RESEARCH ARTICLE

Machine Learning Model for Predicting Number of COVID-19 Cases in Countries with Low Number of Tests

Samy Hashim1, Sally Farooq1, Eleni Syriopoulos1, Kai de la Lande Cremer1, Alexander Vogt1, Nol de Jong1, Victor L. Aguado1, Mihai Popescu1, Ashraf K. Mohamed1 and Muhamed Amin1,*

1Department of Sciences, University College Groningen, Hoendiepskade 23/24 9718 BG Groningen, Netherlands

Abstract:

Background:
The COVID-19 pandemic has presented a series of new challenges to governments and healthcare systems. Testing is one important method for monitoring and controlling the spread of COVID-19. Yet with a serious discrepancy in the resources available between rich and poor countries, not every country is able to employ widespread testing.

Methods and Objective:
Here, we have developed machine learning models for predicting the prevalence of COVID-19 cases in a country based on multilinear regression and neural network models. The models are trained on data from US states and tested against the reported infections in European countries. The model is based on four features: Number of tests, Population Percentage, Urban Population, and Gini index.

Results:
The population and the number of tests have the strongest correlation with the number of infections. The model was then tested on data from European countries for which the correlation coefficient between the actual and predicted cases $R^2$ was found to be 0.88 in the multi-linear regression and 0.91 for the neural network model.

Conclusion:
The model predicts that the actual prevalence of COVID-19 infection in countries where the number of tests is less than 10% of their populations is at least 26 times greater than the reported numbers.

Keywords: Machine learning, Model, COVID-19 cases, Healthcare systems, Testing, RNA viruses.

1. INTRODUCTION

The SARS-CoV-2 or COVID-19 outbreak was declared a global health emergency on 30th January, 2020, by the WHO. COVID-19 is a member of the coronavirus family enveloped positive sense single-stranded RNA viruses. It is thought that COVID-19 transitioned from animal to human hosts in the Huanan seafood market in Wuhan in the province of Hubei, China [1]. The virus spread rapidly initially within China and then worldwide. COVID-19 was declared a pandemic on 11th March, 2020, by the World Health Organization. As of April 25th, 2021, there have been almost 100 million confirmed cases worldwide. Yet PCR (polymerase chain reaction), which can detect the genetic material of the virus, is the most accurate technique for identifying the COVID-19 infection [2]. COVID-19 has exposed several inequalities. In the scramble to obtain medical resources, poorer countries have been left behind. Governments of low- and middle-income countries have struggled to provide sufficient funds to obtain medical resources, such as COVID-19 tests [3]. Furthermore, more geopolitically powerful countries have been accused of hoarding supplies leaving poorer countries unable to access sufficient tests [4]. With a disparity in the number of COVID-19 tests available, we aim to provide a prediction model based on machine learning that mitigates the reliance on clinical tests.

Machine learning has been utilized in contact tracing as a diagnostic and prognostic tool in vaccine and treatment development as a method to forecast and predict COVID-19 cases and deaths [5 - 11]. It has the potential to reduce the strain on healthcare systems that have been heavily burdened...
by the COVID-19 pandemic. For example, machine learning has been used to predict a positive COVID-19 infection in a PCR test [12]. The prediction is based on 8 binary features, including age, sex, contact with individuals known to have had COVID-19, and the appearance of five clinical symptoms. In addition, Sun et al. developed a model to predict the severity of a COVID-19 infection [13]. Furthermore, the model has been utilized to predict the prevalence of COVID-19 patients between one and six days in advance in 10 Brazilian states [14].

In this work, we have built multilinear regression and neural network models to predict the number of COVID-19 cases as of 15/03/2021. The models have been trained on the US States data and tested against the number of infections in the European countries. Then, both the models have been used to predict COVID-19 infection cases in countries with a low number of tests. The model was based on four features: the number of tests, population, urban population, and the Gini index. The model suggests that the actual number of infections is at least 10 times higher than the reported number of infections.

Uncertainties from different sources are not considered in this study; first, the ML model parameter uncertainty, which requires different techniques to be placed, such as Bayesian Neural Networks (https://arxiv.org/abs/2107.03342). This uncertainty is not considered since the DNN used for this study does not deliver certainty estimates or suffers from over- or under-confidence. Second, uncertainty data sources, since the data used in the analysis performed provide neither uncertainty in the PCR tests nor the estimation of the number of populations. Additionally, the PCR test uncertainty used in the US data, used in training, was different from the corresponding uncertainty for the tests used in other countries, and thus was used for the inference. Moreover, different COVID-19 variants significantly change the uncertainty rates, which can be a topic of future studies.

2. MATERIALS AND METHODS

The data were obtained from several official sources, such as from the World Bank World Development Indicators [16 - 19], government websites and publications [20 - 22], Worldometer [23], and from Our World in Data [14]. The data were extracted, standardized, and compiled into a single file. Although several features were considered, only four were included in the model owing to a lack of availability of data and low correlation with COVID-19 cases recorded. The four features used were Population, Tests, Gini Index, and % Urban Population. As the model first needed to be trained on the US States and then tested on European countries, data for all factors included would need to be available for both. This considerably limited the number of features that could be incorporated into the models. Several other factors were also considered, for example, median age and percentage of the population that always wears a face mask. However, median age was excluded from the model as it correlated poorly with the number of infections. The mask-wearing variable was excluded as the proportion of the populations that always wore masks was measured differently between the training and test countries and likely with all other countries for which the models were used to make predictions.

Data used to train the model covered the period from the beginning of the pandemic to February 2021. Later time period data were not used owing to the vast differences among countries not only in terms of the starting date and accessibility of vaccines but also the rate of vaccination. These discrepancies would make predictions for other countries inaccurate. The data used to test the model covered the period up until March 15th, 2021. A later date was considered for the test data than for the training data as most European countries started vaccination after the US.

Although the intention was originally to train the data on Indian states as well as the US states to allow for different models for the developing and developed countries, India was excluded owing to the high prevalence of the new B. 1.617 variant, which has increased transmissibility [24]. Although replacing India with Russia as an additional training data set was considered, the lack of data available made this unfeasible.

Some pre-processing steps had to be taken to clean the data before it could be used for the machine learning algorithm. First, the relevant features and information were extracted from the .csv file, where the data were stored; after that, all commas were removed from individual data points to make sure python could parse them correctly. The data were then normalized via a min-max-scaler, which places all data points between 0 and 1. For each data point in a feature, the MinMaxScaler deducts the smallest value in the feature and then divides this answer by the range, which is the difference between the original maximum and original minimum. The MinMaxScaler retains the original shape of the distribution, thus preserving the information embedded into the initial data set. However, it is important to note that this also means that the MinMaxScaler does not reduce the importance of outliers. Finally, the pre-processing procedure was completed by removing data samples that had missing values for some of their features. This is to make sure that all data can be used for training the model, as missing values can cause errors and unwanted variations within the procedure.

Two different types of machine learning algorithms were used for the analysis of the data multi-linear regression and a multi-layer perceptron artificial neural network (ANN). The multiple linear regression model was built using the Seikit-learn library [16]. The neural network code operated Keras architecture from the Tensorflow library [25] was used to construct the model. The ANN utilizes 1 output layer, 1 input layer, and 3 dense hidden layers (Fig. 1).

All dense layers use the Rectified Linear Unit (ReLU) as an activation function, which is defined as follows:

$$\begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$

The slope is always 0 for negative inputs and always 1 for positive inputs. ReLU was used as it is computationally less intensive and faster than most other activation functions, such as sigmoid and tanh.
The mean squared error (MSE) function is used to calculate the loss in the current iteration of the neural network. This function takes the absolute error of all points and calculates their mean. MAE is calculated via the following equation:

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i| ^ 2 = \frac{1}{n} \sum_{i=1}^{n} |e_i| ^ 2
\]

MSE was used because it is a commonly used metric and relatively robust to outliers suitable for the data used in this study.

The neural network contains a few hyperparameters that had to be set manually before the training. These hyperparameters are chosen by using a random grid search technique. The choice of the ReLU activation function, the number of hidden layers, and the number of nodes in each layer are examples of hyperparameters.

The features currently utilized in the models are as follows: Population, Tests, Gini index, and % Urban population. To observe their collinearity, the number of cases was plotted against these features for the US states (Fig. 2).

The population and the number of tests conducted both show a strong correlation with the prevalence of COVID-19 cases with R² values of 0.95 and 0.81, respectively (Fig. 3 (AB)), and p-values of zero. However, a much lower correlation was obtained for the Gini index and percentage urban population with R² values of 0.12 and 0.16, and p-values of 0.01 and 0.003, respectively. The features that are currently utilized in the models were selected based on their strong correlation with the number of cases. Other features, such as Median age, % of people wearing a facemask outside, and Number of lockdown days, were not used as low correlation was found between these features and the number of cases, and because the data were incomplete for a number of these features. Adding these features to the models would have resulted in a higher error.

3.2. Multilinear Regression

A multiple linear regression model was built and trained on the US States data according to the following equation:

\[
Y = Ax_1 + Bx_2 + Cx_3 + Dx_4 + K \rightarrow (1)
\]

Where, Y denotes the number of cases; A, B, C and D are the regression coefficients obtained from least square fitting; \(x_1, x_2, x_3\), and \(x_4\) are the independent variables (population, number of tests, Gini index, and % of the urban population, respectively), and K is the y-intercept.

The model shows a very strong correlation between the predicted and actual prevalence of COVID-19 cases for both the US States data (the training dataset) and the European data (the test dataset) (Fig. 3).

For the US data, the calculated slope is 1.00 with an intercept of zero and R² of 0.95. For the European data, the correlation coefficient R² is 0.88, and the slope and the intercepts are 1.49 and 12k, respectively, which indicates that the predicted prevalence of infections for the EU is generally higher than the reported. This could result from the differences in the behavior and commitment of the people toward the governmental rules in the US and the EU.

To understand the contribution of each feature to the prediction model, we report the estimated regression coefficients for each of the four features. The calculated coefficients are 0.87, 0.13, -0.01, and -0.03, for the population, number of tests, Gini index, and % urban population, respectively.
Fig. (2). COVID-19 cases vs. (A) Population; (B) Number of tests; (C) Gini index; (D) % Urban population; (E) Median age; (F) % of population that always wears a mask. Each point represents a state.

Fig. (3). To the left, the predictions vs. observed cases for US data (Slope: 1.00; Intercept: 0; $R^2$: 0.95). To the right, the predictions vs. observed cases for European data (Slope: 1.49; Intercept: 12K; $R^2$: 0.88).

The ‘population’ feature has a score close to one, and thus is the major contributor to the prediction model. The scores for the % Urban population and Gini are negative, which suggests that these features are not significant for the regression model.
3.3. Neural Networks

The neural network model is mainly considered to account for possible non-linearities in the Gini index and the percentage of urban populations. A fully connected Deep Neural Network (DNN) is trained and tested with US and EU datasets, respectively. The input layer of the network consists of 128 nodes and is followed by four hidden layers with 128 nodes and an output layer with a single node. The number of nodes of the output layer corresponds to the number of classes. Each layer has a random weight and bias initialization based on the normal distribution initializer, which is necessary to set the first set of numbers of weights and biases and thus initiate the training procedure. The ReLU function has become the default activation function for many types of neural networks because such models are easy to train, and often achieve good performance.

The DNN model is trained with an objective function (loss function) that must be minimized. The Mean Squared Error (MSE) is used as a loss function, and Stochastic Gradient Descent (SGD) optimizer is employed to find the best values for the DNN parameters by minimizing the loss function iteratively over the dataset. The number of iterations (epochs) is chosen to be 100 epochs. The network is trained using data from US states and tested using data from European countries using the same set of features as in the case of multilinear regression, namely Population, Tests, Gini, and the Percentage of urban population. The testing results are illustrated in Fig. (4), which quantify the correlation between the predicted number of infections and the number of infections recorded. The slopes are 0.95 and 0.80, the $R^2$ values are 0.95 and 0.91, and the mean absolute error is 0.03 and 0.06 for the US and EU datasets, respectively. These measurements suggest that the model fits the observed data by learning the relationships between the input variables.

3.4. Prediction of COVID-19 Cases

The reported infections and their corresponding predicted values (using linear regression and NN) are shown in Table 1. Furthermore, according to the training dataset, the US has performed 361 million tests, which is equal to approximately 110% of the US population. Thus, we reported the predicted number of cases for European and other countries with a low number of tests as these countries have had tests equal to 1.1 multiplied by their respective populations (columns 6-8 of Table 1). Although the number of tests for the EU countries is increased by 30%, the slopes of the linear regression and the NN models are increased only by 5% and 11%, respectively.

Table 1. The predicted number of COVID-19 cases for test countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>Actual No. of Tests</th>
<th>Reported No. of Infections</th>
<th>Predictions (Multilinear Regression)</th>
<th>Predictions (Neural Network)</th>
<th>Tests = 110% of Population</th>
<th>Predictions (Multilinear Regression)</th>
<th>Predictions (Neural Network)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albania</td>
<td>506676</td>
<td>117474</td>
<td>210493</td>
<td>171218</td>
<td>3163148</td>
<td>239234</td>
<td>358221</td>
</tr>
<tr>
<td>Austria</td>
<td>6033827</td>
<td>495464</td>
<td>713692</td>
<td>498778</td>
<td>9946648</td>
<td>770166</td>
<td>762327</td>
</tr>
<tr>
<td>Belgium</td>
<td>10110146</td>
<td>808283</td>
<td>906281</td>
<td>527839</td>
<td>12787487</td>
<td>957884</td>
<td>650472</td>
</tr>
<tr>
<td>Bosnia and Herzegovina</td>
<td>702920</td>
<td>142160</td>
<td>263250</td>
<td>223023</td>
<td>3593094</td>
<td>295191</td>
<td>392963</td>
</tr>
<tr>
<td>Croatia</td>
<td>1431342</td>
<td>251174</td>
<td>317183</td>
<td>273416</td>
<td>4496184</td>
<td>352885</td>
<td>437758</td>
</tr>
<tr>
<td>Cyprus</td>
<td>2563270</td>
<td>39651</td>
<td>56012</td>
<td>149468</td>
<td>1334864</td>
<td>44094</td>
<td>503521</td>
</tr>
</tbody>
</table>

Fig. (4). (A) the predictions vs. observed cases for US data (Slope: 0.95; Intercept: 0.0; $R^2$: 0.95); (B) the predictions vs. observed cases for European data (Slope: 1.57; Intercept: 45K; $R^2$: 0.81).
Country    | Actual No. of Tests | Reported No. of Infections | Predictions (Multilinear Regression) | Predictions (Neural Network) | Tests = 110% of Population | Predictions (Multilinear Regression) | Predictions (Neural Network) |
---|---|---|---|---|---|---|---|

Czechia   | 9665502 | 1402420 | 851078 | 546321 | 11795219 | 895219 | 805890 |

Denmark   | 20418687 | 220459 | 526270 | 420297 | 6387490 | 406445 | 662516 |

Estonia   | 1038888 | 86086 | 65321 | 82105 | 1459877 | 69132 | 373473 |

Finland   | 3596402 | 67334 | 375604 | 152357 | 6101438 | 372347 | 644442 |

France    | 57231533 | 4071662 | 5455817 | 3026689 | 71912361 | 5758992 | 406445 |

Germany   | 46319641 | 2578835 | 6752524 | 3338496 | 92369061 | 7397418 | 4629496 |

Greece    | 5856618 | 796133 | 497960 | 11425943 | 871592 | 707136 |

Hungary   | 4104415 | 524196 | 734001 | 437445 | 5479373 | 387925 | 662516 |

Ireland   | 3720861 | 3223142 | 4982964 | 2681373 | 66439272 | 5340019 | 372347 |

Italy     | 44623304 | 3223142 | 4982964 | 2681373 | 66439272 | 5340019 | 372347 |

Latvia    | 1670193 | 93959 | 98590 | 147075 | 2058635 | 103506 | 503588 |

Lithuania | 2218746 | 205644 | 194128 | 127608 | 2965009 | 204527 | 372347 |

Luxembourg| 1670193 | 93959 | 98590 | 147075 | 2058635 | 103506 | 503588 |

Moldova   | 771763 | 524196 | 734001 | 437445 | 5479373 | 387925 | 662516 |

Netherlands| 46623304 | 3223142 | 4982964 | 2681373 | 66439272 | 5340019 | 372347 |

Norway    | 4115415 | 80440 | 364594 | 176433 | 5996009 | 392747 | 609492 |

Poland    | 10668987 | 1917527 | 2962379 | 1413123 | 41599229 | 3346222 | 2136684 |

Portugal  | 8480932 | 814257 | 820324 | 557283 | 11193380 | 868501 | 764967 |

Romania  | 141568 | 183014 | 1316936 | 194200 | 18264429 | 1474642 | 164909 |

Switzerland| 16610468 | 1403000 | 2837478 | 3404783 | 297688125 | 23467745 | 3339198 |

UK        | 10305398 | 4258438 | 3494992 | 5988545 | 3816487 |

Ukraine   | 7328468 | 1467548 | 466872 | 341250 | 764967 |

The negative value reported for Luxembourg is a result of the very low population and the relatively high urban population and Gini index.

Using the same training dataset, we predicted the number of infections in selected countries where the number of tests is less than 10% of their populations (Table 2). The average number of the predicted infections is higher than that reported.
by 26 times for the linear regression model and 4 times for the NN. The discrepancy between the results from multilinear regression and NN models in Table 2 is due to the overfitting feature of the NN. The overfitting indicates that the generalization of the NN model is rather limited. This is due to the minimal dataset, i.e., 52 entries used for the training procedure, which is not enough for the NN model to avoid overfitting.

CONCLUSION

Both the multilinear regression and neural network models predicted the number of COVID-19 cases with a fair degree of accuracy on the European test data set. Considering Table 1, the number of cases predicted by the models was close to the number of cases reported for some countries, such as Italy, Poland and Slovakia. Yet, in most cases, the model predicted more cases than they were reported. The models were trained on data from the US country that tested extensively. Therefore, it seems that due to limited testing in most countries, the number of cases reported was a gross underestimation of the actual number of infections. This disparity was more pronounced in countries that were not testing extensively. The predicted number of infections for these countries was 26 times higher than the reported numbers on average. Therefore, the models can be effective tools for estimating the prevalence of COVID-19 infection in countries where sufficient testing is not available or where it is suspected that governments may not be entirely transparent about the number of COVID-19 infection.

LIST OF ABBREVIATIONS

MSE = Mean Squared Error
DNN = Deep Neural Network
SGD = Stochastic Gradient Descent

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The authors confirm that the data supporting the findings of this study are available within the manuscript.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

table2019.


