Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 181 (2021) 940–947

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies

# A COVID-19 time series forecasting model based on MLP ANN

Pedro Henrique Borghi[a,b], Oleksandr Zakordonets[a], João Paulo Teixeira[a] *

[a]*Research Centre in Digitalization and Intelligent Robotics (CEDRI), Instituto Politecnico de Bragança, Braganca, Portugal*
[b]*Federal University of Technology - Paraná (UTFPR), Cornélio Procópio, 86300-000, Brazil*

**Abstract**

With the accelerated spread of COVID-19 worldwide and its potentially fatal effects on human health, the development of a tool that effectively describes and predicts the number of infected cases and deaths over time becomes relevant. This makes it possible for administrative sectors and the population itself to become aware and act more precisely. In this work, a machine learning model based on the multilayer Perceptron artificial neural network structure was used, which effectively predicts the behavior of the series mentioned in up to six days. The model, which is trained with data from 30 countries together in a 20-day context, is assessed using global and local MSE and MAE measures. For the construction of training and test sets, four time series (number of: accumulated infected cases, new cases, accumulated deaths and new deaths) from each country are used, which are started on the day of the first confirmed infection case. In order to soften the sudden transitions between samples, a moving average filter with a window size 3 and a normalization by maximum value were used. It is intended to make the model's predictions available online, collaborating with the fight against the pandemic.



*Keywords:* COVID-19 Brazil forecast; COVID-19 Italy forecast; COVID-19 worldwide forecast.

* Corresponding author. Tel.: +351 913 835 373; Tel.: +55 14 99716 6671.
*E-mail address: pedromelo@alunos.utfpr.edu.br*


10.1016/j.procs.2021.01.250

## 1. Introduction

In early December 2019, in the city of Wuhan - Hubei, China, a new coronavirus emerged that caused potentially fatal respiratory complications related to severe acute respiratory syndrome (SARS) [1]. The disease caused by the infection with this new coronavirus was named COVID-19. The virus, although new, is part of an already known group, some of which are pathogenic to man [2]. Over the following months, it spread to all nations of the world, being classified by the WHO at 11 March 2020 as pandemic level [3]. Since it has a high rate of contagion and lethality and a vaccine or widely effective treatment is not known yet, it has required measures of social isolation, strict health control actions and partial or complete suspension of daily activities in cities [4]. With the rapid advance of COVID-19, several health systems worldwide have collapsed, leading to serious health crises and causing thousands of deaths [5]. Although containment measures and health system reinforcements have already been recommended by scientific and medical organizations across the globe, many countries delay or do not apply them, making the population highly susceptible and contributing to an accelerated spread of the disease.

Since its inception, constant surveys have been carried out around the world on the number of new people infected and dead by the disease, as well as the number of patients recovered. Systematic descriptions of these data are made available daily on various platforms online [6-12]. Combining the expectations of health agencies, government action measures and the monitoring of updated data, it is noted that the behavior of the time series produced by the number of accumulated cases and the number of accumulated deaths resembles an exponential curve with a subsequent flattening. Observing and predicting future data from these series becomes a relevant activity due to the global scale of the problem, since control and relaxation measures can be decided with more accurate predictions of the evolution of the situation.

Based on this problem and on the potential pointed out, this work presents the development and comparison of two methods to effectively predict the number of new cases and new deaths due to COVID-19. The work seeks to assist in fighting the virus by forecasting data within up to 6 days, providing additional tools that assist in decision making by government officials and health systems and in informing the population. The model, created to deal with country data, can later be adapted for states and cities according to their realities. As they are active series that are constantly being updated, the models have the potential to be similarly updated, since they obtain data automatically in online platforms and are trained in a few seconds. It is expected that the number of forward forecast days can be augmented with the time because longer time series will be available. The implemented models are based on multilayer Perceptron (MLP) artificial neural networks (ANN) [13].

## 2. Theoretical Framework

This section presents the theoretical introductions for the machine learning approach and the moving average filter used to smooth the time series.

### *2.1. Multilayer Perceptron - Artificial Neural Network*

The machine learning method by artificial neural network comes from the basic structure called artificial neuron. This, which is inspired by the behavior of the biological neuron, can be represented by several different models, featuring different typologies. By associating several neurons, neural networks are formed that, similarly to biological nervous systems, perform functions of recognition, control, decisions and others. The functionality of the networks is conditioned according to the training to which they are submitted. This training, supervised or unsupervised, is based on the presentation of examples, and simulates a systematic learning process by determining the difference between the response given by the network and the expected behavior. The experience of the network is stored by the synaptic weights between neurons and its performance is evaluated, for example, by the ability to generalize behaviors, recognize patterns, fix errors or execute predictions [13-15].

For time series forecasts, one of the most popular topologies is the multilayer Perceptron, a feedforward type architecture that is based on the Perceptron neuron model. From an arrangement of successive interconnected neural layers, the information spreads from the input to the output so that on the way it is abstracted by neurons.

The process of determining the most suitable parameters (number of layers, number of neurons in each layer and

activation functions) for networks depends mainly on the type of problem and its complexity [9] [16]. For the forecasting of time series with low randomness, as in the case of the series trends related to the number of infections and deaths due to COVID-19, few resources are needed. Thus, it is expected that with up to two hidden layers and a maximum number of ten neurons in each one will be enough for the MLP model to extrapolate both series simultaneously.

Regarding the number of days to be forecasted, it is known that the longer the period, more context information must be provided to the network. This contextualization can come in the form of an increase in the input data window or in the association of other features that describe the same period. Also, discontinuities and sudden changes in the sequential behaviour of the input must be avoided or reduced, in order to facilitate the interpretation by the network.

### *2.2. Moving Average Filter*

One way to reduce sudden transitions and discontinuities in time series is by applying a moving average filter [17]. Basically, the effect of this processing is to smooth the transitions of consecutive samples by calculating the average between each sample and its neighbors. The increase in the window for averaging implies, in most cases, an increase in the smoothing of the series, since it takes into account a larger number of samples. A wrong choice of the window size can lead to a strong depreciation of the series' behavior, de-characterizing it. Small windows, in the case of the series used in this work, are the most suitable, since they reduce sudden transitions and maintain short-term behavior.

Mathematically, the moving average filter can be described by Equation 1, where $M$ is the series after applying the filter, $N$ is the window size, $x$ is the original series and $w$ is the window function, which weighs the samples of $x$ according to its shape [17]. In the case of w being a pulse train (rectangular window function), Equation 1 becomes an arithmetic mean. For this work, this was the chosen window.

$$M(n) = \frac{1}{N+1} \sum_{m=n-\left(\frac{N}{2}\right)}^{n+\left(\frac{N}{2}\right)} |x(m)|.w(N+1) \tag{1}$$

## 3. Materials

The database used was extracted via the Application Programming Interface (API) available publicly at [7]. From a sequence of requests executed on the programming platform, it is possible to obtain a data package about COVID-19 updated daily for 185 countries. These packages contain time series of data referring to the cumulative number of confirmed infected cases, cumulative number of deaths, cumulative number of recovered patients and cumulative number of active cases. The beginning of the series obtained can be conditioned via a specific request in the API or from the day with the notification of the first confirmed case in the country (standard chosen in this work). In addition, the series have a one-day step between consecutive samples.

Data is sourced from Johns Hopkins University Center for Systems Science and Engineering [18] that is available at [10]. Thus, this paper aggregates the data from the international and national specific data sources, some of which [3, 8, 11, 12].

The proposed system algorithm was implemented and executed in MatLab® 2020a software, using the Deep Learning Toolbox.

## 4. Methodology

Initially, the data present in the database [7, 10] were downloaded and stored as a file. Then, the 30 countries with the largest number of confirmed cases of COVID-19 were selected until 11 may of 2020. The following time series were generated: number of confirmed infected cases accumulated, number of new infected cases, number of accumulated deaths and number of new deaths.

Thus, the four time series have the same number of days for the same country and, eventually, different number of days for different countries.

Then the moving average filter with a window of size equal to three was applied to all series in order to reduce the effect of possible acute increases, and decreases (in the case of series of new cases and new deaths), between consecutive samples. This process implies smoothing the transitions between samples, improving the learning process of the neural network during training. This smoothing of the time series over 3 days attenuates the instability of the daily reported number of cases that several times are corrected by the numbers of next day.

Then, the series amplitude was normalized, all of them being normalized from their maximum value. However, the maximum values of the series of accumulated confirmed infected cases and accumulated deaths were stored as a reference and used in the post-processing stage for the presentation of full-scale forecasts.

Five data sets were then generated using the series from all 30 countries. Three of the five sets, typified as datasets and named: train, test and seer, had the following features: a window of 20 samples of accumulated confirmed infected cases, a window of 20 samples of new infected cases, the number of accumulated deaths for the twentieth sample and a window of 20 samples of new deaths, (in a total of 80 input nodes). Each set of features mentioned is related to the same period of 20 days in a single country. The slide between consecutive segments was defined as one day.

The other two sets, typified as output forecast sets and named: train and test, had the following features: a window of 6 samples of accumulated confirmed infected cases and a window of 6 samples of accumulated deaths. Each set of features mentioned refers to the period of 6 days after the same 20-day window of the respective train and test data sets. With that, it is expected that the neural network learns to predict 6 days later from 20 days of contextualization. In addition, data from all countries are used in all sets so that they may contain the specificities of each country, if any, and thus the forecast can be performed despite these. Therefore, the network was trained under all countries' data and thus can be applied specific to one or generally. When well-tuned, this approach allows a robust application.

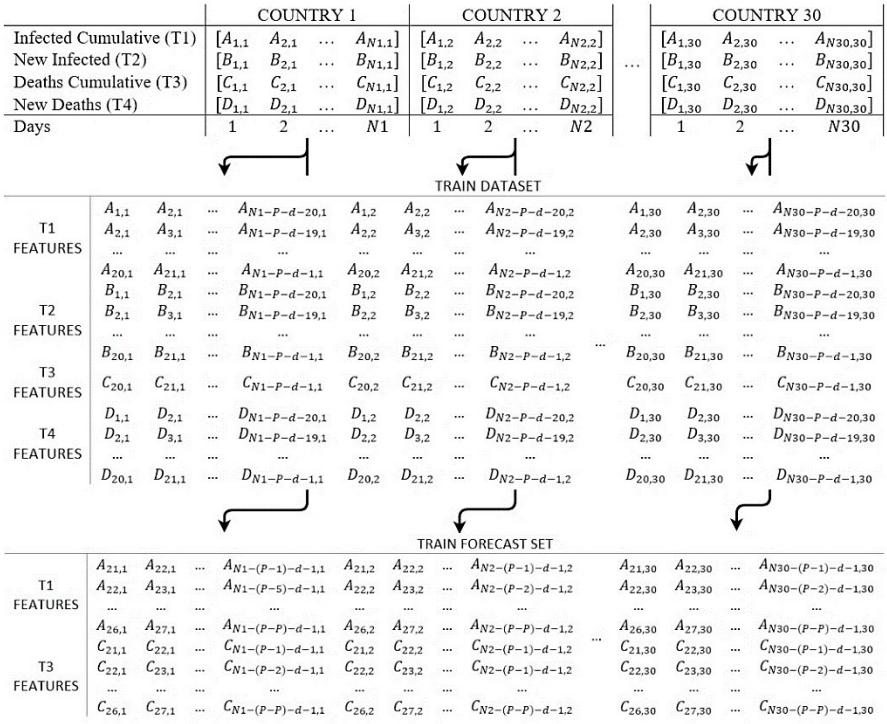


Fig. 1. Generation of train dataset and forecast set based on 20-day context from 30 countries.

Fig. 1 shows the train dataset and forecast set organization, generated from segmentation of original dataset. Test and seer sets was similarly built, being related to posterior periods of train sets. At Fig. 1, $N_n$ is the number days related to country $n$ period, $P$ is the number of days to forecast, in this case equal to 6, and $d$ is the length of test set in days to each country.

The naming between the sets differs by the number of days each contains. For example, the seer set has only the data from the last 25 days of acquisition, that is, the day may eleven (11/05) and its 24 predecessors. Thus, all forecasts

made by the network contain at least one day without the possibility of validation. The test set proceeds from the seer set up to three days. Thus, forecasts can be validated and correspond, for short series of up to 40 days, 15% of the total period. From the lower limit of the test set, all data belongs to the train set. Thus, it is guaranteed that most data will be used in training, helping to reduce the effect of data scarcity.

The stage of construction of the learning model occurred based on a typical structure of a MLP network. In this work, the determination of the most suitable parameters for the application respected the following methodology, where for each variation, new training processes, testing and measurement of the respective errors were carried out.

- Variation in the number of hidden layers from one to two;
- Variation in the number of neurons in the first hidden layer from one to ten;
- Variation in the number of neurons in the second hidden layer, if any, from one to ten;
- Activation function of neurons in the hidden layers as being tangent sigmoid;
- Activation function of neurons in the output layer as being linear;
- Backpropagation training algorithm using the Levenberg-Marquardt method.

In the process of training the networks, the segments of the train dataset were randomly divided between training, validation and testing under a respective proportion of 70%, 15% and 15%. After training, the network was simulated on the test dataset and the predictions obtained were used to calculate the error, respectively to the expected values contained in the test forecast set. From the error matrix obtained, measures of mean square error (MSE) and mean absolute error (MAE) were determined for each output neuron and globally. The results of forecasts and error measures were stored for further analysis of the data. In addition, the forecasts went through a post-processing that readjusted the scale of confirmed cases and deaths according to previously stored references for normalization.

## 5. Results

In the stage of determining the most suitable parameters for the MLP network learning method, after the execution of all the structural variations presented in the methodology, the results of the MSE and MAE measurements local (for each output neuron) and global (mean between all local measurement) were organized upwards from global minimums. From this organization, the six best models (combination of parameters) were selected, whose performances are presented in Tables 1 and 2. The fifth and sixth columns represent the respective errors calculated only on the neurons that predict infected cases and deaths on the sixth day ahead from the most recent entry, that is, theoretically the most difficult forecast to be realized. That predictions are made on sixth and twelfth neurons, respectively. The output data were restored to non-normalized levels according the reference stored on normalization process before the MSE e MAE measures, so that is possible to notice the robustness of models since there are different absolute values of infected and death cases to each country.

From Tables 1 and 2, it is noticed that the error measures for the forecasts of the sixth day ahead are always higher than compared to the respective global measure, which reinforces the hypothesis of the network's difficulty for forecasts more distant from the input data. However, the aggregation of several characteristics about the series, as it was done, improves learning and facilitates the generalization of the input behavior, and it is said that the models tend to represent reality satisfactorily, see Figures 2, 3, 4 and 5.

Regarding the number of layers and the number of neurons in each hidden layer, it is noticed that there is no tendency for improvement as the number of neurons in the layers increases since there are models with few neurons (less than half of the high limit) among the best. However, it noted the prevalence of models with two hidden layers among the best. Anyhow, it is demonstrated that it is possible to achieve good results with just one hidden layer (M2 and M9).

Due to M1 and M2 are within the six best results in both tables and together gathering all the minimum measures, they are defined as being the most suitable for the application. Among both, model M1 is chosen as the best since it gathers 4 from 6 minimum measures. Figures 2, 3, 4 and 5 are generated based only on it and using Brazil’s and Italy’s data.

Table 1. Best models based on MSE measures at the stage of choosing parameters for the MLP network; Errors calculated on the network's output for test dataset (all countries used).

| Model | Number of neurons in the first hidden layer | Number of neurons in the second hidden layer | Global MSE $[10^6]$ | MSE_6 Infected $[10^6]$ | MSE_6 Deaths $[10^3]$ |
|---|---|---|---|---|---|
| M1 | 9 | 4 | **10,30** | **43,35** | **167,5** |
| M2 | 8 | 0 | 11,67 | 45,18 | 783,5 |
| M3 | 8 | 10 | 12,80 | 61,49 | 473,2 |
| M4 | 3 | 1 | 13,94 | 66,48 | 617,7 |
| M5 | 8 | 1 | 14,46 | 62,81 | 663,2 |
| M6 | 1 | 1 | 14,55 | 62,25 | 733,4 |

Table 2. Best models based on MAE measures at the stage of choosing parameters for the MLP network; Errors calculated on the network's output for test dataset (all countries used).

| Model | Number of neurons in the first hidden layer | Number of neurons in the second hidden layer | Global MAE $[10^3]$ | MAE_6 Cases $[10^3]$ | MAE_6 Deaths |
|---|---|---|---|---|---|
| M2 | 8 | 0 | **1,199** | **3,718** | 356,1 |
| M3 | 8 | 10 | 1,256 | 4,373 | 278,5 |
| M7 | 4 | 6 | 1,340 | 4,619 | 403,3 |
| M8 | 8 | 5 | 1,352 | 4,611 | 326,7 |
| M9 | 4 | 0 | 1,356 | 4,802 | 334,8 |
| M1 | 9 | 4 | 1,374 | 4,132 | **227,3** |

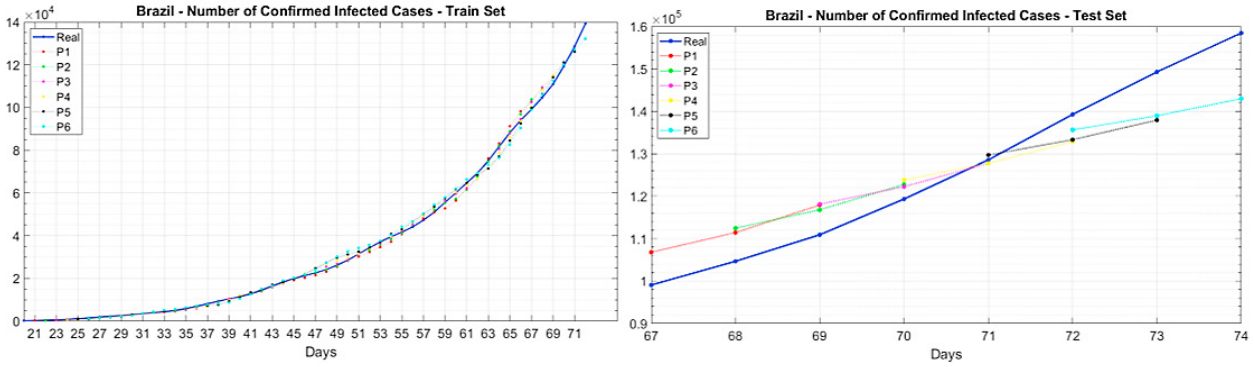


Fig. 2. Real and predicted curves of confirmed infected cases at the (a) train set and (b) test set for Brazil.

Fig. 2, graphically presents the real series, in blue, of the number of confirmed infected cases of Covid-19 for Brazil together with the $P1 - P6$ predictions made by the model on the train (Fig. 2.a) and test (Fig. 2.b) sets. It is noticed that the network's forecasts clearly follow the real series despite possible disruptions over time. It is also noticed, from the test set, that neurons of each output exhibit similar behaviors, despite being displaced at different distances from the real series, suggesting that the series will continue to grow at a similar rate in the following days.

Fig. 3, graphically presents the number of deaths over the days in Brazil, based on the train set (Fig. 3.a) and test set (Fig. 3.b). As for the number of confirmed cases, here the network demonstrates following the behavior of the curve over time. Despite this, there is an increase of error between the real and the forecast data, at the days 55 to 59. This can be justified by the occurrence of a more acute disturbance in the original series in this period when compared to the global behavior. However, it is noticed that the differences are momentary once the model adapts and the series returns to present low disturbance.

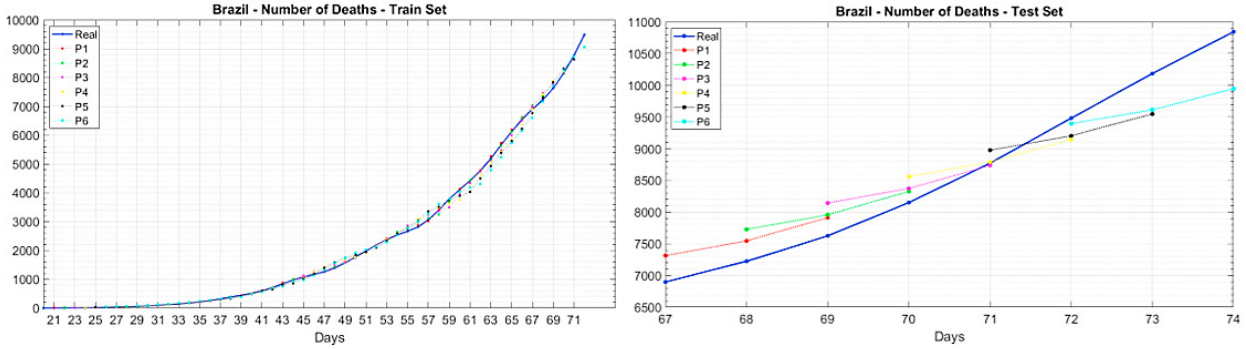


Fig. 3. Real and predicted curves of deaths at the (a) train set and (b) test set for Brazil.

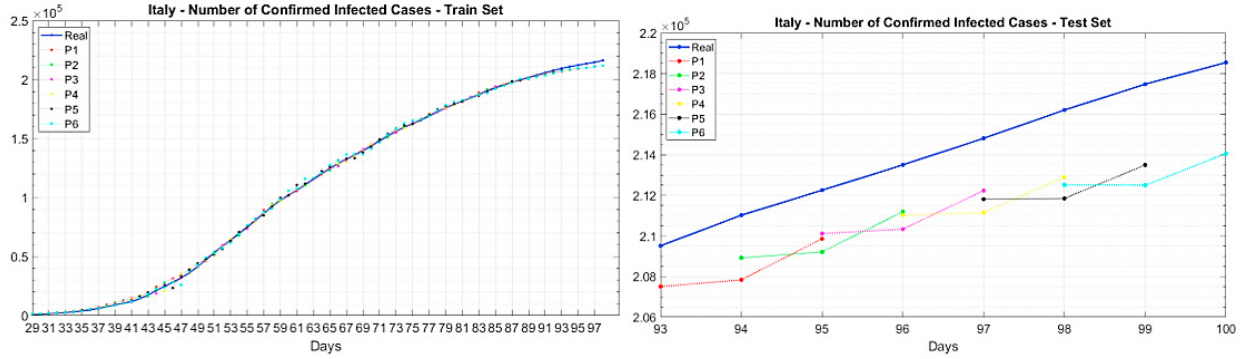


Fig. 4. Real and predicted curves of confirmed infected cases at the (a) train set and (b) test set for Italy.

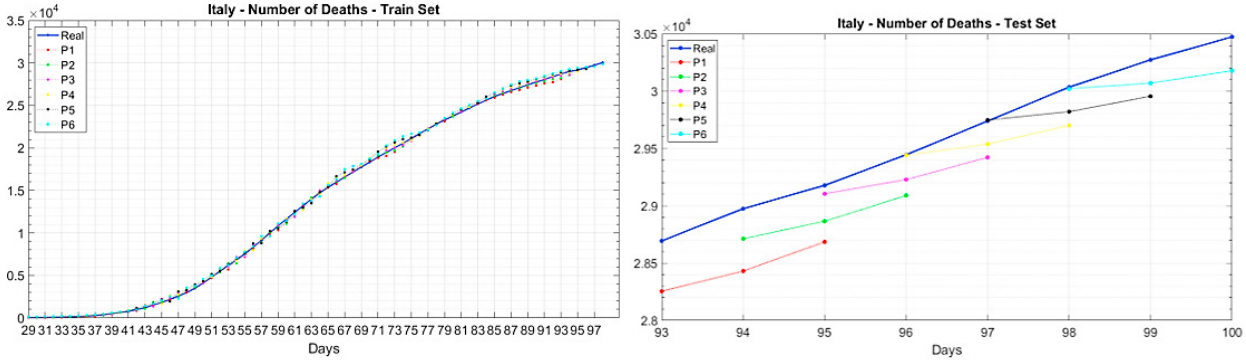


Fig. 5. Real and predicted curves of deaths at the (a) train set and (b) test set for Italy.

For Italy (Figures 4 and 5), the model showed to be able to represent the flattening behavior of the number of infected and dead curves in the train sets. It can be said that the contextualization by window of 20 samples, as was done in this work, is enough for the network to generalize the forecasts and adapt itself to eventual disturbances.

It is relevant to note this because it shows the network's ability to represent both the exponential acceleration behavior and the subsequent deceleration rate. In addition, the presentation of the results of the test sets make it clear that despite the deviation, the network seeks to adjust to the curve, having crossing points between expected and predicted over time.

## 6. Conclusions

The paper presented a model based on a MLP ANN that effectively describes and predicts for up to six days the behavior of time series related to the number of infected cases and deaths by COVID-19.

After searching for a set of parameters for the network that best suited the problem, an architecture was obtained with: two hidden layers, 9 and 4 neurons for layers one and two respectively and sigmoid tangent activation functions for the hidden layers and linear to the output layer. The choice of this architecture was made after comparing global and local measures of MSE and MAE. From these measurements and the observation of the graphical results, it was possible to notice that the 20-day contextualization window and the related features were sufficient for the network to learn and generalize the behavior of the series for all countries. With that, it can be said that, after being trained, the model is robust to be applied to each country alone or for all countries at the same time. In addition, the model satisfactorily described the behavior of exponential acceleration and later flattening of the series.

For future work, it is intended to make available the forecasts on an online platform providing daily updated information about the number of infected cases and deaths by COVID-19, for countries or even municipalities.

## Acknowledgements

This work has been supported by Fundação para a Ciência e Tecnologia within the Project Scope: UIDB/05757/2020.

## References

[1] C. Huang et al. (2020). “Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China,” Lancet, vol. 395, no. 10223, pp. 497–506, doi: 10.1016/S0140-6736(20)30183-5.

[2] Z. Wu and J. M. McGoogan. (2020). “Characteristics of and Important Lessons from the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases from the Chinese Center for Disease Control and Prevention,” JAMA - Journal of the American Medical Association, vol. 323, no. 13. American Medical Association, pp. 1239–1242, doi: 10.1001/jama.2020.2648.

[3] WHO, 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

[4] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao. (2020). “A novel coronavirus outbreak of global health concern,” The Lancet, vol. 395, no. 10223. Lancet Publishing Group, pp. 470–473, doi: 10.1016/S0140-6736(20)30185-9.

[5] D. D. Rajgor, M. H. Lee, S. Archuleta, N. Bagdasarian, and S. C. Quek. (2020). “The many estimates of the COVID-19 case fatality rate,” The Lancet. Infectious diseases. NLM (Medline), doi: 10.1016/S1473-3099(20)30244-9.

[6] D. L. Heymann . (2020). “Data sharing and outbreaks: best practice exemplified,” The Lancet, vol. 395, no. 10223. Lancet Publishing Group, pp. 469–470, doi: 10.1016/S0140-6736(20)30184-7.

[7] “Coronavirus COVID19 API.” [Online]. Available: https://documenter.getpostman.com/view/10808728/SzS8rjbc?version=latest#43e467ac-2cb0-4409-84e8-e18794e47271. [Accessed: 11-May-2020].

[8] “Relatório de Situação - COVID-19.” [Online]. Available: https://covid19.min-saude.pt/relatorio-de-situacao/. [Accessed: 11-May-2020].

[9] Haykin, Simon. (2007). Redes neurais: princípios e prática. Bookman Editora, 2007.

[10] “GitHub - CSSEGISandData/COVID-19: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE.” [Online]. Available: https://github.com/CSSEGISandData/COVID-19. [Accessed: 20-Jul-2020].

[11] “COVID-19 situation update worldwide, as of 20 July 2020.” [Online]. Available: https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases. [Accessed: 20-Jul-2020].

[12] “Tracking coronavirus: Map, data and timeline - BNO News.” [Online]. Available: https://bnonews.com/index.php/2020/04/the-latest-coronavirus-cases/. [Accessed: 20-Jul-2020].

[13] Rodrigues, P. M.; Teixeira, João Paulo. (2010) “Classification of Electroencephalogram Signals Using Artificial Neural Networks”. Proceedings of 3rd International Conference on BioMedical Engineering and Informatics (BMEI’10).

[14] Teixeira, J. P. and Freitas D. (2003). "Segmental Durations Predicted With a Neural Network", Proceedings of Eurospeech’03 – International Conference on Spoken Language Processing, Geneva. Pages 169-172.

[15] Rodrigues, Pedro M. and Teixeira, João Paulo, (2013). “Alzheimer’s Disease Recognition with Artificial Neural Networks” - chapter 7 (pag. 102-119) of the book “Information Systems and Technologies for Enhancing Health and Social Care”, by Ricardo Martinho, Rui Rijo, Maria Manuela Cunha and João Varajão. IGI Global. DOI: 10.4018/978-1-4666-3667-5.

[16] Silva, I.N. da, Spatti, D.H., Flauzino, R.A. (2010). “Redes Neurais Artificiais para Engenharia e Ciências Aplicadas”, Editora Artliber.

[17] Webster, J. G. (2010). Medical instrumentation : application and design. 4th ed. [s. l.]: J. Wiley & Sons. ISBN 9780471676003.

[18] E. Dong, H. Du, and L. Gardner, “An interactive web-based dashboard to track COVID-19 in real time,” Lancet Infect. Dis., vol. 20, pp. 533–534, 2020, doi: 10.1016/S1473-3099(20)30120-1.