Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 181 (2021) 973–980

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

COVID-19 Time Series Prediction

Leonardo Sestrem de Oliveira[a,b], Sarah Beatriz Gruetzmacher[a,b], João Paulo Teixeira[a,c]

[a] *Instituto Politécnico de Bragança, 5300-253, Bragança, Portugal*
[b] *Universidade Técnológica Federal do Paraná, 80230-901, Curitiba, Brazil*
[c] *Research Centre in Digitalization and Intelligent Robotics (CEDRI), Applied Management Research Unit (UNIAG) - Instituto Politécnico de Bragança, 5300-253, Bragança, Portugal*

Abstract

The Artificial Neural Network (ANN) is a computer technique that uses a mathematical model to represent a simpler form of the biologic neural structure. It is formed by many processing units and its intelligent behavior comes from the iterations between these units. One application of the ANN is for time series prediction algorithms, where the network learns the behavior of time dependent data and it is able to predict future values. In this work, the ANN is applied in predicting the number of COVID-19 confirmed cases and deaths and also the future seven days for the time series of Brazil, Portugal and the United States. From the simulations it is possible to conclude that the prediction of confirmed cases and deaths from COVID-19 have been successfully made by the ANN. Overall, the ANN with a specific test set had a Mean Squared Error (MSE) 50% higher than the ANN with a random test set. The combination of the sigmoidal and linear activation functions and the Levenberg-Marquardt training function had the lowest MSE for all cases.

Keywords: Time Series, Covid 19 Prediction, Brazil Covid 19 Prediction, Portugal Covid 19 Prediction, USA Covid 19 Prediction

* Corresponding author. Tel.: +351 273 30 3129.
E-mail address: joaopt@ipb.pt

10.1016/j.procs.2021.01.254

1. Introduction

The ANN is a computer technique that uses a mathematical model to represent a simpler form of the biologic neural structure. The non-linear nature of this net and its ability to learn, with or without supervision, from its environment are the reasons this network is the optimal tool to solve problems of temporal series prediction [1]. The distinction of the learning method with or without supervision is related to knowing the outputs or not, respectively. When learning with supervision, the inputs and the outputs are provided to the network. Then, the network processes the inputs and compares the outputs results with the outputs desired. Errors are then propagated back through the system to adjust the weights which control the network. In the method without supervision, the network is provided with the inputs but not with the outputs desired. The system itself must decide what it will use to group in the input data. This decision is often referred to as self-organization or adaption. The network will need to find statistic attributes with the objective of representation of the motivation that come in the network [2].

An ANN is formed by many processing units, usually connected by communication channels and related to specific weights. The units do the operations using only the input data received from the connections. The intelligent behavior of the network comes from the iterations between these units [3-4]. Fig. 1 shows an example of an ANN. The network is formed by the input layer, where the patterns are presented, the hidden layer, where through weighted sums most of the processing is made and the output layer, where the result are presented.

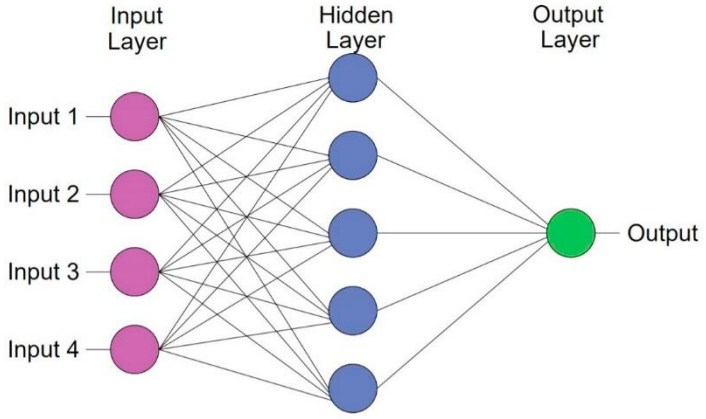

Fig. 1: Example of structure of the generic neural network.

After creating the neural network, the next step is the training. In this step the weights are updated by iterative method, with the objective of minimizing the MSE. In this case the technique used is the stochastic descendent gradient or backpropagation, where the weights are adjusted in the negative gradient direction of the target function [4].

Time series prediction algorithms is widely applied in many diverse areas, e.g., financial market, weather forecasting and complex dynamical system analysis. The analysis of these series is important to study the behavior of time dependent data and to forecast its future values depending on the history of variations in the data [6-7].

The objective of this article is applying the ANN in temporal series prediction, where this tool helps to analyze the dynamic of the system, learning its behavior, allowing to reproduce it and predict future values for the time series [8]. The analysis of the series in the time domain has focus in describing the magnitude of events that occur at certain times and in the relation between the observations in different time periods [8].

The theory and development of the ANN had a great evolution in the last decades and with the actual global scenario of the new coronavirus this tool can be useful to predict the parameters, risks, and effects of such an epidemic [1], [9]. In addition, it is useful for estimating the dynamics of transmission, targeting resources and evaluating the impact of intervention strategies [10]. In opposition, Uhlig et al. says in [11] that forecasting and predictions during a crisis is a double-edged sword, because the risk of incorrect predictions or unreliably large uncertainty intervals can be fatal.

2. Methodology

The first step was to collect the COVID-19 time series data. In this work the daily cumulative number of cases and deaths by the COVID-19 from eight countries were used as input. The countries chosen were the United States, South Korea, Germany, China, Portugal, Spain, Italy and Brazil. Since these numbers were updated daily, it was important to use a tool to update the database. Using the software Matlab, the data for each country was collected using an API [12] and the *webread* function, that reads the content from the web service specified by a *url* and returns it as data. The function reads the data provided by the API through the website *api.covid19api.com/total/dayone/country/x,* where the x mean the countries name. This function returns a structure array with the cumulative data for the confirmed, deaths, recovered and active cases, for each day, since the day one of the infection in that country. With this structure, every time the program runs the data is pulled directly from the website and is always up to date.

To be possible to analyze the information for all those countries together the data needed to be normalized. The numbers for each country were therefore divided by each series maximum value. This way each time series was composed of values ranging from zero to one. The maximum value was saved in a variable, so it was possible to return the values to the original scale after the predictions. The normalized series for the number of confirmed cases is shown in Fig.2.

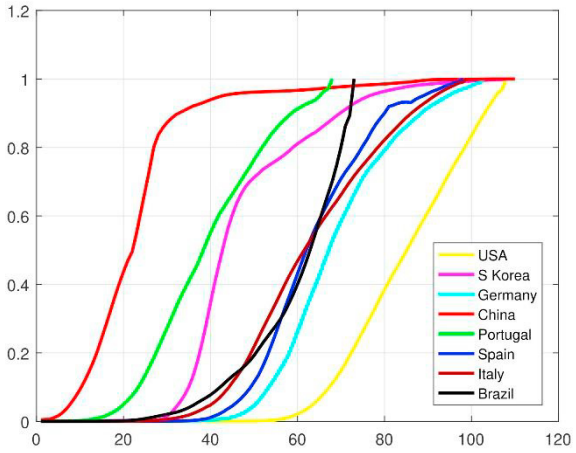

Fig. 2: COVID-19 confirmed cases time series.

Fig.3 shows the normalized number of deaths for the countries under study. The number of deaths by COVID-19 in China showed a large increase between the 16th and 17th of April, in a moment where the numbers were already very low in the previous days. A revision on the official deaths number in Wuhan increased the numbers by 50% and, with it, the overall numbers for China [13]. The time series therefore differs greatly from those of the other countries. For this reason, it was chosen not to use China's death time series in the predictions.

The next step was to organized the confirmed cases and deaths data in matrices to be used as the input for the ANN. For every country, the data was organized in a vector for the number of confirmed cases and one vector for the number of deaths. The objective of this work was to create two ANN, one for predicting the number of confirmed cases and one to predict the number of deaths, using the data of the five previous days to predict the next one. To achieve this, using the for structure, the vectors were organized in matrices where each column represented one input and was composed of the information of the previous five days. The input matrices were created for each country separately and then concatenated together, creating one single input matrix for the confirmed cases and one for the number of deaths.

The output of the ANN, the target t, was defined as a vector with the data since the sixth day of infection, as this was the first prediction made by the network.

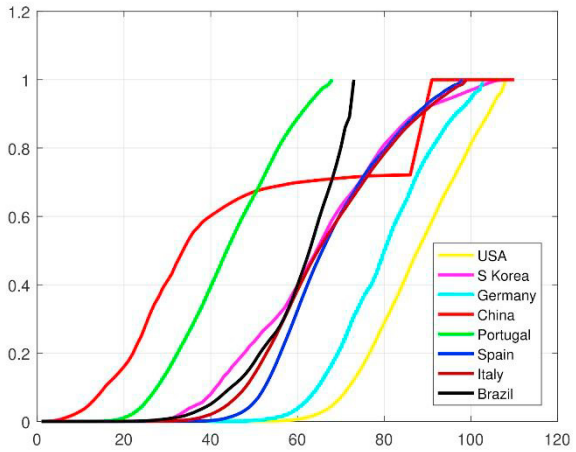

Fig. 3: COVID-19 deaths time series.

Both Neural Networks were created using the Matlab command *newff*, that generates a feed-forward backpropagation ANN. Three different combinations of activation functions with different number of nodes and training functions were tested, first with the training, validation and test sets not defined and, after, defining the validation set as the data for Italy and the test set with the data from Brazil. The MSE was calculated for all those cases and the best performing combination was chosen.

Both ANN, for confirmed cases and deaths, were then trained with the data from all eight countries, using the best combination of activation function, training function and number of nodes. After, inside a for loop the prediction for seven future days was made. As before, the prediction for the next day was made using the five previous days, the predicted number was then incorporated in the data set for the next loop. This prediction was made for the number of confirmed cases and deaths for Brazil, Portugal and United States.

3. Results

Two different training methods were used: Levenberg-Marquardt and the Resilient Propagation. The Levenberg-Marquardt method provides a numerical solution to the problem of minimizing a non-linear function, with a combination of speed and stability in the convergence. The basic idea of the algorithm is that it performs a combined training process around the area with complex curvature to make a quadratic approximation and speed up the convergence significantly, combining two algorithms (steepest decent and Gauss-Newton), where the first is responsible for the stability and the second for the convergence speed [14]. The Resilient Propagation method performs a direct adaptation of the weight step based on local gradient information. The effort of adaptation is not blurred by gradient behavior whatsoever, it only depends on the sign of the derivative not its value and therefore it will converge from ten to one hundred times faster than the simple backpropagation algorithms [15].

3.1. Behavior of the Neural Network

The ANN accomplish its objective of predicting the number of confirmed cases and deaths without having the training, validation and testing sets defined, that was, therefore, picked randomly by the program. The ANN was also

successful when the data for Brazil was not in the training set and was used only for testing, although in this case the performance was worse. In these simulations, with the input of the previous five days in the matrix p it was expected the prediction of the sixth day, presented in the output matrix t. To validate the accuracy of the predictions made by the ANN, the MSE was calculated.

The best results were achieved with the Levenberg-Marquardt training function and the activation sigmoidal and linear functions *logsig* and *purelin*. As the behaviour of the data is crescent, a linear function in the output layer was more adequate [16]. This combination with 12 nodes in the hidden layer was used to estimate the number of confirmed cases and deaths for the future seven days for Brazil, Portugal and United States. It was used a different ANN for the number deaths and the number of confirmed cases; therefore, two predictions were done separately.

3.2. Simulations

The results of the simulations with a defined training and testing sets and also a random sets for the number of infected cases are presented in Table 1, and for deaths in Table 2. These simulations were done using two different training functions, Levenberg-Marquardt and Resilient Backpropagation and three different activation functions, tangent sigmoidal (*tansig)*, logarithmic sigmoidal (*logsig)* and Elliot symmetric sigmoid transfer function *(elliotsig)*. In the output layer the linear activation function (*purelin)* was always used, since the others functions had worst results and as the data have a crescent behaviour, the use of this function is justifiable.

Table 1 – MSE for the prediction of confirmed cases.

	Test set: Brazil			Test set: Random		
Activation Function	**Nodes**	**Levenberg-Marquardt**	**Resilient Propagation**	**Nodes**	**Levenberg-Marquardt**	**Resilient Propagation**
tansig - purelin	8	2.17E-05	6.29E-05	8	3.61E-05	1.26E-04
	12	1.66E-05	6.57E-04	12	2.17E-05	9.55E-05
	30	1.55E-05	3.04E-04	30	1.70E-05	**6.96E-05**
	50	1.55E-05	3.04E-04	50	1.28E-05	1.46E-04
logsig - purelin	8	1.61E-05	5.31E-05	8	1.93E-05	7.93E-05
	12	**1.35E-05**	4.27E-05	12	1.88E-05	9.93E-05
	30	1.68E-05	**2.99E-05**	30	1.97E-05	2.09E-04
	50	1.51E-05	1.44E-04	50	2.15E-05	1.46E-04
elliotsig - purelin	8	1.76E-05	9.54E-04	8	3.76E-05	2.79E-04
	12	2.16E-05	3.86E-04	12	1.35E-05	6.70E-04
	30	1.51E-05	4.09E-04	30	**1.17E-05**	2.07E-04
	50	2.61E-05	9.33E-04	50	2.42E-05	6.06E-04

3.3. Results and Discussion

Analysing the results of the simulations and comparing the two training functions, the Levenberg-Marquardt function had a better performance than the Resilient Propagation function. This is noticeable from the MSE values as the results in the Levenberg-Marquardt are ten times lower than the Resilient Backpropagation results.

In the prediction of the number of confirmed cases, the best results were obtained using the combination of activation function *logsig-purelin* with 12 nodes in the hidden layer, with an MSE of 1.35E-05 when using the data of Brazil in the test set. When using a random test set, the best result was obtained with the combination of activation function *elliotsig-purelin* with 30 nodes in the hidden layer with an MSE of 1.17E-05.

For the deaths prediction when using Brazil's data in the test set, the best results were obtained by the combination of *elliotsig-purelin* as the activation functions and the use of 30 nodes in the hidden layer, resulting in an MSE of 5.10E-03. When using the random test set, the best combination was using *tansig-purelin* with 30 nodes resulting in an MSE of 1.80E-03.

Table 2 – MSE for the prediction of deaths.

	Test set: Brazil			Test set: Random		
Activation Function	**Nodes**	**Levenberg-Marquardt**	**Resilient Propagation**	**Nodes**	**Levenberg-Marquardt**	**Resilient Propagation**
tansig - purelin	8	5.90E-03	6.60E-03	8	5.80E-03	6.90E-03
	12	7.70E-03	7.00E-03	12	2.80E-03	6.10E-03
	30	5.30E-03	7.20E-03	30	**1.80E-03**	7.60E-03
	50	5.40E-03	**6.00E-03**	50	3.60E-03	**4.90E-03**
logsig - purelin	8	7.90E-03	6.80E-03	8	2.70E-03	7.40E-03
	12	5.20E-03	7.10E-03	12	2.20E-03	6.70E-03
	30	5.20E-03	8.70E-03	30	2.50E-03	7.60E-03
	50	5.20E-03	7.30E-03	50	2.30E-03	5.10E-03
elliotsig - purelin	8	9.10E-03	8.90E-03	8	3.10E-03	6.80E-03
	12	7.60E-03	8.50E-03	12	8.70E-03	7.40E-03
	30	**5.10E-03**	7.50E-03	30	2.30E-03	6.30E-03
	50	5.70E-03	6.80E-03	50	8.10E-03	7.90E-03

Even though the ANN accomplishes the objective of predict the number of confirmed cases and deaths of the COVID-19, the predictions of the confirmed cases had an MSE one hundred times lower than the MSE of the number of deaths prediction. This is noticeable when comparing the results presented in the tables with the figures for the estimation in the figures bellow. Fig. 4, 5 and 6 show the predictions made by the ANN for the number of confirmed cases for Brazil, Portugal and the United States, respectively. Figure 7, 8 and 9 show the prediction made by the second ANN for the number of deaths for the same countries. These graphics show that, overall, the curves of the predicted number of confirmed cases has a better fit when compared to the database then the curves for the predicted number of deaths. The prediction of seven future days was also made. The simulations made with the random test set had a better performance than when a specific country data was used in the test set. These results can be understood as in the random case the training of the ANN was made using the data of the all countries, making the prediction easier than when the data for one country is totally unknown by the ANN in the training process.

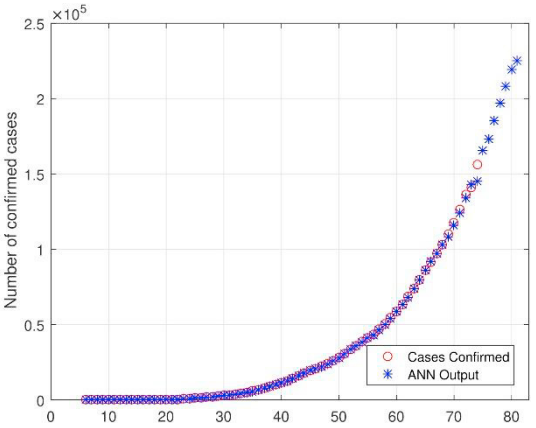

Fig.4 - Brazil confirmed cases prediction.

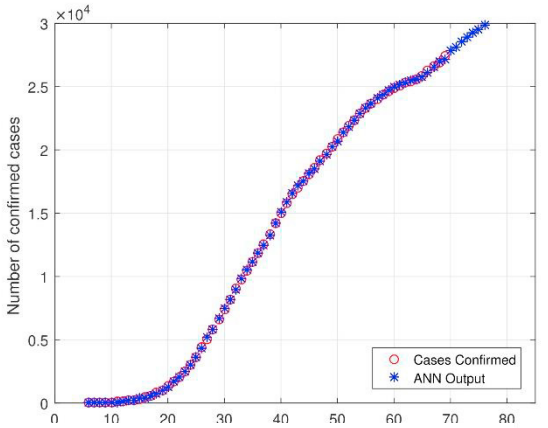

Fig.5 - Portugal confirmed cases prediction.

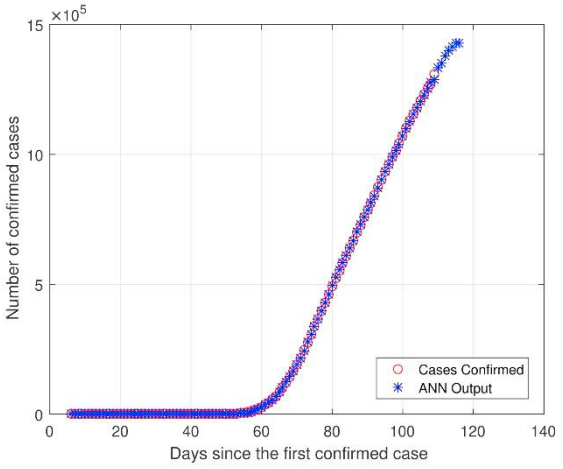

Fig.6 - United States confirmed cases prediction.

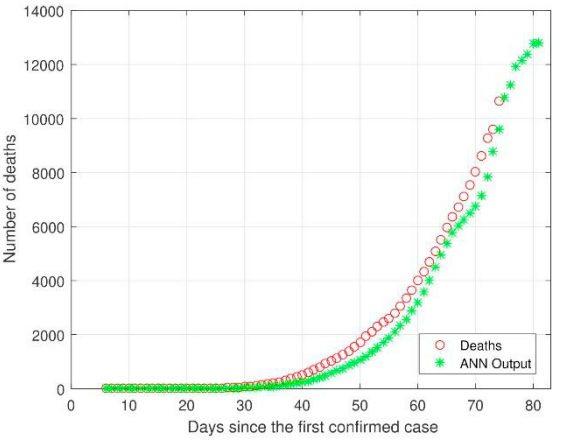

Fig.7 - Brazil deaths prediction.

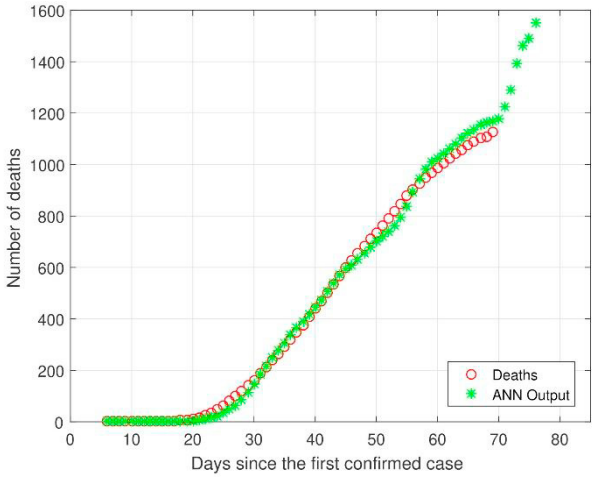

Fig.8 - Portugal deaths prediction.

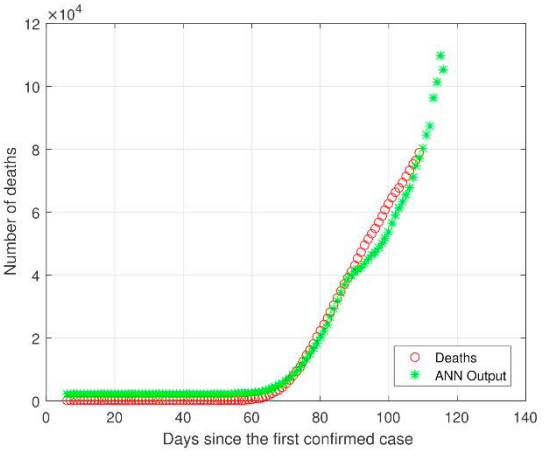

Fig.9 – United States deaths prediction.

4. Conclusion

From the simulations it is possible to conclude that the prediction of confirmed cases and deaths from COVID-19 have been successfully made by the ANN. For the prediction of the seven future days without the confirmed data as the output, the values estimated followed a logical tendency that fitted the respective curves. Overall, the ANN with a specific test set had a worse performance than the ANN with a random test set, as the MSE in the defined test set was, on average, 50% higher than the random set case. Due to the data used as test set also being a part of the training set in the random case, the prediction has a better result.

Two different training functions were tested for both cases with different combinations of activation functions and number of nodes in the hidden layer. The Resilient Propagation function shown mostly an MSE ten times bigger than the MSE using the Levenberg-Marquardt function. The combination of the *logsig-purelin* activation function with the Levenberg-Marquardt training function presented the best results for all cases, and the use of 12 nodes in the hidden

layer presented the lowest MSE or in some cases, as shown in Table 2, the increase of the number of nodes above 12 did not improve the performance of the ANN.

The trained neural networks were than used to predict the number of confirmed cases and the number of deaths for the future seven day in Brazil, Portugal and United States. With a visual analysis it is possible to notice that the prediction for confirmed cases presented better results than the deaths prediction. This result was expected, since the MSE for the confirmed cases ANN was lower, showing that the network estimated the confirmed cases more easily.

Acknowledgements

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

References

[1] Hu, Y. H.; Hwang, J. Introduction to Neural Networks for Signal Processing. CRC Press LLC, United States of America (2002).

[2] Rauber, T. W. Redes neurais artificiais. Universidade Federal do Espírito Santo, 29 (2005).

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

[4] Rodrigues, Pedro M. and Teixeira, João Paulo; "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, 2013. DOI: 10.4018/978-1-4666-3667-5.

[5] Facure, M. Uma abordagem intuitiva as redes neurais, https://lamfo-unb.github.io/2017/06/18/itro-ao-deep-learning/. Last accessed 29 Mar 2020.

[6] Qin, Y. et al. A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. (2017).

[7] Zitao, L.; Milos, H. A regularized linear dynamical system framework for multivariate time series analysis. In AAAI, pages 1798−1805 (2015).

[8] Ferreira, B. B.; Gois, J. A. M. Predição de Séries Temporais Biológicas Utilizando Ferramentas da Logica Nebulosa. Associação Brasileira de Engenharia e Ciências Mecânicas, Anais, Brasil (2009).

[9] Pal, R.; Sekh, A.A.; Kar, S.; Prasad, D.K. Neural Network Based Country Wise Risk Prediction of COVID-19. Preprints. (2020).

[10] Hu, Z.; Ge, Q.; Jin, L.; Xiong, M. Artificial intelligence forecasting of covid-19 in china. arXiv preprint arXiv:2002.07112. (2020).

[11] Uhlig, S.; Nichani, K.; Uhlig, C.; Simon, K. Modeling projections for COVID- 19 pandemic by combining epidemiological, statistical, and neural network approaches. medRxiv (2020).

[12] Coronavirus COVID19 API, https://documenter.getpostman.com/view/ 10808728/SzS8rjbc?version=latest. Last accessed in 11 Apr 2020.

[13] Wuhan officials have revised the city's coronavirus death toll up by 50%, https:// edition.cnn.com/2020/04/17/asia/china-wuhan-coronavirus-death-toll-intl-hnk/ index.html. Last accessed in 09 Apr 2020.

[14] Yu, H.; Wilamowski, B. M. Levenberg-marquardt training. Industrial electronics handbook, 5.12: 1 (2011).

[15] Riedmiller, M.; Braun, H. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In: IEEE international conference on neural networks. IEEE, p. 586-591 (1993).

[16] Dorofki, M., et al. Comparison of artificial neural network transfer functions abilities to simulate extreme runoff data. International Proceedings of Chemical, Biological and Environmental Engineering, 33: 39-44 (2012).