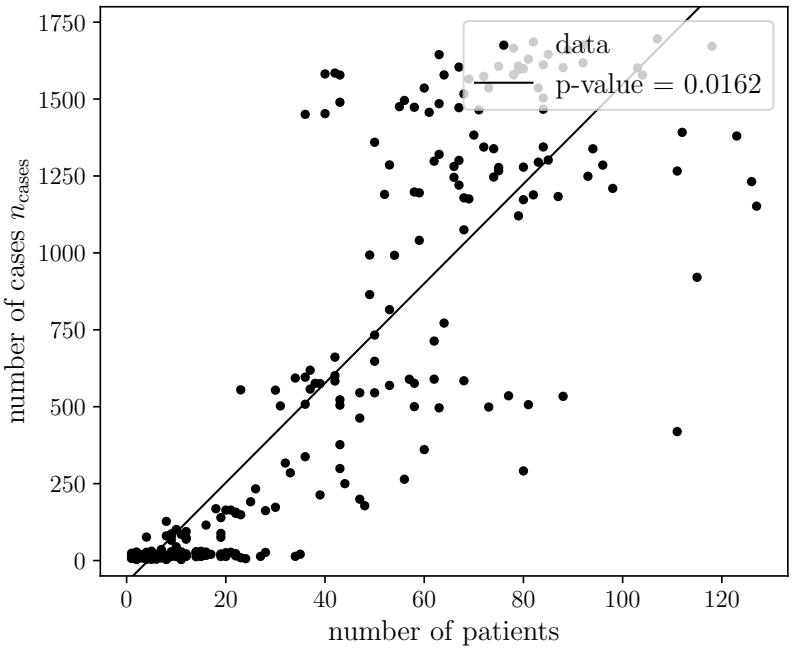


Figure S.II: (A) Bi-weekly mean age of RHUH patients. The solid line represents the median patient age, and the gray shaded area represents the inter-quartile range (25-75 percentile) of the observed patient ages. (B) Bi-weekly mean number of confirmed positive RHUH patients. (C) The grey bars show the weekly running average of the number of cases observed nationwide in Lebanon between March 1st, 2020, and December 30, 2020 (the running average can be computed until December 07).



(a)



(b)

Figure S.III: (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.

S.II Deep learning model architecture

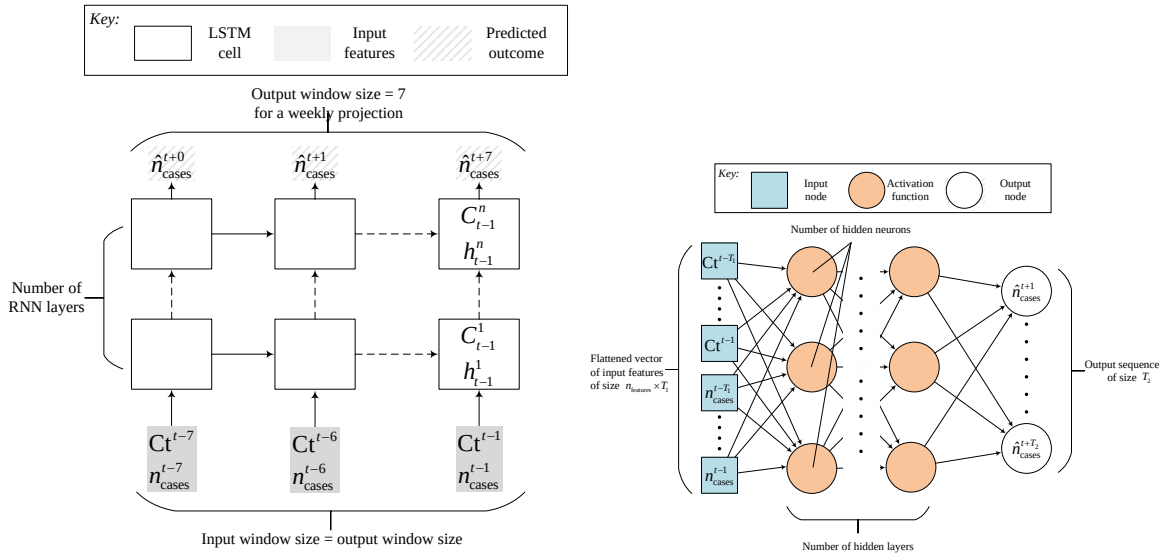


Figure S.IV: Model architecture of the (a) stacked LSTM (SEQ) and the (b) feedforward neural network (DNN) models.

S.III Effect of additional training data on model performance

The model development methodology in Section 3.3 was applied to the entire patient cohort of RHUH (Figure 3A). Groups 1 and 2 were both used to train and cross validate each model and its performance on Group 3 was tested using the MSE criterion. The predictions of these models on the unseen data group (Group 3) are shown in Figure S.V, the train and test errors are listed in Table S.I showing that the sequence-to-sequence (S2S) model outperforms all the other models as given by its low test error. This implies that deep learning model architecture can learn additional model representations as more training data becomes available. The distribution of the train and test errors is also shown by the box plots in Figure S.VII. The optimal hyperparameters of each model are listed in Table S.II.

Table S.I: 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).

Model	Train error	Test error
	Groups 1/2	Group 3
Sequence-to-sequence (S2S)	0.01460916	0.10580273
Stacked LSTM (SEQ)	0.00898817	0.34242207
Feedforward neural network (DNN)	0.01399668	0.20013593
Support vector machine regression (SVR)	0.00454633	0.13963704
Gradient boosting machine (GBM)	0.00035697	1.46476531
Polynomial regression (OLS)	0.00461187	0.16241863

The MSE 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 given by 1221.7 and 1341.2, respectively.

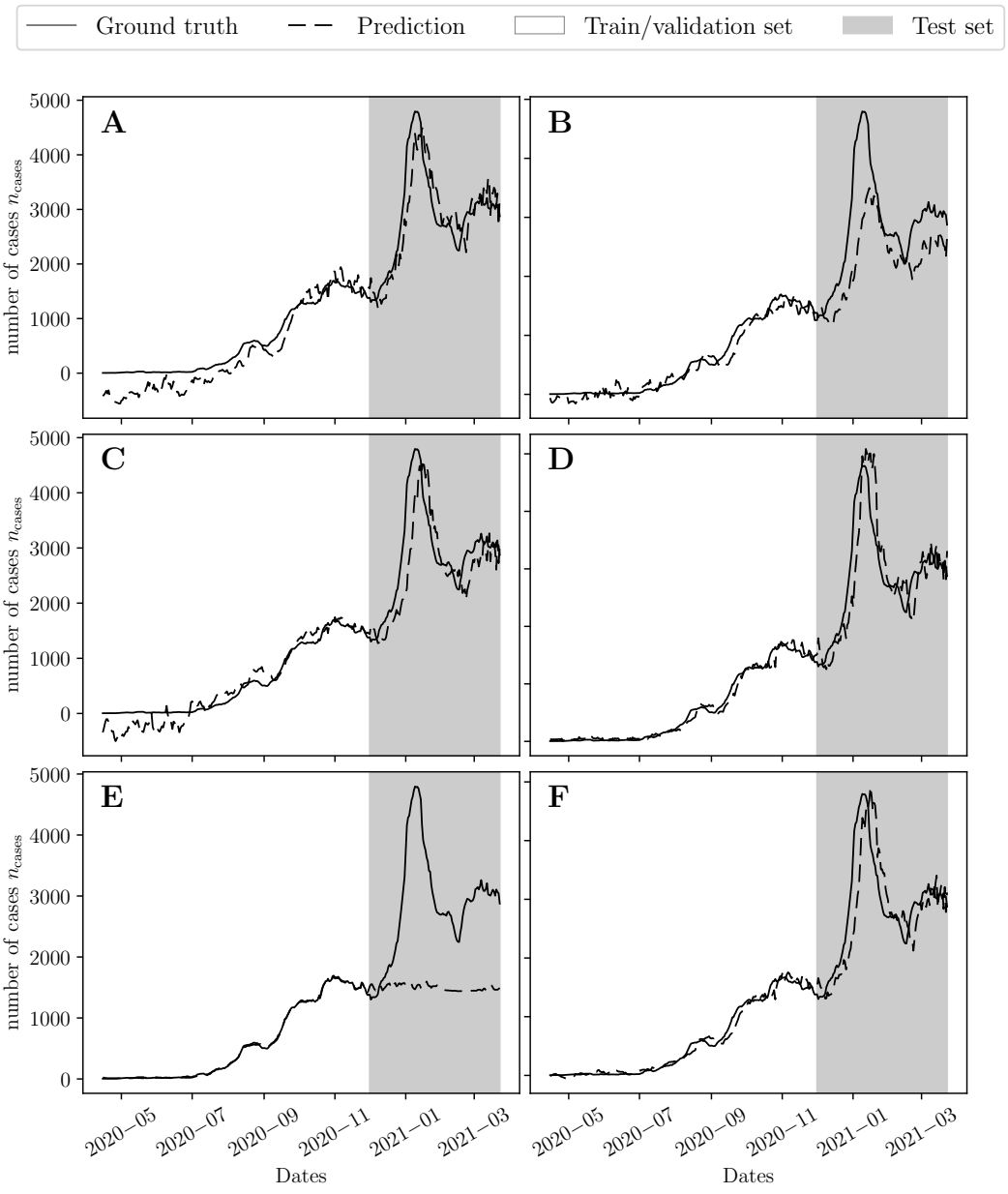


Figure S.V: Predicted 7-day rolling average of daily number of cases on the unseen data group using (A) the sequence-to-sequence (S2S) model, (B) the stacked LSTM (SEQ), (C) The feedforward neural network (DNN), (D) The support vector machine regression (SVR) model, (E) The gradient boosting machine (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.

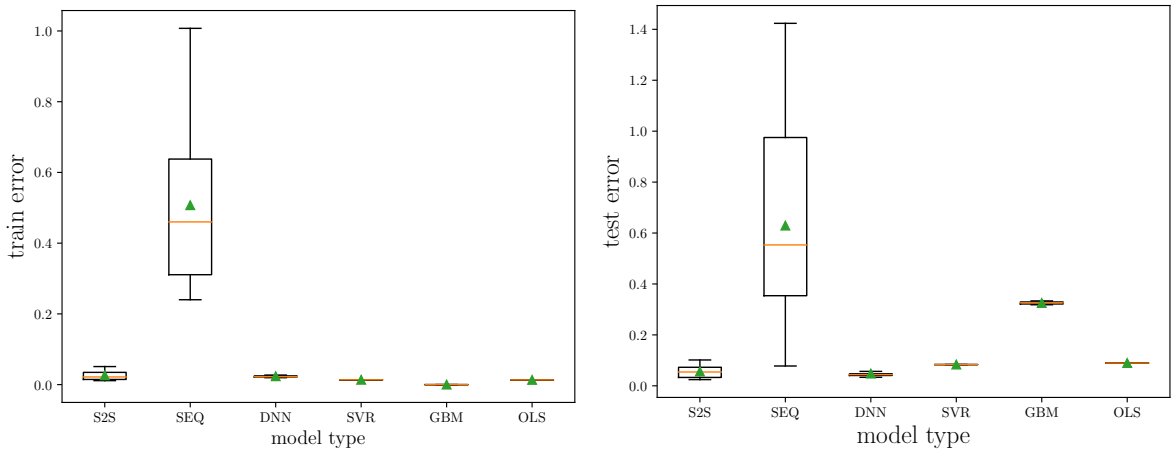


Figure S.VI: Illustration of variance in (a) training errors on discovery group (Group 1) and (b) test errors on the test group (Group 2) for different models. The errors were calculated using the MSE of the predicted and actual trajectories shown in Figure 5. The green triangles represent the mean error of 30 independent training runs for each model type. The orange lines represent the median error.

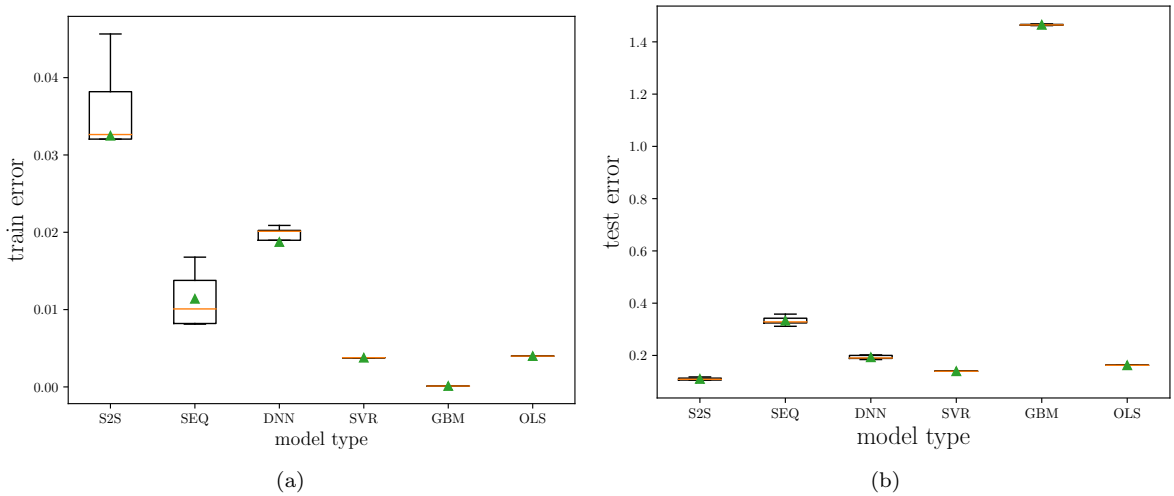


Figure S.VII: Illustration of variance in (a) training errors on combined discovery and test groups (Groups 1 and 2) and (b) the test errors on the unseen group (Group 3) for different models. The errors were calculated using the MSE of the predicted and actual trajectories shown in Figure S.V. The green triangles represent the mean error of 30 independent training runs for each model type. The orange lines represent the median error.

Table S.II: Optimal hyperparameters of models developed using combined discovery and test groups (Groups 1 and 2).

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}	0.0	0.0-0.9
Number of hidden layers	n_{hidden}	2	1-5
Teacher forcing probability	P_{teacher}	0.8	0.0-0.9
Learning rate	l_{rate}	1×10^{-4}	1×10^{-5}-1×10^{-2}
batch size	b_{size}	16	4-128
best epoch	$n_{\text{epochs}}^{\text{best}}$	16	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}	2	1-5
Learning rate	l_{rate}	1×10^{-4}	1×10^{-5}-1×10^{-2}
batch size	b_{size}	32	4-128
best epoch	$n_{\text{epochs}}^{\text{best}}$	15	1-n_{epochs}
Deep neural network (DNN)			
Sliding window size	T_1	6	1-40
Number of hidden neurons	n_{hidden}	1500	1-2500
Probability of dropout	P_{dropout}	0.3	0.0-0.9
Number of hidden layers	n_{hidden}	1	1-5
Learning rate	l_{rate}	1×10^{-4}	1×10^{-5}-1×10^{-2}
batch size	b_{size}	4	4-128
best epoch	$n_{\text{epochs}}^{\text{best}}$	9	1-n_{epochs}
Support vector machine regression (SVR)			
Sliding window size	T_1	11	1-40
Ridge factor	λ	1×10^{-4}	1×10^{-3}-1.0
Margin of tolerance	ϵ	0.01	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}	0.9	0.1-1.0
Maximum portion of features	f_{features}	1.0	0.1-1.0
Decision tree maximum depth	D	2	1-5
Learning rate	l_{rate}	0.01	1×10^{-5}-1×10^{-2}
Maximum number of boosting stages	n_{stages}	3000	50-5000
Polynomial regression (OLS)			
Sliding window size	T_1	11	1-40
Ridge factor	λ	1×10^{-4}	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.

S.IV Gaussian process model for RHUH patient cohort

We used the Gaussian process regression framework developed by Hay et al. to reconstruct the pandemic trajectory in Lebanon. We used a grid search to tune the priors on all the viral kinetics model parameters and Gaussian process parameters ν and ρ which control the bandwidth of the Gaussian kernel function. We attempted to minimize the MSE of the median predicted trajectory relative to the actual case counts and get a good estimate on the pandemic trajectory. We took the average of several runs to account for the randomness of Markov chain Monte Carlo (MCMC) sampling. Figure S.VIIIB shows the resulting predicted trajectory relative to the normalized case counts in Lebanon.

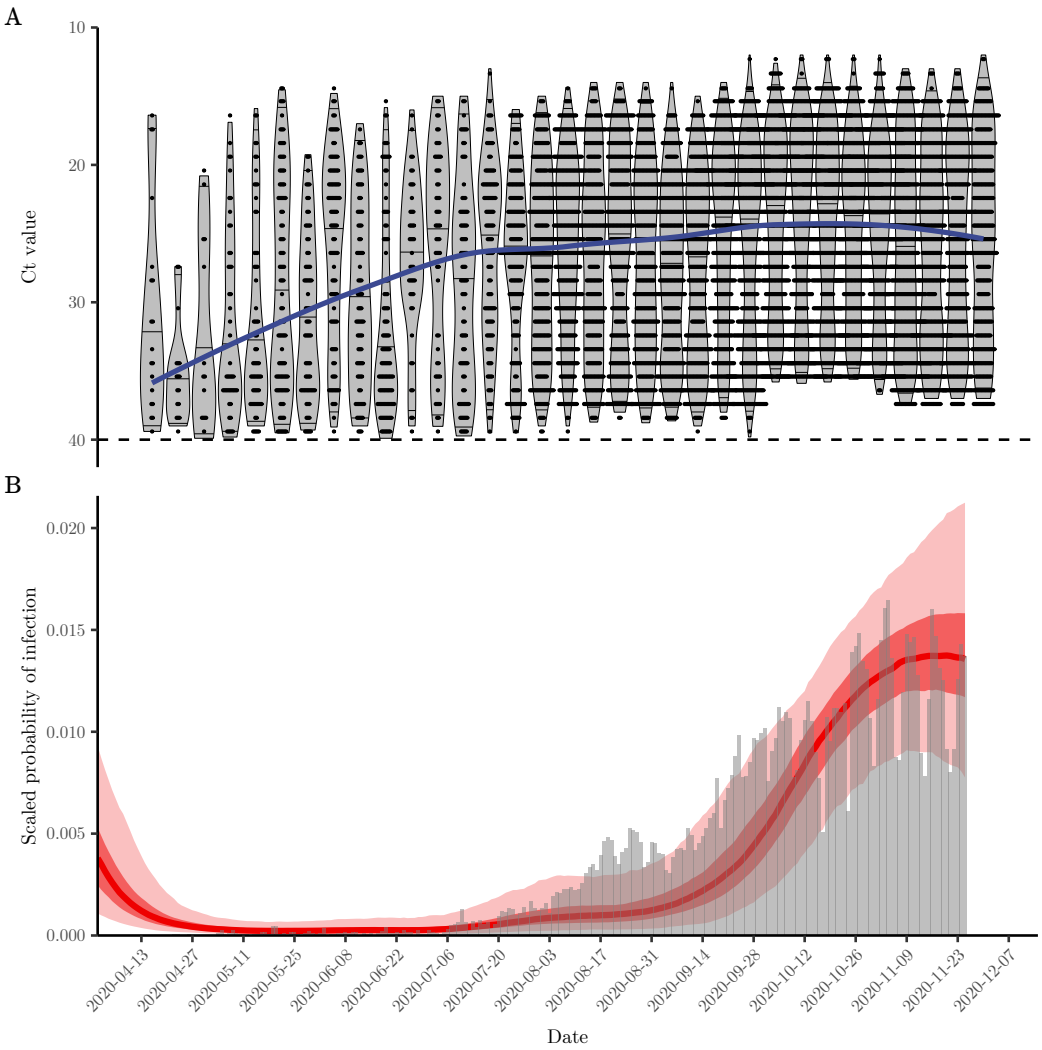


Figure S.VIII: Incidence rate and pandemic trajectory predictions using the predictive framework developed by Hay et al. [1] (A) shows the cross-sectional Ct samples (violin plots) and smoothed average (solid blue line) obtained from RHUH throughout the pandemic in lebanon. (B) Posterior distribution 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 (March 01, 2020 through November 30, 2020).

S.V Inferring the pandemic trajectory in Massachusetts using Brigham and Women's Hospital cross-sectional Ct data

Data collected from BWH by Hay et al. was used to test the performance of the models developed in this paper. Figure S.X shows the predicted incidence rates based on the Ct values observed at BWH (shown in Figure S.IXA). The deep learning models (Figures S.XA, B, and C) had less accuracy than both SVR and OLS models (Figures S.XD and F). Slightly biasing the Ct values resulted in better performance of all deep learning models (not shown in this paper) implying that they are very sensitive to fluctuations in Ct, which could be due to slightly different PCR machine calibration and/or specimen collection methods.

The SVR model shows very good performance on the BWH dataset but we advise caution when using such predictive models without prior cross-validation as is being done in this section.

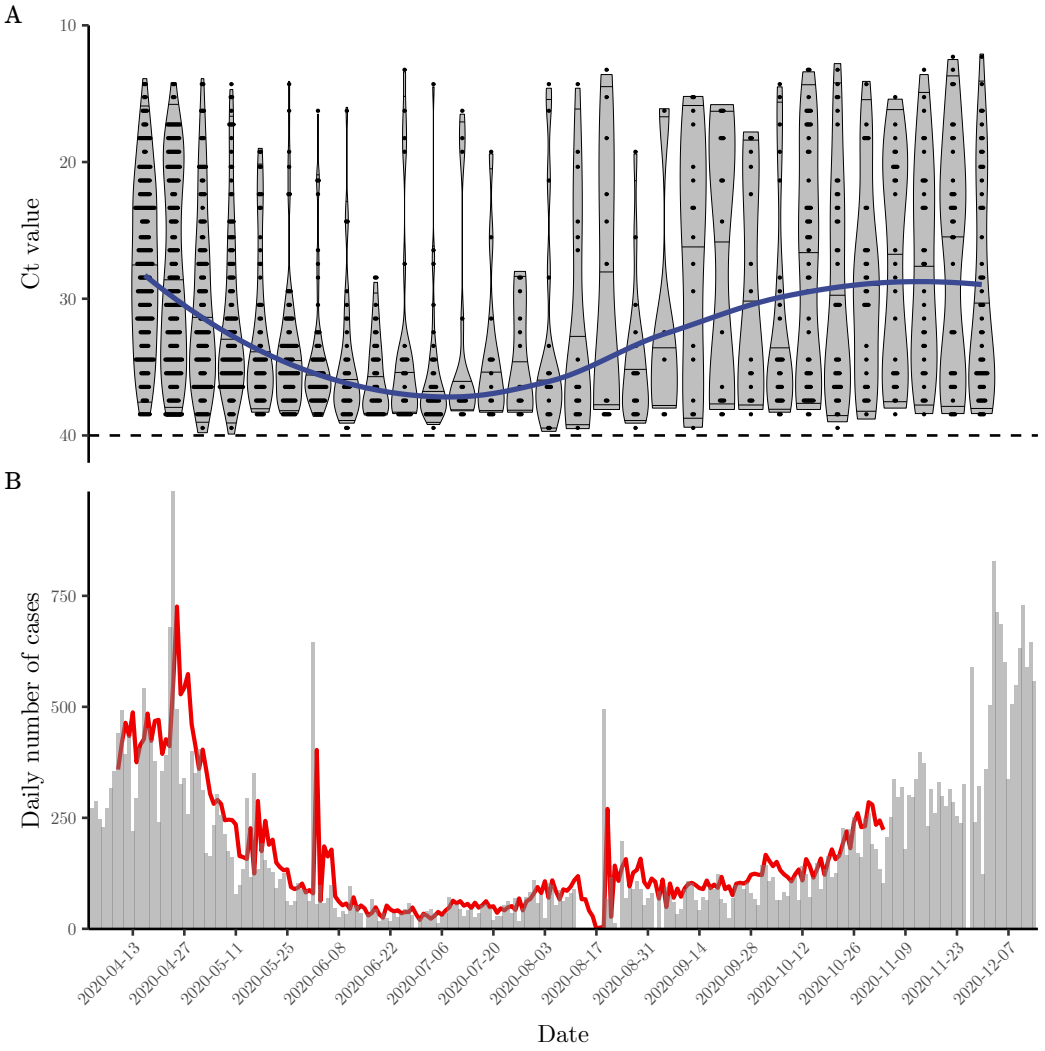


Figure S.IX: Incidence rate and pandemic trajectory predictions using the support vector machine regression (SVR) model (A) shows the cross-sectional Ct samples (violin plots) and smoothed average (solid blue line) obtained from Brigham and Women's Hospital (BWH) throughout the pandemic in Massachusetts. (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.

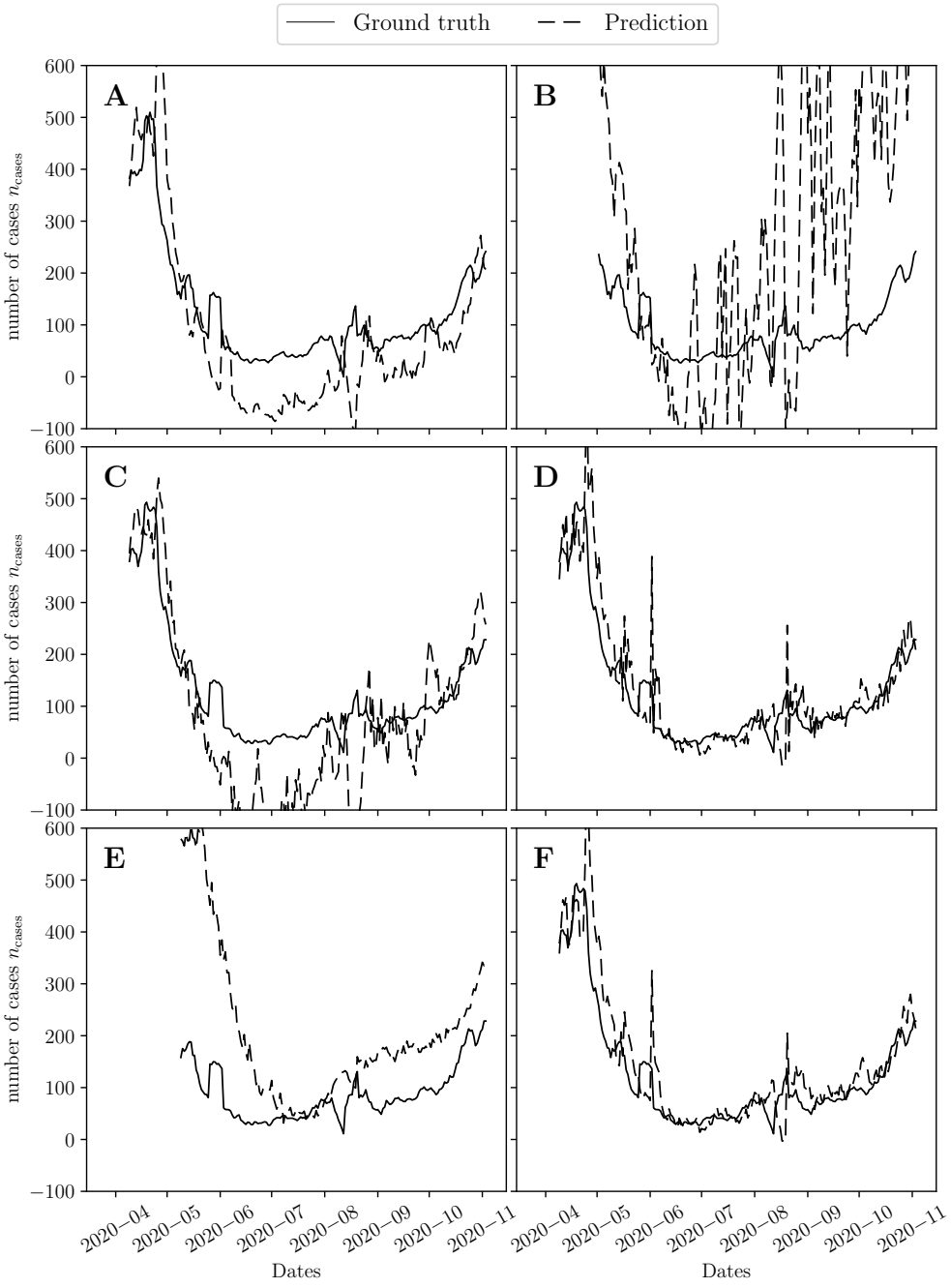


Figure S.X: Predicted 7-day rolling average of daily number of cases in Massachusetts predicted using (A) the sequence-to-sequence (S2S) model, (B) the stacked LSTM (SEQ), (C) The feedforward neural network (DNN), (D) The support vector machine regression (SVR) model, (E) The gradient boosting machine (GBM), and (F) the polynomial regression (OLS) model. The Ct values used in inference were obtained from Brigham and Women's Hospital (BWH) [1].

S.VI Deployment of the predictive model

We deployed the S2S model developed using the entire dataset (see supplementary material Section S.III) in a user-friendly interface and made it publicly available through https://covid-forecaster-lebanon.herokuapp.com [2]. The user-interface allows the user to enter the number of cases and Ct values observed for a certain number of days backward (which represents the optimal sliding window obtained through hyperparameter tuning). The S2S model is used to infer the predicted total number cases for the coming week (i.e., the average predicted case counts multiplied by 7). The data can be entered manually or copied from a spreadsheet. Continuous updates and patches will be applied to the dashboard to incorporate all the other models and provide additional visuals.

References

[1] 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, p. eabh0635. doi:10.1126/science.abh0635.

[2] COVID-19 weekly forecaster. https://covid-forecaster-lebanon.herokuapp.com. [Online; accessed 31-March-2022]