COVID-LIBERTY, A Machine Learning Computational Framework for the Study of the Covid-19 Pandemic in Europe. Part 2: Setting up the Framework with Ensemble Modeling

Nicholas Christakis[a,b,*], Panagiotis Tirchas[a], Michael Politis[a], Minas Achladianakis[a], Eleftherios Avgenikou[a], George Kossioris[a,b]
[a]Department of Mathematics and Applied Mathematics, University of Crete, Greece 700 13
[b]Laboratory of Applied Mathematics, University of Crete, Greece 700 13

*: Corresponding Author's email: nchristakis@tem.uoc.gr

Received: July 10, 2021. Revised: August 27, 2021. Accepted: August 30, 2021. Published: September 1, 2021.

Abstract—The Covid-19 pandemic has caused within a period of one year and eight months over 200,000,000 infections and more than 4,000,000 deaths. It is of paramount importance to design powerful and robust tools in order to be able to predict the evolution of the disease. In this paper, the computational framework COVID-LIBERTY is introduced, in order to assist the study of the pandemic in Europe. In Part 1, important parameters that should be taken into consideration and their parametrizations were given, as well as the details and mathematics of the computational engine of COVID-LIBERTY, a feed-forward, back-propagation Artificial Neural Network. In Part 2, the CPRT index is introduced, the framework setup around the Artificial Neural Network is presented and the algorithm of ensemble modeling is discussed, which improves the accuracy of the predictions. In the simulations, 4 European countries with similar population numbers were considered. The capabilities of the COVID-LIBERTY framework for accurate predictions for periods up to 19 days will be demonstrated.

Keywords—COVID-LIBERTY, Artificial Neural Network, ensemble modeling, parametrizations, CPRT

I. INTRODUCTION

The Covid-19 respiratory disease was declared a public health emergency by the World Health Organization (WHO) on January 20th 2020 [1] and has since affected over 200 countries worldwide with over 200,000,000 confirmed cases and more than 4,000,000 deaths [2]. The global economy has been badly affected by it, as well as all aspects of everyday life. Many scientists from various fields have been putting significant effort in understanding how the virus spreads, discovering effective treatment methods and developing a vaccine.

This past year, Machine Learning (ML) techniques have been extensively utilized in order to assist in Covid-19 predictions (see the introduction in [3]). Indicatively, we mention in Part 2 two of the most recent works found in literature. The authors in [4] used random forests in order to predict the slowdown of Covid-19, giving an indicative date of September 2021 for USA. The authors in [5] compared a number of ML algorithms (i.e., random forests, decision trees, K-nearest neighbors, artificial neural networks, gradient boosting machine, support vector machine) in order to determine important blood parameters which identify severe Covid-19 patients. With all potential disadvantages (poor quality data sets, poor application of ML methodologies, poor reproducibility and introducing bias in design, for a thorough review see [6]) ML remains a powerful tool, which, with proper use, will help in the fight against the Covid-19 pandemic.

In the present paper, the **COVID-LIBERTY** ML computational framework is presented: **C**ases **OV**er tests **ID**entifier-**L**ockdown **I**ndex **B**ased **E**nsemble modeling with effective **R**eproduction and **T**emperature variabilit**Y**. The framework is designed around a feed-forward back-propagation Artificial Neural Network (ANN), which takes into consideration parametrizations of all important factors, in order to predict the evolution of the disease in 4 European countries, namely Austria, Belgium, Greece and Portugal. The choice of countries and the data acquisition procedure have already been discussed in [3], as well as the details of the ANN, all important factors and their parametrizations. In the following sections, the CPRT index is introduced and the setup of the framework is described, with the appropriate parametrizations of crucial factors (i.e., lockdown, seasonality and effective reproduction). Then, results from simulations are presented from the Christmas 2020 period and the algorithm of **ensemble modeling** is introduced, which improves the accuracy of predictions. Finally, results are presented for the period of the third pandemic wave in Greece (March-April 2021) and conclusions are drawn on the ability of the framework to be utilized as a predictive tool.

It has to be noted that the ANN utilized in COVID-LIBERTY has been fully developed and extensively tested by the authors. Changes and modifications may be made directly at any stage in the code, in order to further optimize model performance and improve the accuracy of results. The code itself, is quite fast with all performed simulations (i.e., training, predictions) never exceeding 135 minutes run-time in Ubuntu Linux, version “16.04.07 LTS” environment, on an Intel Pentium E5200 dual-core processor with 2.5 GHz base frequency and 4Gb RAM.

Moreover, in the present work we do not compare our results to results of other similar models. We have used different input parameters and we have formulated new indices in order to parametrize what we believe are important factors that should be considered for Covid-19 predictions. Our aim is to provide the scientific community, health authorities and governing bodies with a tool which can be used for the prediction of future trends of the pandemic in order to assist in decision making. An evaluation of the performance of COVID-LIBERTY and its assessment in comparison with other ML models will follow at a later stage.

II. THE CPRT INDEX

In the proposed framework, a parameter is required, which represents and quantifies the effects of Covid-19, both at input and output. Following the discussion in [7], it is not advisable to utilize ANNs to predict absolute numbers and ask questions such as: “what is the price of a stock going to be tomorrow”. It is more advisable to ask questions regarding trends, that is whether, based on previous days’ trends and other relevant information, a stock price may rise tomorrow. Similarly, in this study we will not ask the framework to predict future numbers of infections, but rather if these numbers may be higher or lower compared to previous days. For this reason, in the patterns we will not consider the actual number of daily cases. It is clear that the number of cases strongly depends on the number of tests that have been performed on that day. For this reason, a new index will be introduced for each country to quantify the effect of the disease, the number of daily reported cases **divided by** the number of daily Covid-19 tests. Let us call this index *Cases Percentage Relative to Tests* (CPRT). In this study, its 7-days moving average values will be utilized. This index has the advantage of coupling the number of infections with the number of tests hence, the question asked to the ANN (given the fact that 7-days moving average values are used) is “what percentage of Covid-19 tests **on average** will turn out positive”. If an average number of positive cases is explicitly requested for a particular period, it could be easily retrieved from CPRT by knowing an estimate for the number of performed Covid-19 tests during that period. It is expected, due to the 7-days moving averages employed in this study, that the calculated percentage of positive tests could be underpredicted when the infections were sharply increasing or overpredicted when the infections were sharply decreasing. Another advantage of CPRT is that it takes values between 0 and 1 and does not require further normalization in an ANN. In fact, it was observed that for all countries in this study, the 7-days moving average values of CPRT never exceeded 0.35, which meant that at all stages of the pandemic so far, for the countries involved in this study, not more than 35% of tests were reported on average as positive. The suitability of CPRT

DOI: 10.46300/91016.2021.8.4

as a good measure of the trend of the disease is depicted in Figure 1, where results are presented until December 20, 2020. CPRT can be seen to follow very closely the number of reported cases with time. It was found through analysis of the data that for all countries the peaks of CPRT and reported cases were at most one day apart. As observed in Figure 1, at the beginning of the datasets, there was a peak in CPRT for most countries, comparable to the second peak of November. This was due to the fact that at the beginning of the pandemic the number of performed tests was limited, hence the high CPRT. It has to be noted that Sweden was not considered in COVID-LIBERTY predictions, since we did not have adequate data regarding testing and thus, no CPRT index could be calculated.

III. THE COVID-LIBERTY FRAMEWORK

COVID-LIBERTY was designed around a feed-forward, back-propagation ANN, the details of which were given in Part 1 of this work [3]. Appropriate patterns needed to be formulated, which would include all factors that have been shown to be of importance in the spread and evolution of Covid-19. As has been demonstrated in Part 1, these are the effective reproduction number, lockdown and seasonal variability. Changes in these factors have been shown to heavily influence the number of reported cases. Their parametrizations have been detailed in [3]. In the previous section, the CPRT index was introduced as the appropriate parameter to describe in both input and output the effects of the virus. Hence, as inputs of the model, all these parameters should be included (namely, CPRT, lockdown, seasonality and effective reproduction) and a new future value of CPRT should requested at output. In order to determine the inputs and outputs of the model, various configurations of the data were tested. Based on experience [8] and extensive testing, it was found that the most appropriate configuration, which resulted in the smallest error during training, was that with 8 inputs and 1 output. The inputs were: CPRTs of 5 consecutive days, a parametrized temperature corresponding to the output day (as discussed in Section 3 of [3]), Lockdown Index (LI) of the output day (as discussed in Section 4 of [3]) and R_t of the output day (as discussed in Section 3 of [3]). The output was the CPRT of the 6th day. Initial tests did not include R_t. However, its inclusion improved significantly the accuracy of predictions, since its variations clearly defined the trend for future evolution of the disease. This is demonstrated in Figure 2, where the evolutions of CPRT and R_t are presented. As can be observed, first comes the peak of R_t, with the one of CPRT following (note how the main peaks of CPRT lag behind the equivalent peaks of R_t by at least 13 days for all examined countries).

Since R_t indices were available from March 19, 2020 onwards, the first available 7-days moving average for R_t was for March 25, hence that was the first date for which predictions could be made and that was the first pattern to be formulated. That was the case for Belgium and Greece. For Austria, data for performed tests were available after April 2, 2020, hence the first available 7-days moving average for CPRT was for April 8 and that was the first date for which predictions could be made for Austria. For Portugal, it was observed that at the very beginning of the dataset, R_t decreased sharply from its highest to its lowest value (without reaching again its maximum value at any point in the available data, see Figure 4 in [3]). Since that occurred during the outburst of Covid-19 in Europe, when knowledge and experience were limited, it was decided to avoid in Portugal's training dataset the first 9 days (that is, half of the time interval it took for R_t to decrease from maximum to minimum). Hence, the training dataset for Portugal started from April 2, 2020 onwards. Patterns with 8 inputs and 1 output were formulated for all countries with data up until December 20, 2020. In total, for Austria 255 such patterns were formulated, for Belgium and Greece 271 patterns and for Portugal 262 patterns. The formulated patterns were randomly split for each country into two groups: two thirds of the data to be used for training and one third to be used for validation, i.e. check how well the framework performed in predicting the 6th CPRT for data that was not trained for.

A. TRAINING AND VALIDATION OF COVID-LIBERTY

For the training phase of COVID-LIBERTY the final errors (as percentage of the initial errors) after 5,000,000 epochs were 0.13%, 0.69%, 0.13% and 0.49% for Austria, Belgium, Greece and Portugal, respectively. Then, we proceeded to validate the ANN with the one third of the dataset, which was retained for that purpose, utilizing the calculated synaptic weights for every country. During validation, patterns were used which the ANN was not trained for, in order to predict for each set of input data the 7-days moving average of CPRT of the 6th consecutive day. Relative errors between predicted and true values of CPRT were calculated and the results are presented in Table 1. All relative errors in this work were calculated based on the formula:

$$\text{Relative Error} = \frac{\left|CPRT_{true} - CPRT_{pred}\right|}{CPRT_{true}} \times 100$$

Table 1. Relative errors between true and predicted values from the validation procedure of COVID-LIBERTY.

Country	Max Relative Error (%)	Min Relative Error (%)	Mean (%)	Median(%)
Austria	56	0.75	18.3	17.14
Belgium	75.9	0.04	10.88	5.69
Greece	90.84	0.01	9.45	2.8
Portugal	46.23	0.1	6.51	4.44

As observed, there were instances where the framework miscalculated the CPRT values, however these instances were extremely few, as shown by the medians and means of the relative error distribution. Taking into consideration the fact that all mean and median relative error values were well under 20% (in most cases, under 10%), the validation was considered as successful.

We now proceed in the following section to simulate for all 4 countries the disease evolution for the Christmas period, i.e. between December 21, 2020 and January 10, 2021. This was a period for which the COVID-LIBERTY framework had not been trained for. In this section, the introduction of the ensemble modeling algorithm in the framework will be described and its performance assessed.

B. MODEL PREDICTIONS AND THE ENSEMBLE MODELING ALGORITHM

In order to ensure that all available information would be utilized, it was decided to perform a new training for every country with a complete dataset until December 2020. New sets of synaptic weights were obtained and these were used in COVID-LIBERTY for the predictions. The final errors of training (as percentage of the initial errors) after 5,000,000 epochs were 0.44%, 0.51%, 0.15% and 0.69% for Austria, Belgium, Greece and Portugal, respectively. These errors were not necessarily smaller than the previous errors, when training was performed for the two thirds of each dataset, since we may have introduced further variability with the addition of this data. However, as can be seen, they were of the same order of magnitude.

The patterns for the predictions were formulated in such a way, so that when we started with known information to predict the 1st CPRT (for December 21, 2020) all information was available, i.e. the CPRTs of the previous 5 days were known, as well as temperature, LI index and R_t value. However, as we proceeded to the following day (December 22, 2020), the predicted value of CPRT for December 20 was required to formulate the new inputs for the framework. Then for December 23, the outputs of the framework for December 21 and 22 were required, and so on. That is, the predicted CPRTs at every step were utilized in the input patterns in order to predict the following ones. It has to be noted that from December 26, 2020 onwards, all following predictions were made with only predicted CPRT values.

The required as input values of parametrized country temperatures already existed for the first 7 days to predict (see discussion in Section 3 of [3]). From then onwards, prognostic data would have to be used to formulate mean temperatures. Since it is established that reliable predictions for typical weather patterns may be obtained for up to two weeks [9], predictions for CPRT should only run for a collective period of 21 days (7 days existing data plus 14 days prognostic data). This explains why we only choose periods of 21 days in total to predict the CPRT trends in the 4 different countries. For the scope of this study, in all following sections temperature parametrizations were derived directly from available meteorological data, even for the last 14 days of simulations (i.e., when the prognostic values should have been used), knowing that if we had used prognostic data, these would be in very good agreement with the actual observed temperatures [9]. If better temperature predictions are required for future applications (i.e., for temperatures which cannot be retrieved from available predictions [10]), there exists the capability of running custom simulations with the use of the WRF numerical model, in order to obtain such data [11].

In the input patterns the LI parameter was also required, which was assumed that could be reliably established for a period of 21 days for each country. In the current simulations, the true values of LI were used.

As final input, the 7-days moving average values of the R_t index for every country were required. As may be seen from the dataset (Figure 4 in [3]), there was no clear trend in the variation of R_t towards the end of 2020. Hence, it was decided to perform 2 different simulations for every country, in order to test the dependence of the predictions on R_t variations. Namely, one simulation where the true values of R_t were used for each day and one simulation where R_t was kept fixed to its last known value (i.e., for December 20, 2020). Last known 7-days moving average R_t values were 1.01, 0.97, 0.93 and 0.96 for Austria, Belgium, Greece and Portugal, respectively.

The results where the true values of R_t were used are presented in Figure 3, where comparisons between the true and predicted values of 7-days moving averages CPRT for all countries are shown. Figure 4 presents the relative error between true and predicted values for every day of the chosen time period for all countries. As may be seen in Figure 3, the trend of CPRT was followed remarkably well in Austria, Belgium and Greece, indicating tendency of CPRT to increase in Austria, to remain constant in Greece and to decrease in Belgium. Only for Portugal the trend seemed not to be correct, with the ANN predicting decrease of CPRT and the actual data showing slight increase. In Figure 4: Austria displayed 3 consecutive peak values above 50% after the first 6 days (where only predicted CPRT values were employed), however the ANN managed to recover after that and the error dropped significantly (average relative error of 30.2%); for Belgium and Greece the relative error was kept on average quite low (Belgium 16.55%, Greece 22.38%), indicating quite good not only qualitative but also quantitative agreement as well; for Portugal, the relative error may be seen to follow an increasing trend (average error of 49.43%). The lack of good agreement for the case of Portugal may be attributed to the fact that although the November peak of CPRT lagged behind an R_t peak by approximately 15 days for Austria, Belgium and Greece, the same was not true for Portugal. It lagged behind by 1 month (see Figure 2). This means that COVID-LIBERTY was trained to understand changes in R_t at a slower rate for Portugal, hence not being able to adjust in time and reverse the trend of R_t from decreasing to increasing fast enough. This difference between Portugal and the three other countries could be due to the fact that the situation with lockdown in Portugal was not very clear during that period; some restrictions were in place, but not to an extent to be classified as soft lockdown (see e.g. [12]). The formulation of LI in [3] does not account for such cases (only hard or soft lockdowns are considered and the respective transitional periods). A possible solution would be to account for such restrictions with a low LI value (lower than soft lockdown LI values). The Portugal effect clearly indicates how important the inclusion of lockdown information in predictive models is, in order to be in a position to obtain accurate results.

Since the variation of R_t cannot be known *a priori*, it was deemed appropriate to run a simulation with R_t kept fixed to its last known value (i.e., that of December 20, 2020) for each country. The results are presented in Figures 5 and 6 (CPRT values and relative errors). Even with fixed R_t values, the predicted CPRT trends were in very good agreement with the true CPRT behaviour, except for Portugal (the reasons have already been explained). In particular, for the case of Austria it improved significantly the relative error compared to that when true R_t variation was used (average error 20.39%). For Belgium though, the average error increased to 30.58%. However, the CPRT trend was again very well predicted). For Greece the average error remained almost constant (22.48%). All average relative errors, from both simulations, are given in Table 2 for direct comparisons. As seen, we cannot definitively conclude which simulation (true or fixed R_t values) gave better results.

Table 2. Average relative errors between true and predicted CPRT values for COVID-LIBERTY predictions.

Country	**Average Relative Error (%)**	
	True R_t values	**Fixed R_t values**
Austria	30.20	20.39
Belgium	16.55	30.58
Greece	22.38	22.48
Portugal	49.43	46.70

Obviously, it would be ideal to know or guess the trend of R_t i.e., would it increase or decrease within the 21 day period we have chosen to study. For this reason, we put forward to proposition of **ensemble modeling**. Ensemble modeling is a technique commonly used in meteorology, where a weather prediction model is initiated many times with different sets of conditions. Then, the results of all simulations are averaged, in order to obtain a prediction (see e.g. [13]). A similar idea is proposed and incorporated in the COVID-LIBERTY framework. Since the variation of R_t is not known *a priori*, at least a **trend** of it may be incorporated in the results if we choose to form **ensembles** (i.e., sets of consecutive days), run different simulations starting on each of these days (predicting thus, 21 days ahead each time) and then average the results. The algorithm of ensemble modeling in COVID-LIBERTY works as follows:

Suppose we want to predict from December 21, 2020 onwards, with the last data having been recorded on December 20, 2020. Let us name this configuration as "**18-19-20 December ensemble configuration**". Also, take note that predictions may be made only for 21 days ahead, as already discussed.

1) The model is run with fixed R_t (with the CPRT of December 20, 2020 being the last of the 5 consecutive CPRTs in the inputs and, subsequently, R_t of December 20, 2020 being the R_t value in the inputs). The model predicts for 21 consecutive days (i.e., until January 10, 2021).
2) The model is run for a second time with fixed R_t, now starting for the previous day (i.e., CPRT of December 19, 2020 being the last of the 5 consecutive CPRTs in the input parameters and, subsequently, R_t of December 19, 2020 being the R_t value in the inputs) and predicts for 21 consecutive days (i.e., from December 20, 2020 to January 09, 2021).
3) The model is run for a third time with fixed R_t, now starting from 2 days before the date in step (1) (i.e., CPRT of December 18, 2020 being the last of the 5 consecutive CPRTs in the input parameters and, subsequently, R_t of December 18, 2020 being the R_t value in the inputs) and predicts for 21 consecutive days (i.e., from December 19, 2020 to January 08, 2021.
4) The common data of all three different runs are averaged and a prediction is obtained for the **19 days** that all three simulations share in common i.e., from December 21, 2020 until January 08, 2021. The averaging is performed through **weighted averages**, considering that the emphasis should be placed on the latest rather than the earliest day, through the formula:

$$CPRT_{average} = \frac{1}{6}CPRT_{earliest} + \frac{2}{6}CPRT_{\text{intermediate}} + \frac{3}{6}CPRT_{latest}$$

In this way, predictions are obtained, which take into consideration changing trends of R_t (from fixed R_t values), with the emphasis always being placed on the latest value, thus forming a trend. If for instance R_t is increasing (decreasing), this will be reflected in the results accordingly, since the prediction with the higher (lower) R_t value has the highest impact in the results. The period of three consecutive days was chosen so, in order to represent the changes in R_t, but without compromising the model prediction confidence interval of 21 days by much. The validity of this assumption will be demonstrated in the results of the simulations. The results of ensemble modeling for all 4 countries (CPRT and relative errors), for the period between December 21, 2020 and January 08, 2021 are presented in Figures 7 and 8. For 3 countries (Austria, Belgium and Greece), the CPRT trend is predicted remarkably well by the model for the period of 19 days. Again, for Portugal there was a failure of the model to predict the correct trend for reasons we have discussed earlier. There is also an improvement in the predictions of the framework, with relative errors for all countries being lower than in the case when a single simulation with fixed R_t values was performed. Table 3 summarizes the results for averaged relative errors between the two different approaches.

Table 3. Average relative errors between true and predicted CPRT values for COVID-LIBERTY predictions.

Country	Average Relative Error (%)	
	Ensemble modeling, fixed R_t values	**Single simulation, fixed R_t values**
Austria	20.09	20.39
Belgium	27.72	30.58
Greece	22.12	22.48
Portugal	41.65	46.70

From the presented results, it could be concluded that the COVID-LIBERTY framework with the algorithm of ensemble modeling performed remarkably well, taking into consideration the fact that data about Covid-19 infections are available only since February-March, 2020 hence, not many training patterns could be formulated. At least for three countries, the framework managed to accurately represent the evolution of the disease for a period of 19 days. Since the examined period was one with low numbers of infections, we decided to test the performance of the framework during the surge of the third pandemic wave in Greece. Greece was chosen for this study because data were readily available and also, due to the fact that the third wave lasted for a longer period of time thus, giving the opportunity to test the framework during all stages of this period (i.e., increase, peaking and decrease of infections).

IV. MODELING THE THRID PANDEMIC WAVE IN GREECE

The third wave of Covid-19 in Greece came in late February 2021, peaked around early April 2021 and started slowly receding afterwards (see Figure 9 for the observed CPRT and daily reported cases). It has to be noted that the 7-days moving average of the daily reported cases started falling below 40% of its April peak value only after the beginning of June 2021 [14]. COVID-LIBERTY with ensemble modeling was employed in order to predict for the period encompassing the third pandemic wave, ranging from March 1 to April 28, 2021. The framework was trained with all data collected until February 28, 2021. It has to be noted that for all subsequent simulations the synaptic weights from this training were used. Then, the whole period was split into 3-day ensembles to predict for 19 days ahead every time. It is necessary to do so, since it is very important to represent the correct R_t trend. And this may only be achieved by updating R_t values every day and performing new ensemble simulations. In this way, for the first prediction (from March 1 to March 19) the ensemble modeling for 26-27-28 February was utilized; for the second prediction (from March 2 to March 20) the ensemble modeling for 27-28 February-1 March was utilized, and so on. In the final prediction (in order to predict from April 10 to April 28), the ensemble modeling of 7-8-9 April was utilized. 41 such ensembles were formulated and predictions for 41 19-days intervals were made. All input patterns were designed as described in the previous section. The whole set of results (CPRT comparisons between true and predicted values and relative errors for all 19-days intervals) are given in the Supplementary Material link that accompanies the paper [15]. All 41 ensembles are split into two different directories (one containing CPRT values, true and predicted, and the other containing the equivalent relative errors) and are numbered consecutively from 01 to 41; 01 corresponding to the 26-27-28 February ensemble with prediction between March 1 and March 19; 02 corresponding to 27-28 February-1 March ensemble with prediction between March 2 and March 20; …; 41 corresponding to 7-8-9 April ensemble with prediction between April 10 and April 28.

The results of these simulations were remarkably good. The model in all instances represented accurately the trend of the infection, with extremely good agreement, both qualitative and quantitative. As shown in Table 4, which presents the statistics of the average relative error for the 41 ensembles, all of the predicted 19-days trends had average error less that 25% and 50 % of these displayed error below 15%!

Table 4. Statistics of average relative error for the 41 ensembles, during the third pandemic wave in Greece

Average Relative Error Distribution	< 10%	<15%	<20%	<25%
Number of ensembles	0	21	33	41

Indicatively, in Figure 10 we present comparisons between true and predicted CPRT values from 3 ensembles, which collectively cover the whole examined period (i.e. from March 1 to March 19-ensemble 1, from March 20 to April 7-ensemble 20 and from April 8 to April 26-ensemble 39). Figure 10(a) presents the CPRT values and Figure 10(b) the corresponding relative errors. The model seems to follow nicely the increasing trend of CPRT for the whole of March. A maximum is reached at the beginning of April (a value of 0.092 for both predicted and true-April 2 for predicted and April 10 for true CPRTs). From then onwards, CPRT is decreasing, reaching gradually a predicted value of 0.057 at the end of the simulation (corresponding to approximately 60% of its peak value, in accordance with the behaviour of the daily reported cases, as recorded in [14]). The average relative error between true and predicted CPRT values for each ensemble is: 17% for ensemble 1, 11% for ensemble 20 and 20% for ensemble 39. This fact indicates that the computational framework managed to reproduce quite accurately not only the behaviour of CPRT, but its actual values as well.

In this way, we may conclude that the designed COVID-LIBERTY framework has operated very well in predicting the trends of the disease under especially complex conditions (the whole period of the third pandemic wave in Greece) and has the potential to be utilized as an operational tool to assist in the fight against Covid-19.

V. CONCLUSIONS AND FUTURE WORK

In this study, the COVID-LIBERTY framework for the study of the Covid-19 pandemic in Europe was presented and the algorithm of ensemble modeling was introduced. Data from 5 European countries with similar populations were gathered, in order to test the framework. In Part 1, the details of the ANN which forms the main computational engine of COVID-LIBERTY were given. Moreover, all important factors were determined, which must be taken into consideration when predicting the disease evolution (lockdown, seasonality, effective reproduction). The importance of these factors was clearly demonstrated and appropriate parametrizations were developed, in order to be utilized in COVID-LIBERTY (i.e., the Lockdown Index and an effective country temperature corresponding to infection date).

In Part 2, the CPRT index was introduced as a parameter which characterizes the evolution of Covid-19. The setup of COVID-LIBERTY was discussed and the algorithm of ensemble modeling was introduced. Ensemble modeling enabled the framework to better consider changing trends in the values of R_t, which cannot be known a priori. It was found that COVID-LIBERTY could follow effectively the infection trend in most countries in this study and a validity range of 19 days for the model predictions was established. Further testing of the framework for the case of the third pandemic wave in Greece showed that COVID-LIBERTY may be effectively utilized on a day-to-day basis to predict quite accurately the evolution of the disease. We believe that the COVID-LIBERTY framework has performed remarkably well, given the limitations that exist in the datasets (341 patterns was the maximum set that could be formulated for the training of the framework, for the case of the third pandemic wave in Greece). The developed framework may be utilized by scientists, governing bodies and health authorities for predictions / planning and constitute a useful tool in the fight against Covid-19.

To the best of our knowledge, COVID-LIBERTY is a unique approach for studying the evolution of the pandemic. It incorporates a number of important parameters (their importance has been shown in Part 1 of this work). All parameters may be obtained directly from publicly available datasets. Moreover, the algorithm of ensemble modeling is coupled within the framework with a feed-forward, back propagation ANN in order to improve the accuracy of predictions.

Since the framework is fully designed and developed by the authors, it is possible to update all model components (in order to make it more efficient) and also, examine the impact of new parameters on disease evolution, e.g. vaccination rates. So far, vaccination rates were not considered, since for the countries that were involved in this study, we are well below herd immunity threshold at the time of completion of this work (see e.g. [14]). It is believed that the effect of the Delta variant (see e.g. [16]) will be incorporated directly in the results through the R_t rate. As more data become available, it is also believed that the accuracy of the framework will improve. Moreover, the model may be scaled down and tested at a more local scale (i.e. a region within a country), if appropriate input local data are available.

REFERENCES

[1] C. Sohrabi, Z. Alsafi, N. O'Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis, R. Agha, World Health Organization declares global emergency: A review of the 2019 novel coronavirus ({COVID}-19), International Journal of Surgery, 76 (2020) 71-76

[2] Worldometers, Information on Coronavirus and European countries statistics https://www.worldometers.info (accessed 19 August 2021)

[3] N. Christakis, M. Politis, P. Tirchas, M. Achladianakis, E. Avgenikou, C. Kalafati Matthaiou, M. Kalykaki, A. Kyriakaki, P. Paraschis, E. Pilios, G. Kossioris, COVID-LIBERTY, A Machine Learning Computational Framework for the Study of the Covid-19 Pandemic in Europe. Part 1: Building of an Artificial Neural Network and Analysis and Parametrization of Key Factors which Influence the Spread of the Virus. Submitted to: International Journal of Neural Networks and Advanced Applications (2021)

[4] Z. Malki, E-S. Atlam, A. Ewis, G. Dagnew, O.A. Ghoneim, A.A. Mohamed, M.M. Abdel-Daim, I. Gad, The COVID-19 pandemic: prediction study based on machine learning models. Environ Sci Pollut Res (2021). https://doi.org/10.1007/s11356-021-13824-7

[5] A. Aktar, M.M. Ahamad, M. Rashed-Al-Mahfuz, A. Azad, S. Udin, A. Kamal, S.A. Alyami, P. Lin, S.M.S. Islam, J.M. Quinn, V. Eapen, M.A. Moni, Machine Learning Approach to Predicting COVID-19 Disease Severity Based on Clinical Blood Test Data:Statistical Analysis and Model Development. JMIR Med Inform, 9, 4, e25884 (2021). doi: 10.2196/25884

[6] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A.I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J.R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, AIX-COVNET, J.H.F. Rudd, E. Sala, S.-B. Schönlieb, Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence 3 (2021) 199-217

[7] N. Christakis, V. Barbaris, A. Spentzos, A New Approach in Financial Modelling with the Aid of Artificial Neural Networks. Journal of Algorithms and Computational Technology, 5, 3 (2011) 513-529

[8] N. Antoniou, A Big Data Analytics Application in Transportation and Astrophysics, MSc Dissertation,

Department of Physics, University of Crete, Greece, 124 pages (2017)

[9] F. Zhang, Y.Q. Sun, L. Magnusson, R. Buizza, S.J. Lin, J.H. Chen, K. Emanuel, What is the predictability limit of midlatitude weather? Journal of the Atmospheric Sciences, 76 (2019) 1077-1091

[10] Accuweather, Predictions of future weather conditions https://www.accuweather.com (accessed 27 July 2021)

[11] N. Christakis, T. Katsaounis, G. Kossioris, M. Plexousakis, On the performance of the WRF numerical model over complex terrainon a high performance computing cluster. In: Proceedings of 2014 IEEE International Conference on High Performance Computing and Communications (HPCC), Paris, France, (2014). DOI: 10.1109/HPCC.2014.53.

[12] Wikipedia, lockdown information retrieval https://www.wikipedia.org (accessed 15 May 2021)

[13] H. Zhang, Z. Pu, Beating the Uncertainties: Ensemble Forecasting and Ensemble-Based Data Assimilation in Modern Numerical Weather Prediction. Advances in Meteorology, 2010, 10 pages (2010). doi: 10.1155/2010/432160

[14] Reuters Covid-19 Tracker, general Covid-19 information for countries worldwide https://graphics.reuters.com/world-coronavirus-tracker-and-maps (accessed 19 August 2021)

[15] Supplementary Material, CPRTs and Corresponding Relative Errors for 41 ensembles http://users.math.uoc.gr/~nchristakis/Supplement_Covid/ (accessed 27 July 2021)

[16] L. Di Domenico, V. Colizza, Epidemic scenarios of Delta variant in France in the summer 2021. Report #31, www.epicx-lab.com (2021).

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Nicholas Christakis conceived the idea, wrote the code in Fortran77 for the ANN, organized all simulations and wrote the paper.

Panagiotis Tirchas performed all simulations and analyses for the ensemble modelling and the 3rd pandemic wave.

Michael Politis retrieved all weather data and produced all graphs.

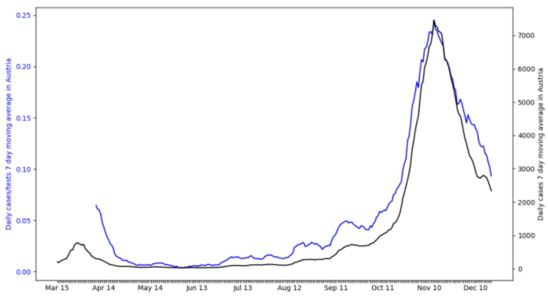
Minas Achladianakis and Eleftherios Avgenikou analysed the data for the initial training and validation of the ANN during the 2nd pandemic wave.

George Kossioris helped in organizing the layout of the paper and simulations and wrote the conclusions.

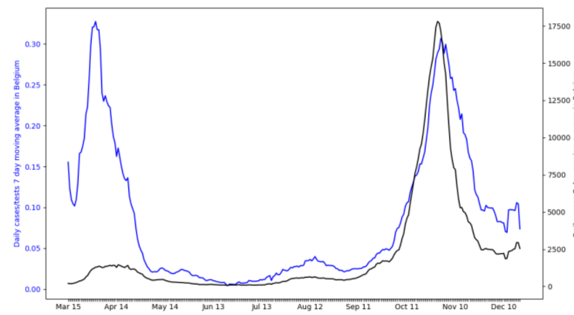
Sources of funding for research presented in a scientific article or scientific article itself

This work was not performed under any specific project and did not receive any grant.

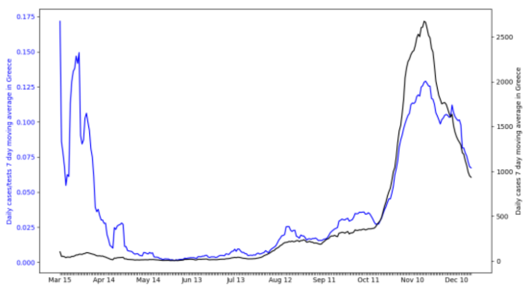
Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)



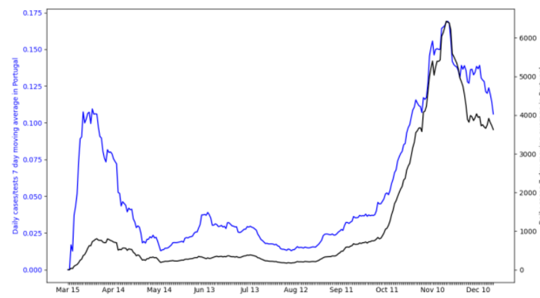
(a)



(b)

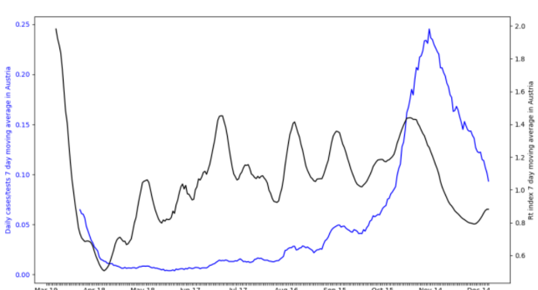


(c)

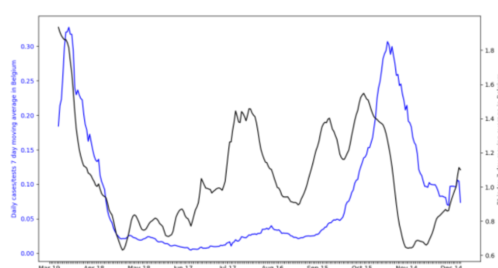


(d)

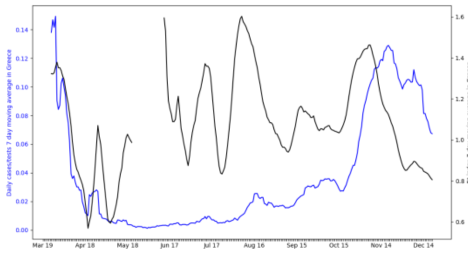
Figure 1. 7-days moving averages of daily cases/tests (CPRT-right) and daily reported cases (left) for the data for the period between March 15 and December 20, 2020 for (a) Austria, (b) Belgium, (c) Greece and (d) Portugal.



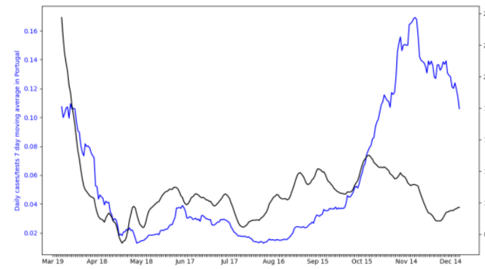
(a)



(b)

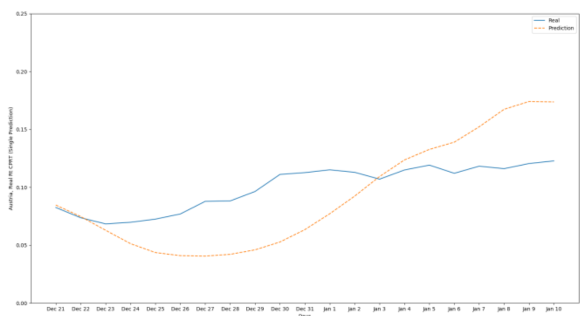


(c)

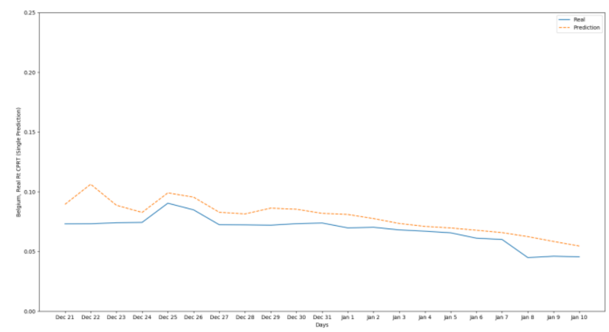


(d)

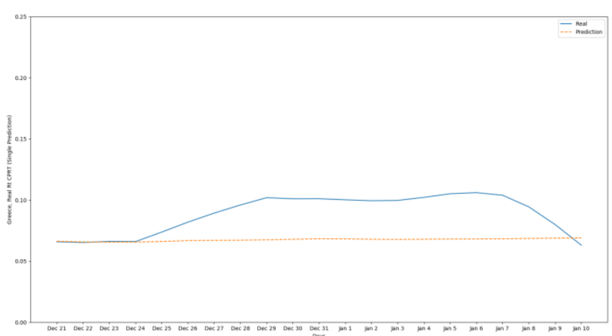
Figure 2. 7-days moving averages of CPRT and R_t for the period between March 15 and December 20, 2020 for (a) Austria, (b) Belgium, (c) Greece and (d) Portugal.



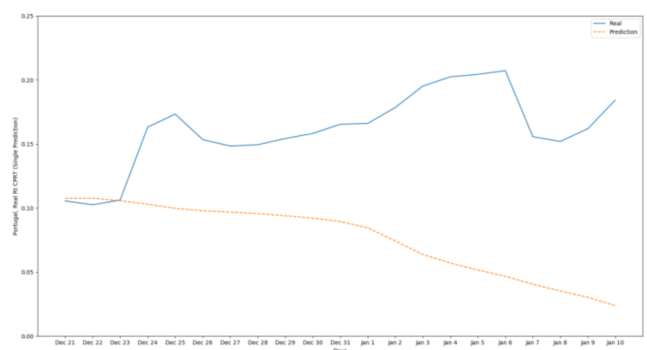
(a)



(b)



(c)



(d)

Figure 3. Comparisons between true and predicted CPRT values for a period of 21 days (December 21, 2020 – January 10, 2021), when the true R_t variations were considered for (a) Austria, (b) Belgium, (c) Greece and (d) Portugal.

International Journal of Neural Networks and Advanced Applications
DOI: 10.46300/91016.2021.8.4 Volume 8, 2021

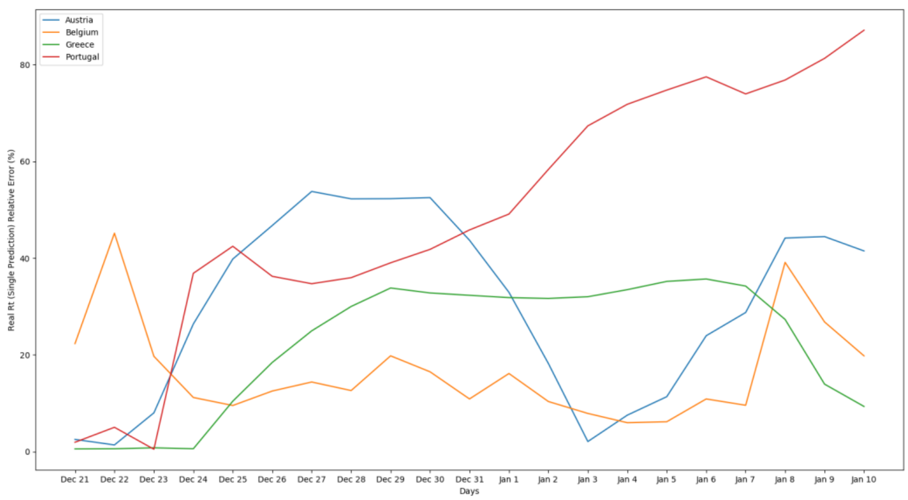
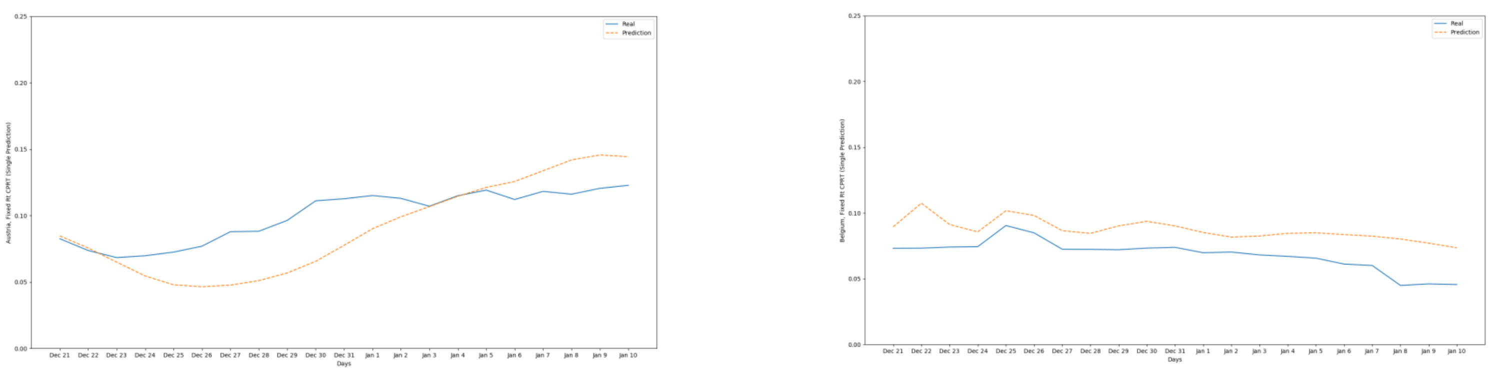
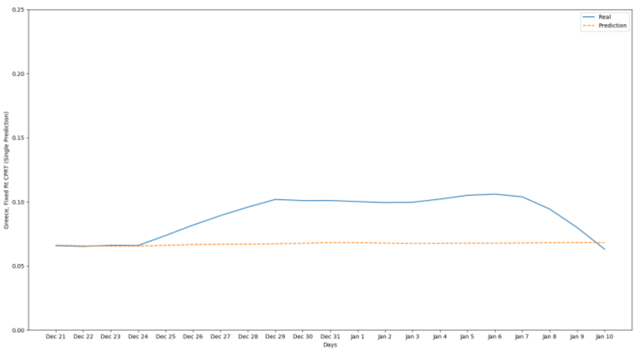


Figure 4. Relative errors between true and predicted CPRT values for a period of 21 days (December 21, 2020 – January 10, 2021), when true R_t variations were considered.

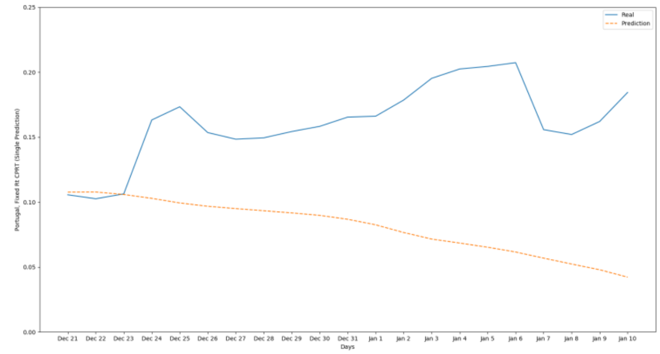


(a) (b)

International Journal of Neural Networks and Advanced Applications
DOI: 10.46300/91016.2021.8.4 Volume 8, 2021



(c)



(d)

Figure 5. Comparisons between true and predicted CPRT values for a period of 21 days (December 21, 2020 – January 10, 2021), when fixed R_t values were considered.

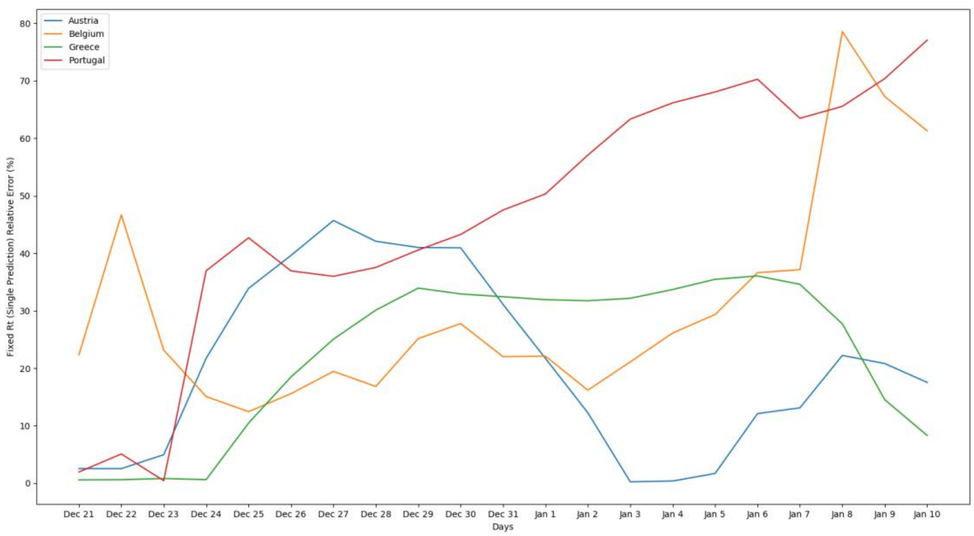
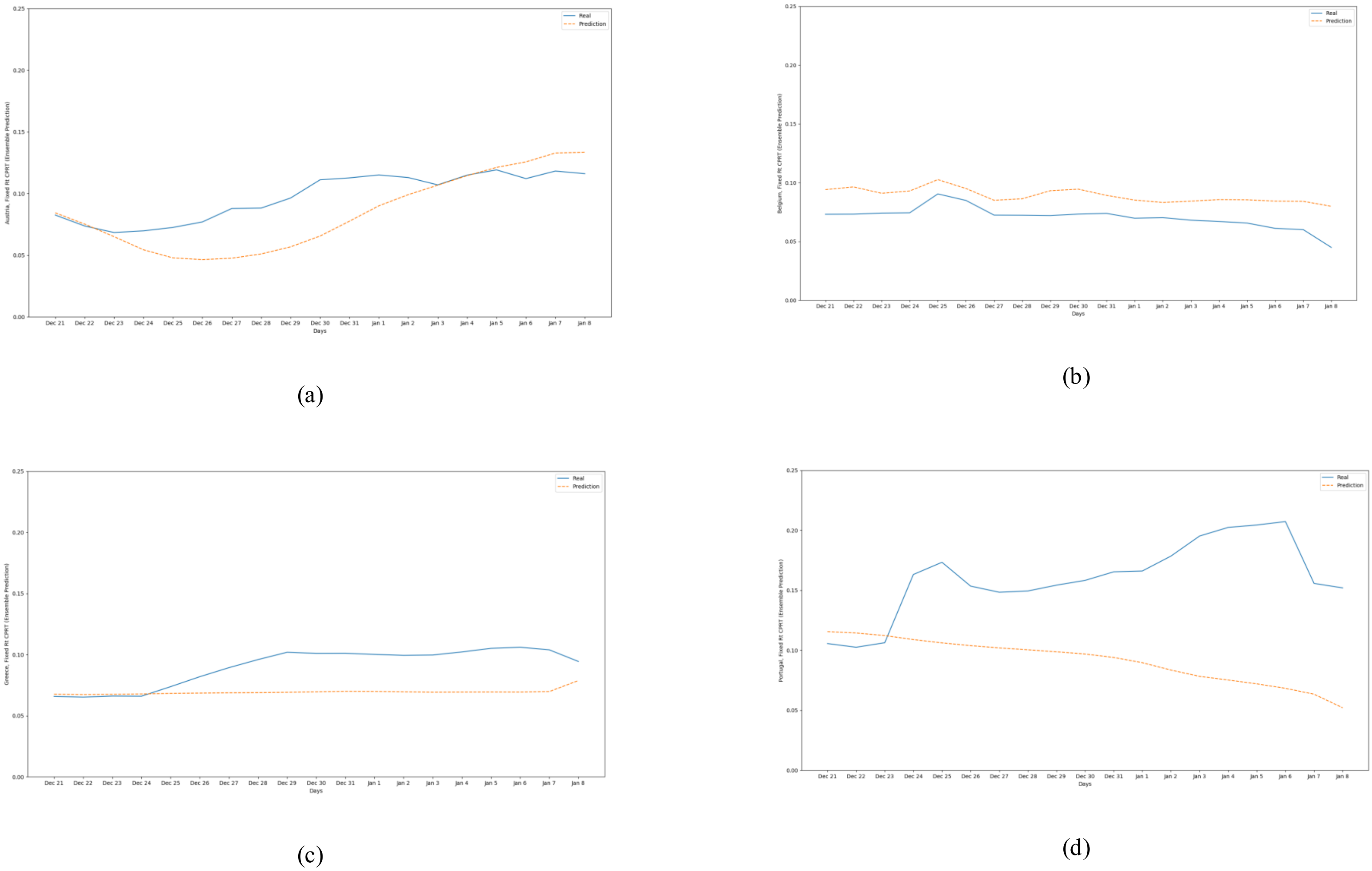


Figure 6. Relative errors between true and predicted CPRT values for a period of 21 days (December 21, 2020 – January 10, 2021), when fixed R_t values were considered.



(a) (b)

(c) (d)

Figure 7. Comparisons between true and predicted CPRT values for a period of 19 days (December 21, 2020 – January 08, 2021), with the ensemble modeling algorithm.

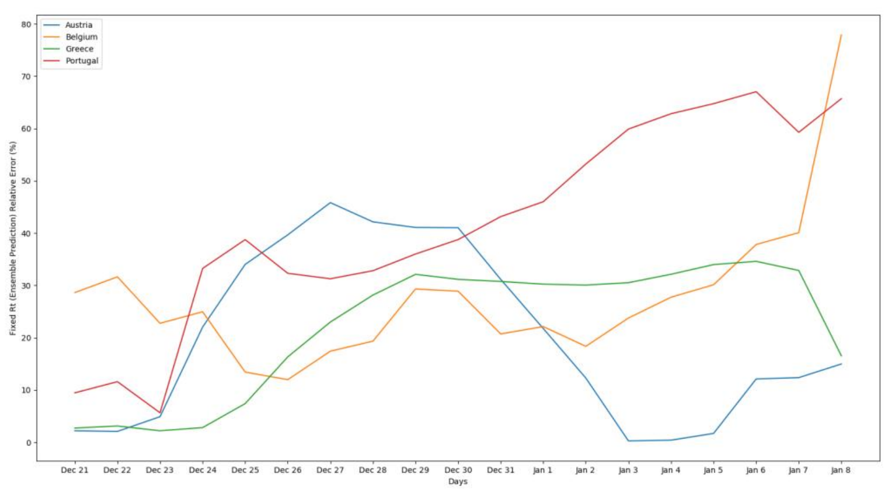


Figure 8. Relative errors between true and predicted CPRT values for a period of 19 days (December 21, 2020 – January 08, 2021), with the ensemble modeling algorithm.

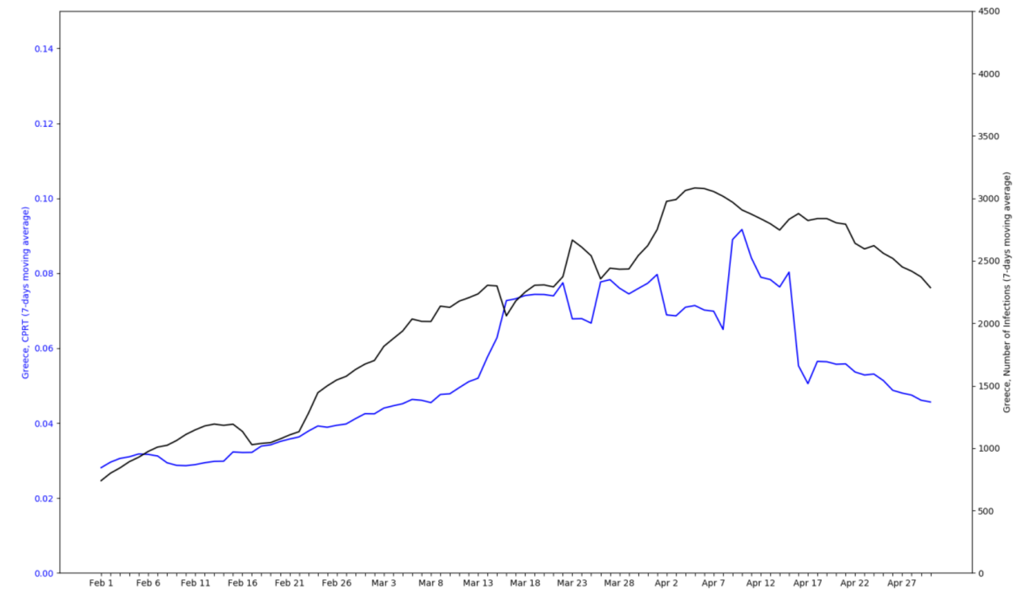
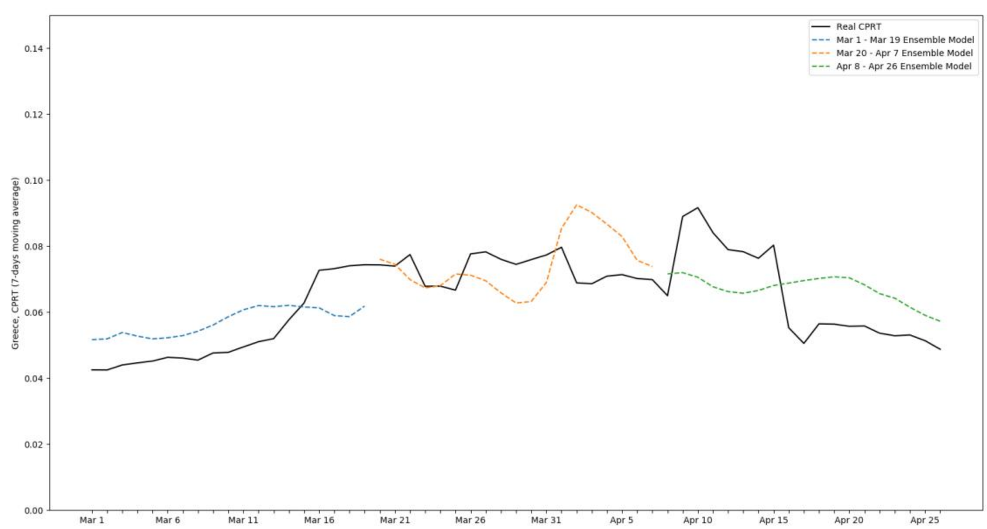
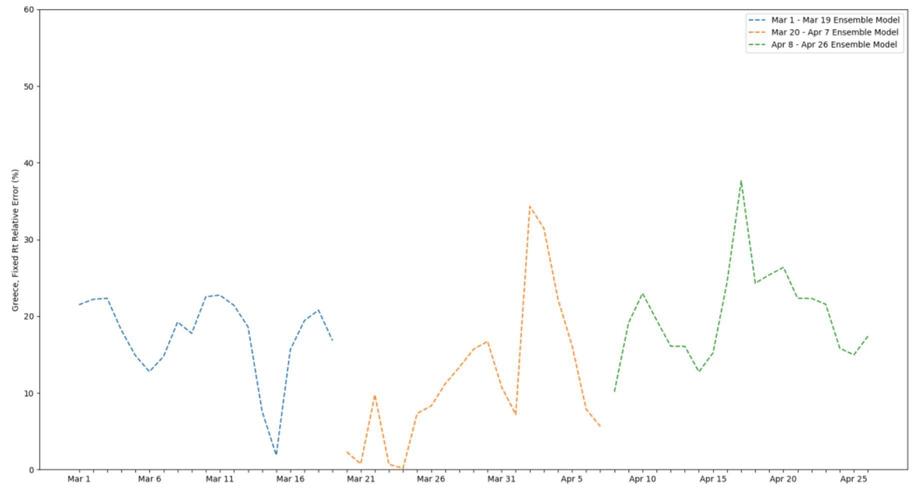


Figure 9. 7-days moving averages of CPRT values and daily reported cases for Greece between February 01 and April 30, 2021.



(a)



(b)

Figure 10. Comparisons between true and predicted CPRT values during the peak of the third pandemic wave in Greece (March -April, 2021): (a) 7-days moving average CPRT values, true (solid line) and predicted (from 3 ensemble configurations-dotted line) (b) Relative errors between true and predicted values of the 3 ensemble configurations. The 3 periods for the ensemble modeling considered are: March 1- March 19 (left dotted line), March20-April 07 (central dotted line), April 08-April 26 (right dotted line).