COVID-19 spread forecast using recurrent auto-encoders

Radu Beche
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
radubeche@protonmail.ch

Romina Băilă
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
romina.baila16@gmail.com

Anca Marginean
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
anca.marginean@cs.utcluj.ro

***Abstract*—Since December 2019, the COVID-19 disease has become one of the most concerning issues across the Globe. The world has experienced serious prevention measures, from lockdown of cities to entire countries, as it is still unknown when will this pandemic end. Thus, it is no surprise that predicting the spread of this novel coronavirus has quickly become a hot topic among Artificial Intelligence researchers. In this paper we try to solve this task, by leveraging the capabilities that Recurrent Auto-Encoders have on time-series and implying a semi-supervised training process. Furthermore, the concept of *nearest neighbour* countries is introduced to estimate the cumulative number of confirmed cases for any country. The results are promising, showing that our proposed method is capable of making reliable predictions for a 30-days period. It is worth mentioning that, while this study uses information related to COVID-19, the proposed method can be used to estimate the evolution of any kind of disease, provided that the associated data comes in form of time-series.**

I. Introduction

Coronaviruses are a large family of viruses affecting both animals and humans. The impact they have on humans is usually in the form of respiratory infections. In December 2019, a novel coronavirus, officially named SARS-CoV-2, appeared in Wuhan, the capital of Hubei province, China. Although its source is still unknown, it is thought to have a zoonotic origin, namely coming from bats. The disease associated with it, COVID-19, spread outside of China so rapidly, that the World Health Organization (WHO) has declared it a pandemic. The disease is now present all over the world, affecting over 200 countries and territories. So, it is no surprise that a majority of nations have imposed strict social distancing measures, like banning social gatherings, restricting free movement and even closing the borders of cities and countries.

Because currently there is no vaccine or specific treatment to cure COVID-19, it is extremely important for health ser-vices and governments across the world to have information regarding the virus' transmission rates. For these reasons, the prediction of COVID-19's development could give authorities an insight on how to adapt the restrictive measures in their countries and manage their resources to effectively contain its spread.

In this paper we propose a method for predicting the total number of confirmed cases made by the novel coronavirus, using a Recurrent Auto-encoder. The data comes from the research paper Dong et al., 2020 [1], from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, Baltimore, USA. A distinct model is trained every time we make a prediction for a new country. We introduce the preprocessing concept of *nearest neighbours*, to select a country's train data only from those countries where the spread of the virus is similar. Moreover, a *semi-supervised training* process is implied, to better model the real-life evolution of the coronavirus. The results that we obtained for our predictions have a margin of error of approximately 2% over a period of 30 days.

The main contributions of this paper can be summarized as follows:

1) A new preprocessing concept of nearest neighbours, to select only data from similar countries (in terms of virus spread) when making predictions for a specific country.
2) Applying an unsupervised learning process, to obtain a more stable model and to guide it for longer periods of time.

The code is made publicly available at the following link: GitHub

II. Related Work

Viruses outbreaks have always been of interest to researchers, whose main goal is to predict the evolution of their spread. The new SARS-CoV-2 is no exception, therefore many studies have already been conducted and predictions for different time periods have already been made. These studies can be divided into two categories, based on their underlying method of prediction. The first category uses non-AI models, such as *epidemiological models*. The second category involves the use of AI models, simple or in combination with improving algorithms. Thus, we will present in this section papers from both categories, noting that some of them refer to other diseases, not COVID-19.

A. Non-Deep Learning models

Forna et al., 2019 [2] estimated the case fatality ratio of West African Ebola, using a Boosted Regression Tree model. Ceylan, 2020 [3], applied a number of Auto-Regressive

Integrated Moving Average (ARIMA) models, using different parameters, to determine the future evolution of COVID-19 in Italy, Spain and France. Flaxman et al., 2020 [4] proposed a semi-mechanistic Bayesian hierarchical model to predict the total number of cases and the number of deaths in 11 European countries. They also conclude whether the preventive interventions adopted in these countries were effective or not. Zhao et al., 2020 [5] used a mathematical Poisson process to estimate the unreported number of cases in China from the beginning of the pandemic. Lastly, Tang et al.,2020 [6], Peng et al., 2020 [7], Chen et al., 2020 [8], Anastassopoulou et al., 2020 [9] and Fanelli et al., 2020 [10] all used simple or modified SIR models, commonly used for epidemics analysis, to forecast the evolution of COVID-19 and determine the effectiveness of the preventive measures implemented by authorities.

The main shortcoming of these models are the number of parameters needed. For example, in [6] a SEIR model is used and 12 parameters are needed (most of which have to be computed). Then, a lot of assumptions are made, like in [4], the authors assume that the preventive measures have the same impact in all 11 countries studied and, because they have more data available, advanced countries (e.g. Italy, Spain) have a more significant influence on the final prediction for each country, than countries with less data. Moreover, in papers like [3] using ARIMA models, it is proved that a different model configuration is needed for each country, that is, for each country, a new set of optimal parameters have to be determined.

B. Deep Learning models

In their paper, Liu et al., 2019 [11] tested a Back Propagation Neural Network (BPNN) to forecast the trend of tuberculosis in China. Dandekar et al., 2020 [12] used a neural network augmented model to predict the evolution of COVID-19 in Wuhan, Italy, South Korea and USA and concluded that the implemented quarantine measures had efficiently stopped the spread of the disease. In another paper, Zeng et al., 2020 [13] used a multi-model ordinary differential equation set neural network (MMODEs-NN) to predict the ending of the COVID-19 transmission in China. Many researchers, like Tomar et al., 2020 [14], Pal et al., 2020 [15], Yang et al., 2020 [16] used Long Short-Term Memory (LSTM) based models, known to be suitable for time-series, to forecast the spread of the novel coronavirus and asses the effectiveness of the prevention methods. Finally, Hu et al., 2020 [17] proposed in their paper a modified stacked auto-encoder for modelling the transmission dynamics of COVID-19 and for real-time forecasting of the confirmed cases across China.

After analyzing the related research, we decided to use a **recurrent auto-encoder**, described below, to leverage the existing AI capabilities of working with time-series.

III. Proposed approach

A. Data

The data format that is used in this work is collected by the John Hopkins institute in collaboration with multiple international agencies such as World Health Organization (WHO), European Centre for Disease Prevention and Control (ECDC), Centers for Disease Control and Prevention (CDC) and others. They separately provide a per day, per country number of cumulative confirmed cases, fatalities and recovered persons (for larger countries, data is provided per region/county/state). The dataset contains entries for almost every country in the world, since 11 January 2020 and is updated on daily basis, as the pandemic progresses. This work focuses on the prediction of the cumulative number of confirmed cases as reported.

B. Data preprocessing

The section is divided into 2 main stages, one which aims to align the data to a temporal interval considering the day of reaching a certain number of confirmed cases in a specific country and another for grouping together countries that have evolved in a similar manner.

Zero-day alignment is the process of translating the spread data for each country from an absolute to a relative timeline, given a certain alignment threshold $\mathcal{T}$. A pandemic does not start at the same time in all countries, and even after the apparition of the first few cases, there is an amount of time for which they are contained (see Fig. 1). These elements can be considered to have a random nature and they are not relevant for predicting an advanced spread. By applying zero-day alignment we can compare and group together the countries that experience a similar growth, which is useful when making a forecast.

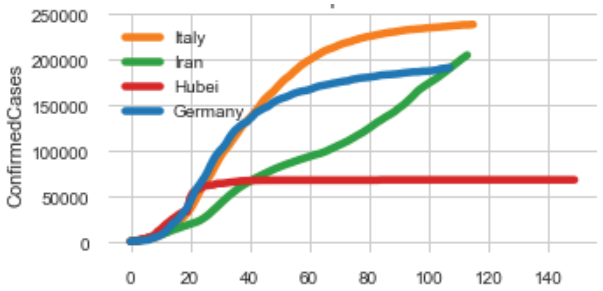

Fig. 1. Example of COVID-19 spread presented after the zero-day alignment step. On y axis is are the number of cumulative number of confirmed cases and x axis represents number of days since a country reached the alignment threshold T

Nearest neighbours algorithm is used to group together the countries with similar growth. To obtain these neighbours for a target country $\mathcal{C}_s$ we proceed as follows: first, we apply the zero-day alignment having a threshold $\mathcal{T}$ for every country in the dataset ($\mathcal{C}_s$ included). Then, we consider a candidate country, $\mathcal{C}_t$ that is more evolved than $\mathcal{C}_s$ (it reached $\mathcal{T}$ earlier). We start sliding $\mathcal{C}_s$ over $\mathcal{C}_t$ beginning with the first day it reached the threshold, until $\mathcal{C}_t$ ends. For each step, an error $\mathcal{L}_{step}(\mathcal{C}_s, \mathcal{C}_{t_i})$ is computed and stored. The smallest error $\mathcal{L}_{step}$ will be the final error ξ associated with $\mathcal{C}_t$. We do this for all candidates in the dataset, taking one individual feature f, $f \in \{confirmedcases\}$ at a time. This will result in a collection of neighbours, from where the training data for the model is

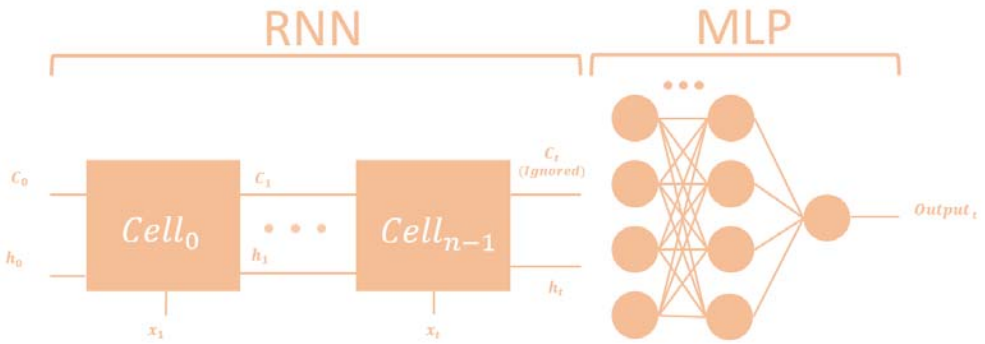

Fig. 2. Recurrent auto-encoder architecture comprised from a group of reccurent cells and a multi layer perception having multiple hidden layers.

going to be selected, given a certain error threshold $\mathcal{T}_{error}$ (only keep data below this threshold).

C. Recurrent auto-encoder

The model architecture (Figure 2) is an encoder-decoder structured as follows:

Recurrent Neural Network - composed of recurrent cells, which embed the temporal information from the input series. For this we are going to make use of Long Short-Term Memory cells[18]. Such a unit is composed of multiple gates, an input gate, an output gate and a forget gate. The exact operations that a cell does are:

$$\begin{aligned} i_t &= \sigma(w_i[h_{t-1}, x_t] + b_i) \\ f_t &= \sigma(w_f[h_{t-1}, x_t] + b_f) \\ o_t &= \sigma(w_o[h_{t-1}, x_t] + b_o) \\ \tilde{c}_t &= tanh(w_c[h_{t-1}, x_t] + b_c) \\ c_t &= f_t * c_{t-1} + i_t * \tilde{c}_t \\ h_t &= o_t * tanh(c_t) \end{aligned} \tag{1}$$

Where, c and h are the hidden and cell state passed from the last timestamp, * is the Hadamard product operator and σ is the sigmoid activation function.

Multi layer perceptron composed of multiple hidden linear units, which decode the aforementioned latent representation to reconstruct the input series.

D. Learning process

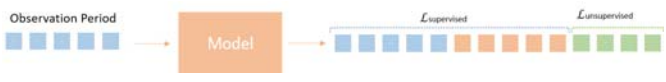

Fig. 3. Model output interpretation

During training, the model takes as input the data for an observation period T_x, and produces an output series for multiple days into the future. This can be divided into 3 parts:

1) P_r the reconstruction of the input sequence; each prediction is based on the exact timestamp from the input sequence, this is the auto-encoding branch.
2) P_n is the prediction for the next n days into the future, given the last day of P_r; each prediction is based on the prediction from the previous timestamp.
3) P_l is the prediction for the next m days after P_n; each prediction is based on the prediction from the previous timestamp; we do not have label for this period.

For each of the aforementioned categories, a certain criterion is optimized as follows:

Reconstruction loss is going to minimize the prediction error. It is measured using the Huber error function [19] between the target y and the reconstructed sequence P_r and the future forecast P_n. The motivation for choosing this is the fact that it is less sensitive for outliers and leads to a better model stability during training. For a given timestamp, the loss can be defined as:

$$f_\delta(y_i, x_i) = \begin{cases} \frac{1}{2}(y_i - x_i)^2, & |y_i - x_i| \leq 1 \\ |y_i - x_i| - \frac{1}{2}, & otherwise \end{cases} \tag{2}$$

and for an entire period, the error becomes the average loss measured at each timestamp:

$$\mathcal{L}_r(y, x) = \frac{1}{T}\sum_{0}^{T} f_\delta(y_i, x_i) \tag{3}$$

where T denotes the number of time stamps fed into the network.

Saturation loss is going to limit the increase of the prediction over large periods of time and make it reach a saturation plateau. It is limiting the gradients of the forecast over larger periods of time by being smaller than a certain weighted value h computed on the training set. The weight of h decreases as the according timestamp is larger following a linear decay:

$$y_i = \alpha - \frac{i}{t} * \beta \tag{4}$$

where $\alpha, \beta \in \mathbb{R}$ are considered hyperparameters, t denotes the total number of time-stamps and i denotes the timestamp number. The unsupervised loss can be expressed as:

$$\mathcal{L}_s = -\frac{1}{t}\sum_{i=1}^{t} h * y_i - \nabla x_i \tag{5}$$

Where x_i denotes the prediction at timestamp $i, i \in t$.

The **total loss** is the sum of $\mathcal{L}_r$, $\mathcal{L}_s$ using the according scale factors λ_r and λ_s:

$$\mathcal{L}_{total} = \lambda_r \mathcal{L}_r + \lambda_m \mathcal{L}_m \tag{6}$$

IV. Experiments

A. Implementation details

Data The data was first transformed using zero-day alignment. We found out that a good alignment threshold was $\mathcal{T}$=400 confirmed cases. To compare the countries we choose the error function $\mathcal{L}_{step}$ to be the **Mean Average Percentage Error**. We compared multiple criterions and obtained very similar results, so we chose this one for the ease of interpretability and for the convenience when comparing our results with the ones found in other related research papers. The threshold for $\mathcal{T}_{error}$ was chosen to be 40% and everything above this was discarded.

Model For the model architecture, the recurrent neural network encoder is composed from a single LSTM cell, followed by a 3-layered MLP. The prediction for a country implies training a new model. Thus, for distinct countries we will have distinct models. Our model was trained using the Adam [20] optimizer for 150 epochs, with a batch size of 16. We chose $\lambda_r = 1.0$, $\lambda_i = 0.5$, Each of the time windows considered for reconstruction and for future prediction were of 5 days. We discerned that the long term predictions for a country were very dependable on the amount of data available for training. Thus, we concluded that a forecasting period of 15 days is considered to be stable for all situations (both when we have more data and less data).During training, the gradients of the model were clipped such that they have a norm of 1. This was done to avoid the exploding gradients problem.

The experimental part was implemented in Python using the Pytorch library.

B. Results

To validate our model, we trained it using multiple time frames: up until April, May and June respectively, each time adding one more month. The mean average percentage error (MAPE) was used to compute the validation error. The rest of the available data (up until the current date), were used for validation. We are confident that the results we obtained have a 2% margin of error. In Table I, several results that were obtained using our model for a number of European countries (arbitrarily chosen) are presented. Moreover, Figure 4 illustrates the spread prediction of COVID-19 in Romania, for a 30-day period.

Comparison of results It is difficult to make comparisons between our model and other proposed methods, because there are many factors that need to be taken into consideration. First of all, there is the evaluation metric, which, in our case is MAPE. Thus, we can consider only those methods using the same evaluation metric. Then, we have the dataset, which we described before. Also, we need to take into consideration the prediction period, meaning for how many days was the prediction made and for which month(s). The month is important because when trying to predict for an earlier month, fewer data is available for training, so, obviously, the model will perform worse. Having all this in mind, the following are rough comparisons between our results and the ones from related research papers, noting that in all cases the training datasets are different than ours. In [3], an ARIMA (0,2,1) model is used to make predictions for a 10 days period (in April) for Italy, having MAPE = 4.752. Şahin et al., 2020 [21] used a nonlinear grey Bernoulli model to make predictions for a 35 days period (March-April), also for Italy, having MAPE = 9.68. In comparison, our model, for a 30 days period prediction for Italy has MAPE = 0.19. In [22], an improved ANFIS model is used to make predictions for China for a 30 days period (May-June), having MAPE = 4.79. In [23], the authors used a Single Exponential Smoothing (SES) model to make predictions for all available countries, for a 35 days period (March-Apri), with a MAPE = 0.917. In contrast to this, our model, when making predictions for all countries, for a 30 days period, has a MAPE = 0.4.

TABLE I
THE RESULTS PRESENTED IN THE TABLE CONTAIN THE MAPE FOR A MODEL PREDICTING ON VARIOUS PERIODS OF TIME. THIS RESULTS WERE OBTAINED BY BACKTESTING.

Country	3 months [%]	2 months [%]	1 month [%]
Romania	7.5	1.3	0.44
Germany	4.6	0.7	0.61
Italy	6.8	1.6	0.19
Poland	9.2	2.9	0.55
Spain	5.5	1.02	0.24

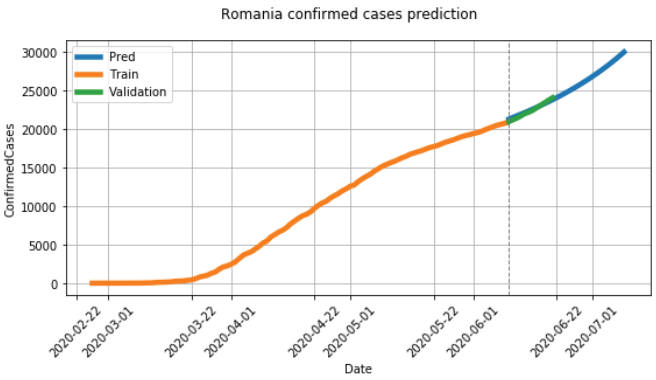

Fig. 4. Model prediction for a period of 30 days for Romania.

V. Conclusion

In this paper, an auto-encoder forecasting model was proposed to predict the spread of COVID-19 in multiple European countries. The results of this research can help authorities to plan in advance prevention measures and allocate resources efficiently, such that the impact of the new disease is minimal in their country. The proposed method can be used to predict the spread of a pandemic, like COVID-19. It is worth mentioning that our proposed model uses only statistical data (number of

confirmed cases), without taking into account any other sort of data, like mitigation measures, travel impact and so on.

REFERENCES

[1] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet Infectious Diseases*, vol. 20, no. 5, pp. 533–534, May 2020. [Online]. Available: https://doi.org/10.1016/s1473-3099(20)30120-1

[2] A. Forna, P. Nouvellet, I. Dorigatti, and C. A. Donnelly, "Case Fatality Ratio Estimates for the 2013–2016 West African Ebola Epidemic: Application of Boosted Regression Trees for Imputation," *Clinical Infectious Diseases*, vol. 70, no. 12, pp. 2476–2483, 07 2019. [Online]. Available: https://doi.org/10.1093/cid/ciz678

[3] Z. Ceylan, "Estimation of COVID-19 prevalence in italy, spain, and france," *Science of The Total Environment*, vol. 729, p. 138817, Aug. 2020. [Online]. Available: https://doi.org/10.1016/j.scitotenv.2020.138817

[4] S. Flaxman, S. Mishra, A. Gandy, H. Unwin, H. Coupland, T. Mellan *et al.*, "Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in 11 european countries," 2020. [Online]. Available: http://spiral.imperial.ac.uk/handle/10044/1/77731

[5] S. Zhao, S. S. Musa, Q. Lin, J. Ran, G. Yang, W. Wang, Y. Lou, L. Yang, D. Gao, D. He, and M. H. Wang, "Estimating the unreported number of novel coronavirus (2019-ncov) cases in china in the first half of january 2020: A data-driven modelling analysis of the early outbreak," *Journal of Clinical Medicine*, vol. 9, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2077-0383/9/2/388

[6] B. Tang, X. Wang, Q. Li, N. L. Bragazzi, S. Tang, Y. Xiao, and J. Wu, "Estimation of the transmission risk of the 2019-ncov and its implication for public health interventions," *Journal of Clinical Medicine*, vol. 9, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2077-0383/9/2/462

[7] L. Peng, W. Yang, D. Zhang, C. Zhuge, and L. Hong, "Epidemic analysis of covid-19 in china by dynamical modeling," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/early/2020/02/18/2020.02.16.20023465

[8] B. Chen, M. Shi, X. Ni, L. Ruan, H. Jiang, H. Yao, M. Wang, Z. Song, Q. Zhou, and T. Ge, "Visual data analysis and simulation prediction for covid-19," *arXiv preprint arXiv:2002.07096*, 2020.

[9] C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos, "Data-based analysis, modelling and forecasting of the covid-19 outbreak," *PLOS ONE*, vol. 15, no. 3, pp. 1–21, 03 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0230405

[10] D. Fanelli and F. Piazza, "Analysis and forecast of COVID-19 spreading in china, italy and france," *Chaos, Solitons and Fractals*, vol. 134, p. 109761, May 2020. [Online]. Available: https://doi.org/10.1016/j.chaos.2020.109761

[11] Q. Liu, Z. Li, Y. Ji, L. Martinez, Z. U. Haq, A. Javaid, W. Lu, and J. Wang, "Forecasting the seasonality and trend of pulmonary tuberculosis in jiangsu province of china using advanced statistical time-series analyses," *Infection and Drug Resistance*, vol. Volume 12, pp. 2311–2322, Jul. 2019. [Online]. Available: https://doi.org/10.2147/idr.s207809

[12] R. Dandekar and G. Barbastathis, "Neural network aided quarantine control model estimation of global covid-19 spread," *arXiv preprint arXiv:2004.02752*, 2020.

[13] T. Zeng, Y. Zhang, Z. Li, X. Liu, and B. Qiu, "Predictions of 2019-ncov transmission ending via comprehensive methods," *arXiv preprint arXiv:2002.04945*, 2020.

[14] A. Tomar and N. Gupta, "Prediction for the spread of covid-19 in india and effectiveness of preventive measures," *Science of The Total Environment*, vol. 728, p. 138762, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0048969720322798

[15] R. Pal, A. A. Sekh, S. Kar, and D. K. Prasad, "Neural network based country wise risk prediction of covid-19," *arXiv preprint arXiv:2004.00959*, 2020.

[16] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin *et al.*, "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in china under public health interventions," *Journal of Thoracic Disease*, vol. 12, no. 3, pp. 165–174, Mar. 2020. [Online]. Available: https://doi.org/10.21037/jtd.2020.02.64

[17] Z. Hu, Q. Ge, L. Jin, and M. Xiong, "Artificial intelligence forecasting of covid-19 in china," *arXiv preprint arXiv:2002.07112*, 2020.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[19] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 03 1964. [Online]. Available: https://doi.org/10.1214/aoms/1177703732

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] U. Şahin and T. Şahin, "Forecasting the cumulative number of confirmed cases of covid-19 in italy, uk and usa using fractional nonlinear grey bernoulli model," *Chaos, Solitons & Fractals*, vol. 138, p. 109948, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960077920303477

[22] M. A. A. Al-qaness, A. A. Ewees, H. Fan, and M. Abd El Aziz, "Optimization method for forecasting confirmed cases of covid-19 in china," *Journal of Clinical Medicine*, vol. 9, no. 3, 2020. [Online]. Available: https://www.mdpi.com/2077-0383/9/3/674

[23] H. H. Elmousalami and A. E. Hassanien, "Day level forecasting for coronavirus disease (covid-19) spread: analysis, modeling and recommendations," *arXiv preprint arXiv:2003.07778*, 2020.