Proposing a Novel Artificial Neural Network Based Methodology for Forecasting Risk of Covid-19 Pandemic

Taha Osama Binhuraib[1[0000-0002-0578-6968]], Gül T. Temur[2[0000-0003-3853-0974]]

[1,2] Bahcesehir University, Yildiz Mah. Ciragan Cad., 34349 Besiktas/Istanbul Turkey
tahaosamaa.binhuraib@bahcesehir.edu.tr
gul.temur@eng.bau.edu.tr

Abstract. The corona virus formally known as Covid-19 has taken the world by storm. In this article we aim to analyze how harnessing the prowess of different computational methods –Machine Learning in particular; thus, helping policy makers take effective decisions by utilizing efficient approaches. Firstly, popularly known approaches (linear regression and logistic growth) are tried for existing data but it is figured out that accuracy rates are not satisfactory enough. Therefore, secondly, an artificial neural network (ANN) based methodology that consists of input data in different characters is proposed. It is figured out that the test results are more accurate with an R-squared score of (0.81) on an unseen data set, and handling the defined forecasting problem by using this multi-faceted data set is beneficial for policy makers, doctors and/or health managers that goal to have foresight on target groups at higher risk.

Keywords: Artificial Neural Network, Covid-19, Forecasting, Pandemic

1 Introduction

In the related literature, differential equation models such as Susceptible, Infected, Recovered (SIR) model have been traditionally applied for infectious diseases [1]. In countries, where less governmental intervention was applied, the accuracy of SIR models showcased relative accuracy [2]. Besides that, researchers mostly have preferred to consider past data on number of cases or date of the cases only [3]. However, it is explicit that human behaviors, sociological events and the many unknown variables have impact on cases and their risks. Therefore, this study proposes different machine learning techniques and analyzes their relative performances in predicting the number of cases. Firstly, simple regression models are considered. Then a single layer artificial neural network (ANN) that has multi-dimensional input data is postulated.

In the light of these explanations, the study is designed as follows: Section 2 clarifies the relevant literature on forecasting the number of cases. Section 3 explains the traditional computational methods used. Section 4 describes the proposed ANN methodology. Section 5 criticizes the results.

2 Literature Review on Forecasting Covid-19 Cases

In the literature related to mathematical epidemiology, the most widely used models for predicting the spread of a given disease are as the following [4]:

a. The SIR Model
b. The SIS Model
c. The SIER Model

The models aforementioned are considered to be mathematical models that don't utilize machine learning algorithms and are mostly solved using ordinary differential equations.

As the nature of viruses are dependent on an array of factors. Machine learning algorithms have been utilized in order to better predict spread viruses. In Table 1, we mention the most notable algorithms used to solve different problems:

Table 1. The Most Notable Algorithms

AUTHORS	ML ALGORITHM	OUTBREAK
[5]	Neural Network	H1N1 flu
[6]	Random Forest	Swine fever
[7]	Bayesian Network	Dengue/Aedes
[8]	Random Forest	Influenza

To our knowledge, there is a lack of studies considering extra parameters besides number of cases or date of the cases. In order to fill this gap, this study utilizes a comprehensive input data including day, population, population density, pollution, tourist number, number of universities and BCG vaccine regulation existence. The contribution of the study is originated from consideration of multi-faceted factors simultaneously.

3 Application of Traditional Computational Models

As a first step, forecasting procedure is conducted using conventional methods to test whether a more advanced method is needed. For this reason, linear and polynomial regressions, and logistic growth are chosen. The time series data was collected from the website Worldometers. https://www.worldometers.info/coronavirus/

Linear Regression: It is probably the first algorithm that comes to mind when dealing with time series data. Numerous predictions have been made using this approach, unfortunately due to the highly exponential and complex nature of viruses, the predictions made were not as accurate as intended. Linear regression simply aims to find a relation between two variables. Considering the simplicity of how this algorithm, it is without a doubt powerful tool that can give some useful insight to the data at hand. Bellow we share the results of linear regression using the Scikit learn library.

1 **Mean Absolute Error**	14539
R-Squared	0.944

Although The R-Squared is relatively high, but since there is a high correlation between the days passed and the cumulative number of cases, a high R-squared score is expected.

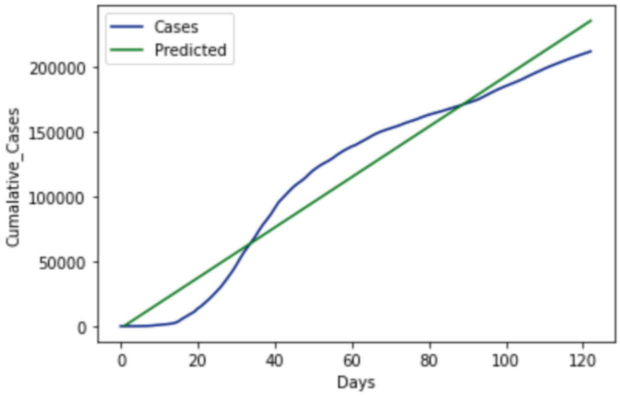

Fig. 1. Linear Regression(predicted) plotted against real data

Polynomial Regression: modeling the relationship between an x independent variable and a y dependent variable as an nth degree polynomial with respect to x. The main goal of polynomial regression is to find a non-linear relationship between the independent variable x and the dependent variable y.

The general formula for an nth degree polynomial regression model is as the following:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon \qquad \text{Eq. (1)}$$

The Sklearn python library was used in order to optimize for the equation given above. Which aims to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation:

$$||X_w - y||_2^2 \qquad \text{Eq. (2)}$$

Using the data gathered since the first case appeared in Turkey until the 123rd day. The results of the 3rd and 7th degree polynomials is as the following:

2nd degree Root Mean Squared Error	9934.2
3rd degree Root Mean Squared Error	9683.2
4th degree Root Mean Squared Error	4964.74
5th degree Root Mean Squared Error	2665.44
6th degree Root Mean Squared Error	2376.01
7th degree Root Mean Squared Error	1453.312
8th degree Root Mean Squared Error	1254.883
10th degree Root Mean Squared Error	3654.7

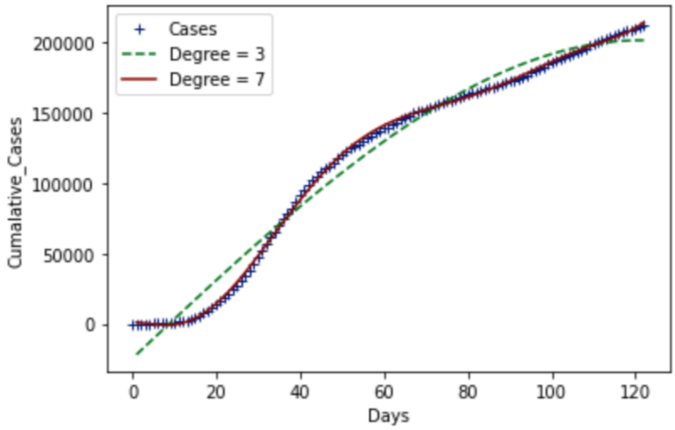

Fig. 2. Polynomial Regression of 3rd & 7th order plotted against the real data

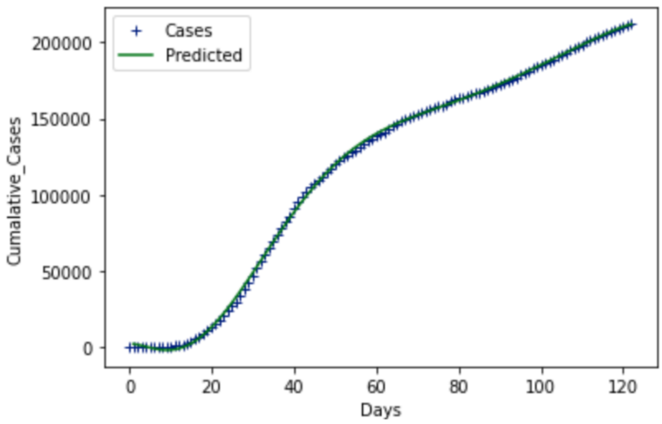

Fig. 3. Polynomial Regression of 8th order plotted against the real data

Logistic Growth: Upon studying the nature in which viruses spread[9], it is observed that the rate increases "exponentially" until it reaches a point of maximum growth rate, which is called the inflection point. The rate then starts decreasing until the number of infected reaches the total number of susceptible people "*L*".

$P(t)$ = Logistic function of *t; t* is the nth day after the first case

$$P(t) = \frac{L}{1 + Ae^{-bt}} \qquad \text{Eq. (3)}$$

The description above resembles that of a logistic growth model. Many predictions have been made using the logistic growth model, achieving higher accuracy rates and with lines that better fit the data.

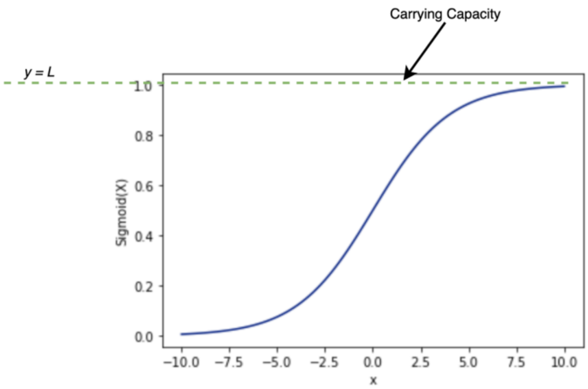

Fig. 4. Logistic growth function

The results for the above model were obtained using the SciPy scientific library which uses a Nonlinear Least Squares optimization technique, which aims to minimize the following:

$$S = \sum_{i=1}^{m} r_i^2 \qquad \text{Eq. (4)}$$

The minimum value is achieved when the gradient is equal to zero:

$$\frac{\partial S}{\partial \beta_j} = 2\sum_i \frac{r_i \partial r_i}{\partial \beta_j} = 0 \quad j = (1, \ldots, n) \qquad \text{Eq. (5)}$$

$$\beta = (\beta_1, \beta_2, \ldots, \beta_n)$$

Using an iterative method is then used to optimize for the vector β parameters of size n.

The Mean absolute error	9738.352

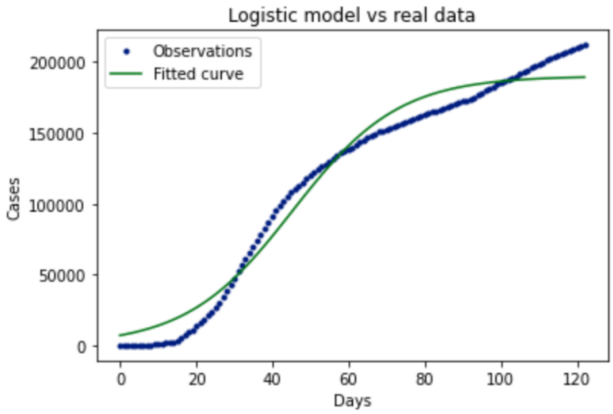

Fig. 5. Logistic growth fit against observations

4 Design of a Novel Artificial Neural Network Approach

In the previous section, it is figured out that traditional methods are not strong enough to give satisfactory results. Therefore, as an alternative method, ANNs is utilized for design of a novel approach for forecasting the number of cases. Inspired by biological neural models and highly resemblant of how animals learn by experience, ANNs aim to learn and extract patterns from large amounts of data[10] . Today ANNs are considered to be the most powerful tool in machine learning. The main advantage in using this model is that it gives data scientists the option to be creative in determining the feature space, and using a neural network model might provide useful insight into the data and help governments determine the variables that affect the spread of the disease, giving them the opportunity to take policy decisions accordingly. Although simpler methods such as a Pearson coefficient could be used in determining variables that negatively or positively affect the number of cases. We are aiming for a more robust method that in essence could help us find hidden patterns in the data.

Considering the success of these models in prediction, we aim to develop a simple 1 hidden layer neural network in order to predict the number of Covid-19 cases.

Data Set Design: The unpredictable nature of the corona virus with its dependence on an array of variables has rendered traditional methods incapable of accurately predicting unseen data. The majority of models are most likely to overfit and thus generalizations of the trained models are not possible. This has motivated us to try and develop a neural network that will encompass different variables such as socio-economic, weather, vaccine regulations, density, and population. The feature for the network is shown in Fig. 6.

INPUT LAYER:
$x1: Day$
$x2: Population$
$x3: Population\ density$
$x4: Pollution$
$x5: Tourists$
$x6: Number\ of\ universitie$
$x7: BCG\ vaccine\ regulatio$

Fig. 6. The Feature vector for the proposed model

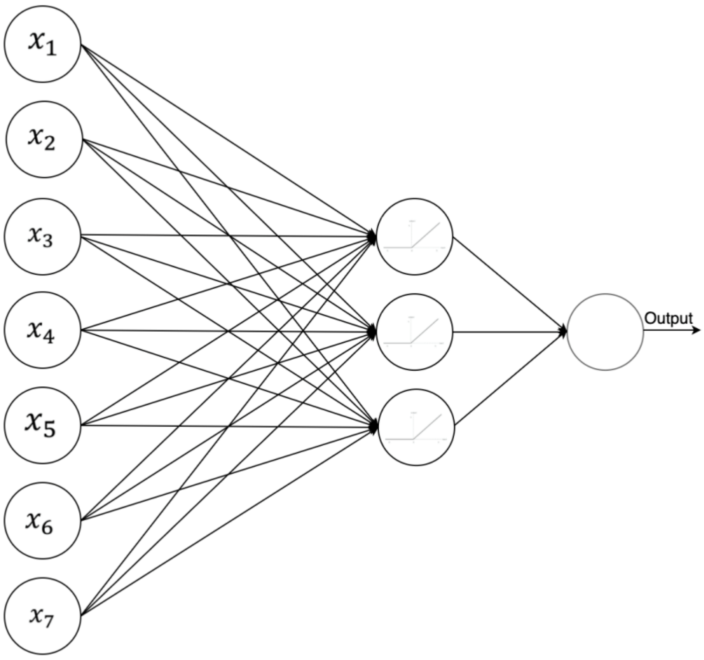

Fig. 7. Proposed Neural Network Architecture

$$Actication\ function = f(x) = \max(0, x) \qquad \text{Eq.(6)}$$

The Sklearn library was used in order to generate the neural network. The solver 'adam' refers to a stochastic gradient-based optimizer proposed by Kingma, Diederik, and Jimmy Ba [11].

The explanations for each data are given below:

Day: nth day after the first case in given city.

Population: population of city. This input feature was normalized.

Population density: Density (per km²)

Pollution: Fine particulate matter (PM2.5) is an air pollutant that can penetrate into the lung and impair lung function[12]. The data was collected from a Real-time Air Quality Index website

Tourists: Number of tourist in millions per year.

Number of universities: number of universities in given city

BCG vaccine regulation: 1: The country currently has universal BCG vaccination program. 2: The country used to recommend BCG vaccination for everyone, but currently does not. 3: The country never had universal BCG vaccination programs. One-hot encoding could be used in future implementations [13] .

Number of cases: number of cases in given day.

Due to a lack of data only 200 input points will be used in order to determine the soundness of the method.

Due to the limited space, only a part of data is shared at Table 2.

Table 2. Example of data used.

city_name	day	Population	density	pollution pm2.5	Tourist in millions/yr	No of universities	BCG vaccine	No_ofCases
New York	32	19453561	159	44	66	240	3	83506
New Jersey	1	8882190	467	30	101	47	3	4
Istanbul	27	15190000	2523	70	6	58	1	22171
London	16	8908081	4543	52	30	48	2	1588
London	17	8908081	4543	52	30	48	2	1965
London	18	8908081	4543	52	30	48	2	2189
New Jersey	6	8882190	467	30	101	47	3	23
New Jersey	7	8882190	467	30	101	47	3	29

The neural network converged after 28377 with a mean absolute error of 15746.21.

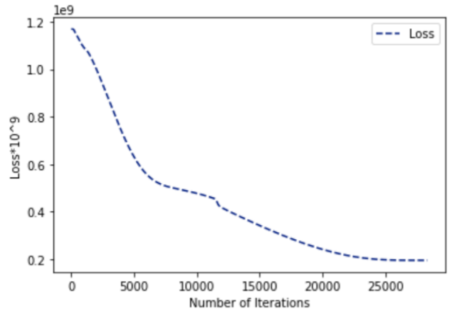

Fig. 6. The number of iterations plotted against the loss.

As seen from the plot above it could be said that the neural network has converged. Although the mean absolute error of this model is worse than that of the polynomial fit, we believe that the generalization of this model is possible, if a larger data set is provided.

5 Discussion

From a technical stand point, the clear disadvantage in this model and the type of data provided is the low variance between the input point within the same city—the population, density, vaccine regulations and socio economic factors do not change in a time series problem. That is why it is important to collect data from as many cities as possible. We must concede that the algorithm will most likely give the highest weight to the "*day*" variable. That is why a larger feature space could offset that bias and help us determine variables that significantly play a role in the spread of the virus.

References

1. LJS Allen, F Brauer, P Van den Driessche, J Wu – (2008) mathematical epidemiology. pp. 45-52
2. Sina F. Ardabili , Amir Mosavi , Pedram Ghamisi , Filip Ferdinand , Annamaria R. Varkonyi-Koczy , Uwe Reuter , Timon Rabczuk , Peter M. Atkinson(2020). COVID-19 Outbreak Prediction with Machine Learning
3. Tao Zhou, Quanhui Liu, Zimo Yang, Jingyi Liao, Kexin Yang, Wei Bai, Xin Lu, Wei Zhang. (2020)
   Preliminary prediction of the basic reproduction number of the Wuhan novel coronavirus 2019-nCoV
   https://onlinelibrary.wiley.com/doi/full/10.1111/jebm.12376
4. LJS Allen, F Brauer, P Van den Driessche, J Wu – (2008) mathematical epidemiology. pp. 19-52
5. Koike, F.; Morimoto, N. Supervised forecasting of the range expansion of novel non-indigenous organisms: Alien pest organisms and the 2009 H1N1 flu pandemic. Global Ecol. Biogeogr. 2018, 27, 991-

   1000, doi:10.1111/geb.12754.
6. Liang, R.; Lu, Y.; Qu, X.; Su, Q.; Li, C.; Xia, S.; Liu, Y.; Zhang, Q.; Cao, X.; Chen, Q., et al. Prediction for global African swine fever outbreaks based

on a combination of random forest algorithms and meteorological data. *Transboundary Emer. Dis.* 2020, *67*,935-946, doi:10.1111/tbed.13424.
7. Raja, D.B.; Mallol, R.; Ting, C.Y.; Kamaludin, F.; Ahmad, R.; Ismail, S.; Jayaraj, V.J.; Sundram, B.M. Artificial intelligence model as predictor for dengue outbreaks. *Malays. J. Public Health Med.* 2019, *19*, 103-108.
8. Tapak, L.; Hamidi, O.; Fathian, M.; Karami, M. Comparative evaluation of time series models for predicting influenza outbreaks: Application of influenza-like illness data from sentinel sites of healthcare centers in Iran. *BMC Res. Notes* 2019, *12*, doi:10.1186/s13104-019-4393-y.
9. Milan Batista(2020)
Estimation of the final size of the second phase of coronavirus epidemic by the logistic model
https://www.medrxiv.org/content/10.1101/2020.03.11.20024901v1.full.pdf
10. D Graupe – (2013)
Principles of artificial neural networks. pp. 1-9.
11. Diederik P. Kingma, Jimmy Ba (2015)
Adam: A Method for Stochastic Optimization
12. Internet source
https://www.health.ny.gov/environmental/indoors/air/pmq_a.htm
13. Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-237426