A Case Study on Forecasting New Daily Cases of COVID-19 at Different Scales in Brazil

Elloá B. Guedes, Catherine Martins, Paulo Ribeiro L. Júnior and Edmar C. Gurjão

Abstract— **Forecasting new COVID-19 daily cases is a task that favors actions that mitigate the damages caused by the virus. In this perspective, the present work evaluates different Machine Learning models for this purpose, considering different scales in Brazil (country, state, and city levels). After a grid search with 860 configurations, the results indicate an Artificial Neural Network well suited for national level forecasting. On the other hand, with a lower performance at the larger scale, Elastic Net was good at predicting smaller ones. This case study highlights the difficulties of model reusing for COVID-19 forecasting and also the necessity of model choice and adjustment depending on the data scale.**

Keywords— **COVID-19; Machine Learning; Forecasting, Model Selection.**

I. Introduction

COVID-19 pandemic emerged as an overwhelming health burden affecting at the time of this research 151.812.556 individuals all of the world. By May 2nd, 2021, it led to 414,399 deaths in Brazil [14] and 3.186.817 deaths all over the world [20]. The SARS-CoV-2 is a highly infectious virus with the human-to-human transmission, which can be reduced by social distancing, and sanitization protocols [19]. Worryingly, genetic variants of SARS-CoV-2 associated with more easily and quickly spreading have been emerging and circulating the world [4]. The early tracking of the disease spread in a specific location is essential to guide public health actions such as optimizing vaccination strategies, lockdown policies, testing, and others action to control the disease spread, hospitalizations, and deaths. Fighting COVID-19 pandemic is an ongoing challenge of worldwide proportions. Public health emergencies like COVID-19 require an effective response in an accurate time.

Epidemiological models have been used to manage this crisis [7], and various modified and straightforward SIR-based (Susceptible, Infectious, Recovered) models can predict the course of COVID-19 worldwide [2], [8], [11]. However, their accuracy and suitability for predicting the COVID-19 disease course is a matter of debate. Although some models were able to predict the course of disease in the countries with controlled epidemics, such as China, they failed in predicting more prominent peaks in others countries from the northern hemisphere [1]. Thus, more sophisticated modeling strategies and solutions based on computational approaches are needed to forecast the pandemic.

Effective surveillance from the local community level to the regional level is the key to find effective solutions for the containment of the virus. However, the COVID-19 surveillance system needs to be improved using Artificial Intelligence, and Information Technology [19]. In this context, forecasting techniques can assist the design of better strategies and effective decisions. These predictions might help to prepare against possible threats and consequences.

Machine Learning, a subarea of Artificial Intelligence, is an essential pillar for COVID-19 forecasting due to its accuracy [12]. The present paper aims at showing the results obtained from a case study on forecasting new daily cases of COVID-19 at different scales in Brazil: from a country level, then to a state, and lastly to a city level, to answer the following research question: *Can a given predictive Machine Learning model of new daily cases of COVID-19 have the same performance for different scales of the training data?* Answering this question is essential to establish criteria in the selection of forecasting models.

To address the scale prediction question, new COVID-19 daily cases were collected from the Brazilian Health Ministry surveillance panel, and a forecasting task was designed considering Brazil, the Paraíba state, and the city of Campina Grande. This case study scenario was chosen, taking into account practical demands for predicting COVID-19 cases in such locations to help raise population awareness of the pandemic, encouraging them to respect social distancing and sanitization policies.

This article is organized as follows: methodology is depicted in Sec. II, focusing on the description of experimental data and forecasting models; results and discussion are presented in Sec. III, comparing different models under the different scales considered. Lastly, the conclusions of the work are presented in Sec. IV.

II. Methodology

Aiming at evaluating ML models on forecasting COVID-19 cases at different scales, we model the problem as a regression task under the Supervised Learning paradigm. To do so, we considered as target variable the 7-days ahead cumulative new cases with an average of three days, rounded to the closest larger integer. Since long-term predictions are challenging for most models [13] and the event upon prediction is new,

Elloá B. Guedes, Research Group on Intelligent Systems, Amazonas State University, Manaus-AM, ebgcosta@uea.edu.br; Catherine Martins, University Hospital Alcides Carneiro, catherinesfmartins@gmail.com; Paulo Ribeiro Lins Júnior, Research Group on Communications and Information Processing, IFPB – *Campus* Campina Grande, Campina Grande-PB, paulo.lins@ifpb.edu.br; Edmar C. Gurjão, Campina Grande Federal University, Campina Grande-PB, ecg@dee.ufcg.edu.br. Elloá B. Guedes thanks FAPEAM and CNPq support (Grant PPP 004/2017). All authors acknowledge FAPESQ financial support (Grant 003/2020 - SEECT/FAPESQ/PB).

ongoing, and has complex dynamics, such a short-term horizon may result in more realistic reliable models.

A. Experimental Data: Collection, Cleaning and Preparation

The experimental data for this case study was collected on Brazilian's Health Ministry COVID-19 surveillance guide on April 24th, 2021 [14]. It comprises both new and cumulative confirmed cases and deaths from all federation units at different scales: from the country as a whole, to state, and to city-wise. Given the continental dimensions of Brasil, due to the complexity and dynamics of gathering such data, there is a disclaimer on some fragilities: daily data may be reviewed to minimize inconsistencies, and new events may be registered with delays.

From the dependent variable in a country level, there were $14.122.795$ confirmed cases starting from February 25th, 2020 up to the date in which the obtained dataset. The state under consideration, Paraíba, had 283.147 cases confirmed so far. At a city level, which takes Campina Grande in Paraíba, there were 25.955 confirmed cases. Graphics in Fig. 1 show the evolution of cumulative daily cases in each place, Fig. 2 shows the new cases daily variation, where the colors used have the same semantics, and Fig. 3 summarizes such variation on new daily cases along the months. For the objective of this paper, data of other states and cities as well as on deaths was discarded.

Transformations in the time series of the observations as a time series make it adequate to an automated ML forecasting scenario: the dates were replaced by integer indexes referring to the day of occurrence since the initial observation; for each day, say the t-th day, the seven days ahead ceil of the three-day mean average cumulative confirmed cases was obtained, i.e.: $y_t = \lceil (X_{t+7} + X_{t+8} + X_{t+9})/3 \rceil$, where X_t denotes the number of cumulative confirmed cases on the day t. As it can be seen in Figs. 2 and 3, there is a significant variation in the number of daily, and it introduces a challenge for forecasting. By defining a target variable y_t as such, it is possible to amortize such variation and model the event in monotonic growth dynamics.

As a result of the described data preparation, each granularity in this case study contains 424 examples with X_t, X_{t-1}, X_{t-2} as independent variables and y_t as the dependent variable.

B. Models, Parameters and Hyperparameters

In this work, we addressed the proposed regression task using the following learning models and frameworks. They were chosen based on the theoretical background on function approximation, on the recent results on practical problems, and on delivering good results with few training data.

1) **Elastic Net**. Based on a linear approach to map the relationship between the target variable and the explanatory variables, this model also considers a linear combination of L_1 and L_2 penalties on the sum of squared errors. It combines effective regularization with feature selection [10].
2) **Multilayer Perceptron Artificial Neural Networks (ANN)**. Consist of massive parallel distributed models made up of fully connected layers of simple processing units that have a natural propensity for storing experiential knowledge and making it available for use [9]. According to the Universal Approximation Theorem, given samples of any function, there is a neural network that can infer an approximate implementation of it;
3) **XGBoost Regressor (XGBoost)**. A scalable machine learning system for tree boosting, it has been widely recognized in a number of machine learning and data mining challenges. It considers a novel tree learning algorithm for handling sparse data, a theoretically justified weighted quantile sketch procedure that enables handling instance weights in approximate tree learning, and a parallel and distributed computing implementation that speedups learning [5];
4) **Gaussian Process Regression (GPR)**. It is a powerful, non-parametric Bayesian approach towards regression problems that can be utilized in exploration and exploitation scenarios. It can capture a wide variety of relations between inputs and outputs by utilizing a theoretically infinite number of parameters and letting the data determine the level of complexity through the means of Bayesian inference [18]. Theoretical and practical developments over the last decade have made Gaussian processes a serious competitor for real supervised learning applications [17].

A grid search on parameters and hyperparameters of the models considered was applied, aiming at optimization. For the Elastic Net, we used 15 equally spaced values from 0.05 to 0.95 for α and L_1-ratio where the first is a constant that multiplies the penalty terms and the second is a mixing parameter of the two penalties considered, number of iterations up to 20.000 and random or cyclic strategies for the coefficient update. In the case of ANNs, we applied geometric pyramid and Baum & Haussler rules for estimating neurons in the hidden layers [15], ReLU (Rectified Linear Unit), sigmoid and hyperbolic tangent activation functions and at most 5000 epochs subject to the early stopping of patience equal to 50 to avoid overfitting. In the case of XGBoost regressor, we tried gbtree, gblinear and dart boosters, and 100, 500, 1.000, and 5.000 possible boosting rounds. On GPR, multiple kernel functions such as Constant, Matérn, Dot Product, RBF, and Periodic Kernel with different hyperparameters were tested. Moreover, we also tested some randomly chosen combinations of these kernel functions over sum and product.

C. Performance Metrics and Validation

The R^2-Score is the performance metric under consideration for this regression task. It denotes the proportion of the variance in the dependent variable that is predictable from the independent variable. The best possible score is 1.0, and it can be a negative value for arbitrarily worse models [3].

Different configurations for each scenario were validated according to a time series split cross-validation with 5 splits and 10 runs. In this approach, a variation of k-fold cross-validation for time series problems, data samples are observed

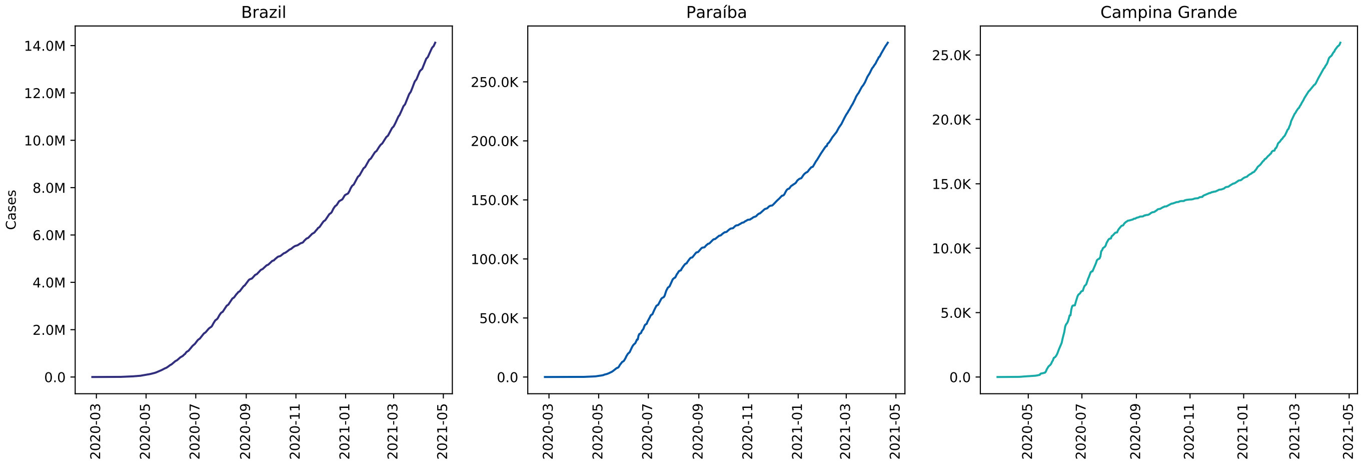

Fig. 1. Cumulative COVID-19 cases.

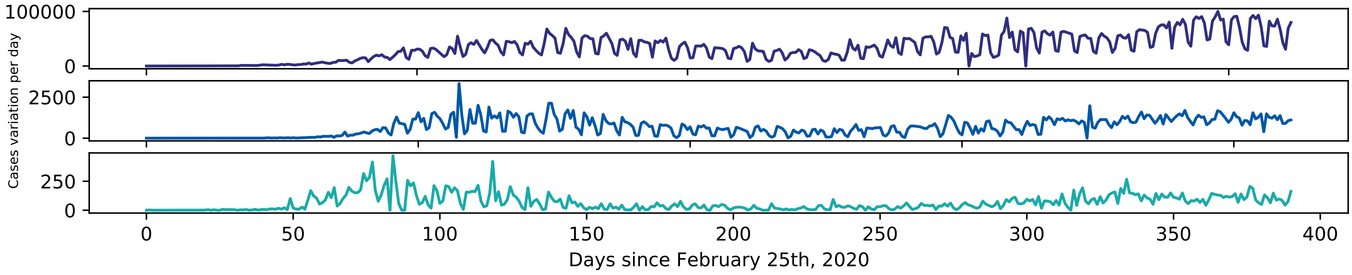

Fig. 2. Daily variation of new cases.

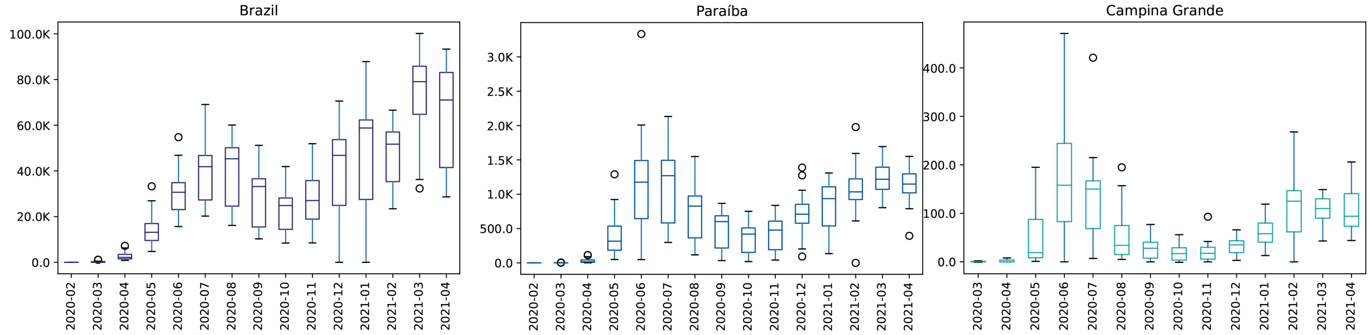

Fig. 3. Variation of new daily COVID-19 cases per month.

at fixed time intervals and, in each split, test indices must be higher than before [13]. Thus performance metrics for each configuration are in terms of average and standard deviation from each split and run.

III. Results and Discussion

The steps described in the previous section were implemented with the Python programming language platform using Numpy, Pandas, Scikit Learn, and Matplotlib frameworks. Results obtained can be found in Tables I-II.

At the first level, 860 different configurations for each model were trained and tested. Taking a closer look at the confidence intervals for R^2 of best configurations per model type in a country level, as shown Fig. 4, it is possible to notice that the ANN was the best regressor for the task. By depicting its performance as more data becomes available along with the splits, vide Fig. 5, it was possible to notice that this model was also the most stable amongst those observed in the same scenario.

The predictions of the best ANN are in Fig. 6 in contrast with observed values (ground truth), as in the last time series split in the cross-validation proposed where 80% of data is for training and the rest for testing. It can be seen that it performs well in all scales and is most likely due to this model's capabilities for capturing non-linear relationships [9].

After examining the generalization capabilities to the state and city under consideration, the performance was diffuse for all models subject to all scales. The best Elastic Net identified

TABLE I
CONFIGURATIONS EVALUATED AND RESULTS FOR BEST CONFIGURATION OF EACH MODEL.

		Best Configuration			
Model	**#**	**$\mathbf{R^2}$ at Brazil**	**$\mathbf{R^2}$ at Paraíba**	**$\mathbf{R^2}$ at Campina Grande**	**Parameters**
Elastic Net	450	0.9262 ± 0.0485	0.5910 ± 0.4202	0.1872 ± 0.8069	$\alpha = 0.2428, L_1 = 0.8857$, cyclic selection
XGBoost	12	0.6330 ± 0.2037	-0.0936 ± 0.5724	-0.5752 ± 0.7043	gblinear booster, 5.000 estimators
ANNs	342	0.9860 ± 0.0118	0.5143 ± 0.7655	-204.7508 ± 409.5896	two hidden layers with $(33, 76)$ neurons and ReLU activation function
GPR	56	0.8010 ± 0.1321	0.1235 ± 0.7587	-0.0238 ± 0.8929	Dot product kernel with $\sigma_0 = 4$

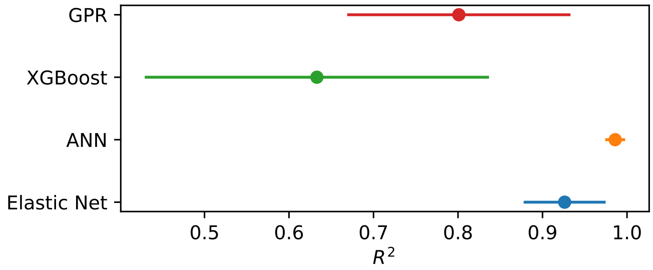

Fig. 4. Confidence Intervals of Best Model for Brazil per Split.

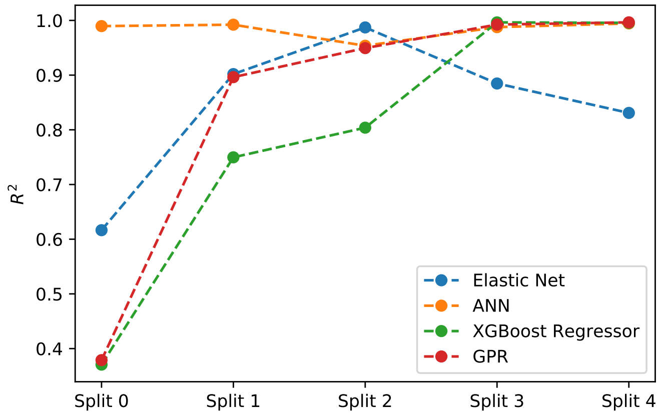

Fig. 5. Results of Best Model for Brazil per Split.

TABLE II
RESULTS OF R^2 AT EACH TIME SERIES CROSS-VALIDATION SPLIT.

Model	**Scale**	**Split 0**	**Split 1**	**Split 2**	**Split 3**	**Split 4**
Elastic Net	**BR**	0,9181	0,8537	0,9807	0,9789	0,8996
	PB	0,4238	−0,1615	0,8340	0,9437	0,9151
	CG	0,8686	−1,0168	−0,5480	0,7461	0,8861
XG Boost	**BR**	0,9273	0,5437	0,7762	0,3329	0,5847
	PB	−0,6470	−0,8515	0,5680	−0,0017	0,4643
	CG	0,7372	−1,2304	−0,9215	−0,4475	−1,0135
ANNs	**BR**	0,9893	0,9933	0,9879	0,9965	0,9631
	PB	−1,0015	0,6906	0,9811	0,9839	0,9173
	CG	−1023,9293	−0,5427	−0,5071	0,9674	0,2577
GPR	**BR**	0,9090	0,5723	0,8702	0,7344	0,9192
	PB	−0,6360	−0,9501	0,6637	0,6676	0,8722
	CG	0,7369	−1,5456	−0,5433	0,6461	0,5869

at the country level, for instance, has a R^2 score -6.45% worse than the best ANN in the same scenario. On the other hand, this Elastic Net had significantly better performance in the state ($+12.97\%$) and especially in city levels ($+100.09\%$) in comparison to this ANN. A hypothesis for these contrasting findings may be related to the dynamics in these small scales scenarios, as shown previously in Fig. 1, that seem not to follow the same distribution as the larger one albeit being part of it.

An important concern for COVID-19 forecasting models is their capability to provide accurate predictions at the beginning of an epidemic outbreak when there is no data at all and then limited as time passes, making predictions widely uncertain [16]. The best ANN, for instance, was not capable of providing good results for a city level with few data, as observed in the first split in Table II. On the other hand, GPR delivered an excellent prediction in this scenario, and its performance grew as more data became available. Given GPR's low computational cost, it has the potential of delivering good forecasts on an ongoing pandemic scenario.

IV. CONCLUSIONS

In this work, we presented the results of a case study on forecasting one week new daily cases of COVID-19 with different Machine Learning models by assessing their capability to generalize the results for smaller scales (country level to city level). After carrying out an extensive grid search with a time series cross-validation approach, it was possible to identify two hidden layers ANN with excellent results on a country level but whose performance is not stable under the small levels considered. For state and city levels, it was not possible to find stable performance across the models investigated.

The results obtained suggested that finding a good model for forecasting new COVID-19 cases that can be reused in other scenarios remains a non-trivial task, still demanding speciallized human effort, mathematical, computational and statistical tools to assess and deliver good predictions.

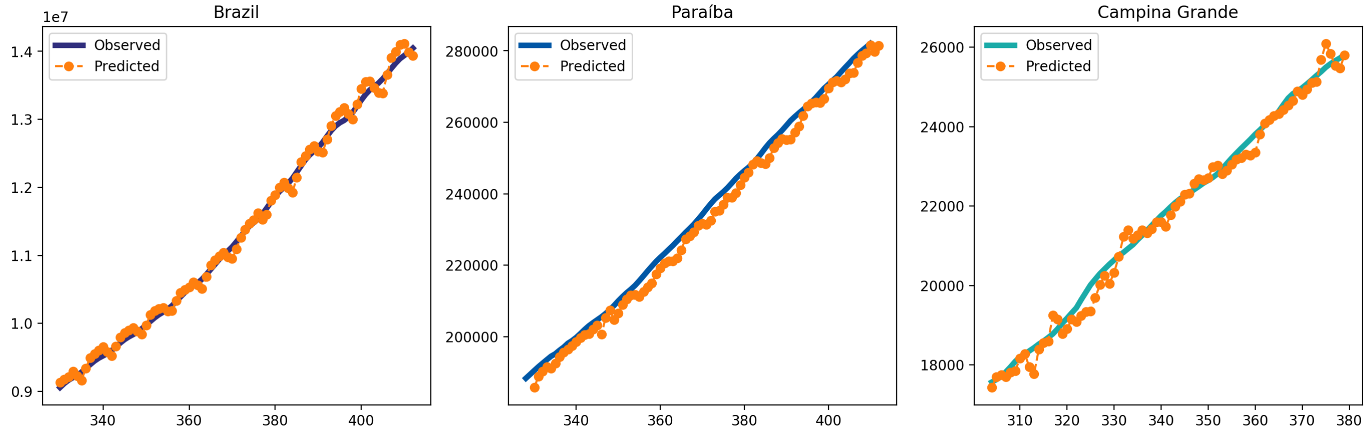

Fig. 6. Predictions of the best ANN with 80% of data for training.

The rapid spread of COVID-19 in Brazil is attributable to many factors, including urban density, the timing of the implementation and maintenance of social distancing policies, and limited testing capacity [6]. Considering this context, a result of this work is an ANN architecture that accurately predicts new COVID-19 cases for Brazil based on past data, using an autoregressive approach that not demands other endogenous or exogenous variables. Such model can be further evaluated to support strategic decisions regarding Public Health, helping authorities and managers to (i) create, adopt, revise and sustain social distancing policies (schools reopening, temporarily closing stores and reducing hours, for example); (ii) rationalize COVID-19 rapid and RT-PCR tests, especially when there is outbreak potential; (iii) improve hospitals' logistics, by anticipating the demand for beds, medicine, equipment, etc.; (iv) plan and implement vaccination campaigns, especially focusing on a vulnerable population, among others. In general, promoting prevention strategies is economically less expensive than implementing therapeutic interventions.

In future work, we aim to investigate the problem of forecasting autoregressive time series phenomenon with few available data, assessing how Machine Learning techniques can help predict a near future with high confidence on predictions. This is particularly useful when there is an outbreak of a new disease.

References

[1] Semra Ahmetolan, Ayse Humeyra Bilge, Ali Demirci, Ayse Peker-Dobie, and Onder Ergonul. What can we estimate from fatality and infectious case data using the susceptible-infected-removed (SIR) model? a case study of covid-19 pandemic. *Frontiers in Medicine*, 7, 2020.

[2] Cleo Anastassopoulou, Lucia Russo, Athanasios Tsakris, and Constantinos Siettos. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PLOS ONE*, 15(3):e0230405, 2020.

[3] A. Colin Cameron and Frank A.G. Windmeijer. An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77(2):329–342, 1997.

[4] CDC. Sars-cov-2 variant classifications and definitions. Available at `https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html`, 2021. Accessed August 12, 2021.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[6] Mayra Monteiro de Oliveira, Trevon L. Fuller, Patricia Brasil, Claudia R. Gabaglia, and Karin Nielsen-Saines. Controlling the covid-19 pandemic in brazil: a challenge of continental proportions. *Nature Medicine*, 26:1505–1506, 2020.

[7] Thiago de Paula Oliveira and Rafael de Andrade Moral. Global short-term forecasting of COVID-19 cases. *Scientific Reports*, 11(1), 2021.

[8] Giulia Giordano, Franco Blanchini, Raffaele Bruno, Patrizio Colaneri, Alessandro Di Filippo, Angela Di Matteo, and Marta Colaneri. Modelling the COVID-19 epidemic and implementation of population-wide interventions in italy. *Nature Medicine*, 26(6):855–860, 2020.

[9] Simon Haykin. *Neural Networks and Learning Machines*. Pearson Prentice-Hall, Estados Unidos, 3 edition, 2009.

[10] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, New York, 1 edition, 2013.

[11] Qianying Lin, Shi Zhao, Daozhou Gao, Yijun Lou, Shu Yang, Salihu S. Musa, Maggie H. Wang, Yongli Cai, Weiming Wang, Lin Yang, and Daihai He. A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in wuhan, china with individual reaction and governmental action. *International Journal of Infectious Diseases*, 93:211–216, 2020.

[12] Parikshit N Mahalle, Nilesh P Sable, Namita P Mahalle, and Gitanjali R Shinde. Data analytics: Covid-19 prediction using multimodal data. In *Intelligent Systems and Methods to Combat Covid-19*, pages 1–10. Springer, 2020.

[13] Aileen Nielsen. *Practical Time Series Analysis – Prediction with Statistics & Machine Learning*. O'Reilly, Canada, 1 edition, 2020.

[14] Ministry of Health of Brazil. Health surveillance secretariat: Covid-19 epidemiological surveillance guide, 2020. Available at `https://covid.saude.gov.br/`. Accessed on August 12, 2021.

[15] Ajoy K. Palit and Dobrivoje Popovic. *Computational Inteligence in Time Series Forecasting - Theory and Engineering Applications*. Springer, Londres, 1 edition, 2005.

[16] Fotios Petropoulos and Spyros Makridakis. Forecasting the novel coronavirus covid-19. *PLoS ONE*, 15(3):1–8, 2020.

[17] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Massachusetts Institute of Technology Press, United States, 1 edition, 2006.

[18] Eric Schulz, Maarten Speekenbrink, and Andreas Krause. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1 – 16, 2018.

[19] Nishant Srivastava, Preeti Baxi, R. K. Ratho, and Shailendra K. Saxena. Global trends in epidemiology of coronavirus disease 2019 (COVID-19). In *Medical Virology: From Pathogenesis to Disease Control*, pages 9–21. Springer Singapore, 2020.

[20] WHO. COVID-19 weekly epidemiological update. Technical Report 4 May 2021, WHO, 2021.