Evaluation of machine learning based COVID-19 forecasting models

Aljaž Sebastjan Ahtik

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
aa1353@student.uni-lj.si

In the recent years the problematic of COVID-19 has been an important topic as it has been putting strain on hospitals, the schooling system, and businesses. The purpose of this article is to evaluate machine learning methods for the task of forecasting the incidence of COVID-19. In our experiment we used different configurations of MLP, LSTM, and GRU models and found that GRU models outperform the other two with LSTMs performing the worst. We also found that simple feature sets outperform the complex ones.

***Index Terms*—COVID-19. machine learning, incidence forecasting, MLP, LSTM, GRU**

I. Introduction

In the recent years the problematic of COVID-19 has been an important topic, both in media and in the scientific world. This widespread pandemic has disrupted everyday life and, perhaps more importantly, put intense strain on hospitals, the schooling system, and businesses. These institutions would perhaps be better prepared to handle such a pandemic if the spikes of infections could be predicted. The purpose of this article is to evaluate machine learning methods for the task of forecasting the incidence of COVID-19. In the article we will first present some related work, then describe the ideas behind the machine learning models used in the paper, and lastly present the results using different model configurations and feature sets.

II. Related work

As COVID-19 has recently been an important issue, there have been many articles written about predicting the incidence of the disease in advance to help hospitals better prepare for an increase in COVID cases. Many of such articles use machine learning methods for this purpose.

The authors of [1] use a least-squares boosting classification algorithm to predict the two-week incidence rate based on the previous two-week incidence, date, and location. The possible classes of the algorithm are as follows: less than 200, between 200 and 1000, and above 1000. The model proposed in the article predicted the number of globally confirmed cases of COVID-19 with the accuracy of 98.45%. As the incidence is a numerical value most other article use a regression based approach.

Article [2] uses LSTM neural networks to predict the incidence of the disease in Canada. The model proposed in the article uses the number of cases, fatalities and recovered patients, with the date to predict the 2, 4, 6, 8, 10, 12 and 14th day incidence of the disease. Similarly, the article [7] attempts to predict the number of COVID-19 cases using LSTM models for USA and Canada, however it focuses more on the comparison of the performances of different LSTM variants. The article compares stacked LSTM, bi-directional LSTM, and convolutional LSTM models where the convolutional LSTM outperforms the rest. Different types of neural networks are compared in article [8], where the author evaluates the performance of simple RNN, LSTM, bi-directional LSTM, GRU, and VAE models. This article uses the number of confirmed and recovered cases from Italy, Spain, Italy, China, USA, and Australia. A more in depth analysis of the factors causing COVID-19 spread is presented in article [6]. This article uses medical, socioeconomic, and environmental factors to model the incidence rate using MLPs.

III. Methods

In this paper we will compare the MLP, simple RNN, LSTM, and GRU models for the task of COVID-19 incidence forecasting. The ideas behind these models will be described in the following sections.

A. MLP

The multi layer perceptron is a fully connected feed forward neural network [4]. The network consists of a input layer, an output layer, and any number of hidden layers with an arbitrary number of neurons (Figure 1).

B. LSTM

As per article [5] the problem with standard recurrent neural networks is that the error signals flowing backward in time tend to blow up or vanish. The authors of the aforementioned paper present a new RNN architecture called LSTM (long short-term memory) which can bridge time intervals longer than 1000 steps, even in the case of noisy sequences. This is achieved using three layer memory cells, consisting of a cell, input gate, output gate, and a forget gate, visualised in figure 2.

C. GRU

A gated recurrent unit architecture was proposed in article [3] as a simplified alternative to LSTMs. This model merges the input gate and the forget gate from the LSTM into the update gate [8]. Simplifying the architecture reduces the number of parameters and thus theoretically improves its performance. The comparison between LTSM and GRU models is shown in figure 3.

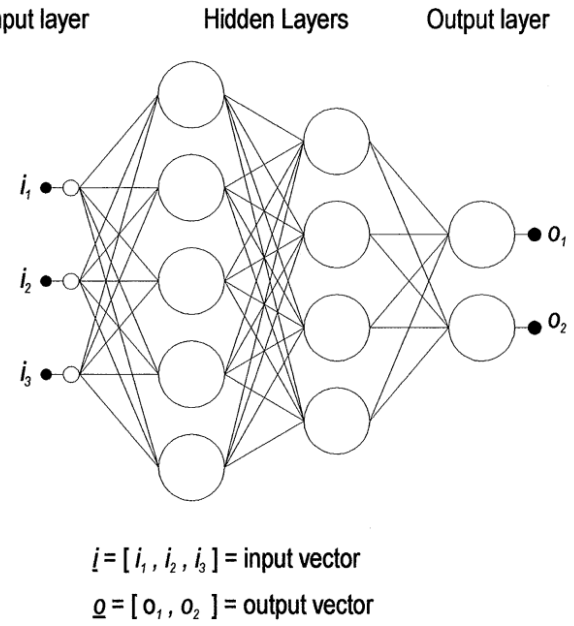

Fig. 1. A multilayer perceptron with two hidden layers [4].

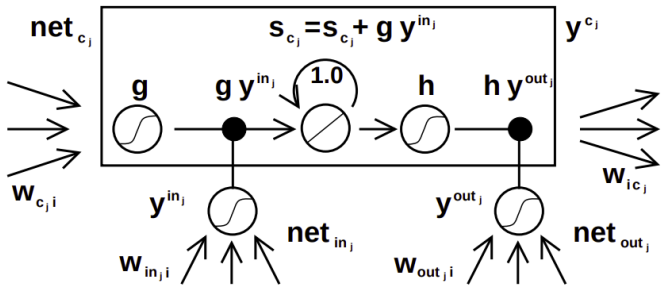

Fig. 2. Architecture of a memory cell and its gate units. The self-recurrent connection indicates feedback with a delay of one time step [5].

IV. Experiments

A. Data

The data used in this experiment was collected from the following sources:

- COVID-19 Sledilnik[1] (COVID data for Slovenia)
- METEO ARSO[2] (weather data for Slovenia)
- data.gv.at[3] (COVID data for Austria)
- HISTALP[4] (weather data for Austria)

Before use the data had to be processed. From the Slovenian weather dataset only stations with elevation lower than 500m were kept to show a better picture of the temperatures of populated areas and then the mean of minimal, maximal, and average daily temperatures over all the remaining stations was calculated. From the Slovenian COVID-19 dataset the

[1]https://covid-19.sledilnik.org/
[2]https://meteo.arso.gov.si/
[3]https://www.data.gv.at/daten/covid-19/
[4]https://dataset.api.hub.zamg.ac.at/

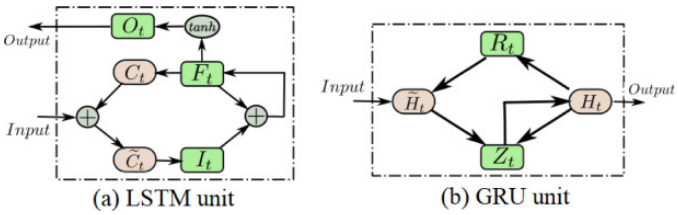

Fig. 3. Basic structure of LSTM and GRU models. I_t, F_t, and O_t represent the input, forget, and output gates of the LSTM, R_t and Z_t represent the reset and update gates of the GRU [8].

important features were extracted and the holiday, lockdown, and day of week features were added. The weather and COVID datasets were then merged by date.

The Austrian weather dataset contained hourly temperatures therefore the minimal, maximal, and average temperature was firstly calculated for all stations under 800m elevation in the four selected states. Then the means for the stations in each state was calculated. The Austrian COVID dataset splits the number of cases by age, which we do not take into account in this paper, therefore daily data was summed over all age-groups for each state. This data was then merged with hospital data and the same features as the Slovenian dataset were added. Then the COVID and weather data were merged.

The data for each state was taken into account separately to increase the number of data available. We only kept states with the population of 1000000 or greater, as they are more comparable to Slovenia. The day of week was transformed into cosine and sine values, the day was normalised, and the cumulative COVID statistics were normalised by dividing the values by state / country population. Rows with NaN values were removed. The fist ten rows for the compiled Slovenian dataset are shown in table I.

B. Models

In our experiment we used different configurations of MLPs, LSTMs, and GRUs. Due to relatively few data we limited our architecture search to relatively simple models with at most two layers. We trained the models with a 30 day input and a 7 day single-shot prediction using early stopping with the patience of 10 and a max epoch number of 100. To validate the model for early stopping we used a validation set of 30%. In the case of the MLP, the 30 day input was flattened and passed to the model in a single step.

We tested the models using 10-fold cross validation. We tested the model performance on data from the entire duration of the pandemic, because we wanted to check whether a similar model could be used to forecast the incidence for a possible future pandemic.

V. Results

In this section we will present the results of our experiments. The results of the 10-fold cross validation is shown in table II. We can see that MLP and GRU outperform LSTM, with the performance generally increasing with the amount of hidden neurons. There are also some outliers in the LSTM evaluation where the mean is 12.706 but the minimum is 0.012. The

TABLE I
THE FIRST TEN ROWS OF THE COMPILED AND NORMALISED DATASET FOR SLOVENIA.

ICU_beds	cases	day	deceased	dow_cos	dow_sin	holiday	lockdown	normal_beds	population	recovered	temp_avg	temp_max	temp_min	tests
3.80E-06	1.71E-04	-1.73E+00	0.00E+00	-9.01E-01	-4.34E-01	0.00E+00	1.00E+00	1.81E-05	2.11E+06	1.90E-06	-1.74E-01	2.30E-01	-7.61E-01	5.90E-04
7.60E-06	3.56E-04	-1.73E+00	0.00E+00	-2.23E-01	-9.75E-01	0.00E+00	1.00E+00	3.85E-05	2.11E+06	4.28E-06	-4.65E-01	-1.62E-01	-4.41E-01	1.00E-03
1.28E-05	5.57E-04	-1.72E+00	0.00E+00	6.23E-01	-7.82E-01	0.00E+00	1.00E+00	5.80E-05	2.11E+06	6.18E-06	-1.14E+00	-1.19E+00	-7.70E-01	1.35E-03
1.76E-05	7.69E-04	-1.72E+00	4.75E-07	1.00E+00	0.00E+00	0.00E+00	1.00E+00	7.98E-05	2.11E+06	1.09E-05	-1.31E+00	-1.44E+00	-1.28E+00	1.95E-03
2.28E-05	9.96E-04	-1.72E+00	9.50E-07	6.23E-01	7.82E-01	0.00E+00	1.00E+00	1.07E-04	2.11E+06	1.85E-05	-1.28E+00	-1.51E+00	-1.12E+00	2.54E-03
2.95E-05	1.23E-03	-1.71E+00	1.43E-06	-2.23E-01	9.75E-01	0.00E+00	1.00E+00	1.38E-04	2.11E+06	3.47E-05	-1.48E+00	-1.71E+00	-1.07E+00	3.10E-03
3.71E-05	1.47E-03	-1.71E+00	1.90E-06	-9.01E-01	4.34E-01	0.00E+00	1.00E+00	1.69E-04	2.11E+06	5.75E-05	-1.22E+00	-1.57E+00	-9.67E-01	3.61E-03
4.75E-05	1.71E-03	-1.70E+00	2.85E-06	-9.01E-01	-4.34E-01	0.00E+00	1.00E+00	2.07E-04	2.11E+06	7.84E-05	-5.89E-01	-5.04E-01	-5.59E-01	4.27E-03
5.84E-05	1.95E-03	-1.70E+00	3.80E-06	-2.23E-01	-9.75E-01	0.00E+00	1.00E+00	2.44E-04	2.11E+06	9.45E-05	-5.58E-01	-9.72E-02	-9.40E-01	4.74E-03
6.98E-05	2.19E-03	-1.70E+00	4.75E-06	6.23E-01	-7.82E-01	0.00E+00	1.00E+00	2.84E-04	2.11E+06	1.11E-04	-2.86E-01	-3.05E-02	-7.19E-01	5.03E-03

plot of mean absolute errors of the test folds for the 1 layer with 64 neurons models is visible in figure 4. We can see that every model has a spike at the second fold, which is at the data point of the biggest spike in cases in Slovenia. Similarly some smaller spikes can be seen at the fourth, sixth, and eighth folds at the cases spikes of the Austrian states.

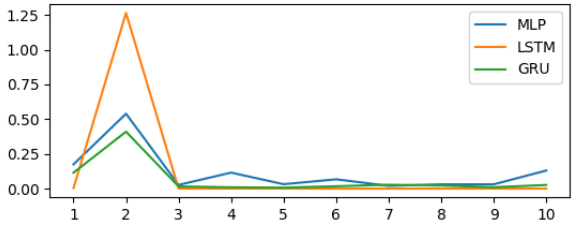

Fig. 4. A plot showing mean absolute errors for each fold in 10-fold cross validation with the errors for LSTM scaled down by a factor of 100.

We used the best models (MLP and GRU with one hidden layer with 64 neurons) and trained them using a bigger feature set. We also checked the models with doubled hidden layer neurons. The results are visible in table III.

Lastly, we trained a 1 hidden layer with 64 neurons GRU model for 1000 epochs using early stopping with patience of 100 on both feature sets (using a 0.6-0.3-0.1 train-validation-test sets) and found that using the smaller feature set outperforms the extended one with a 5.55E-5 MAE to 13.01E-5 MAE. The results of the tests are visualised in figure 5.

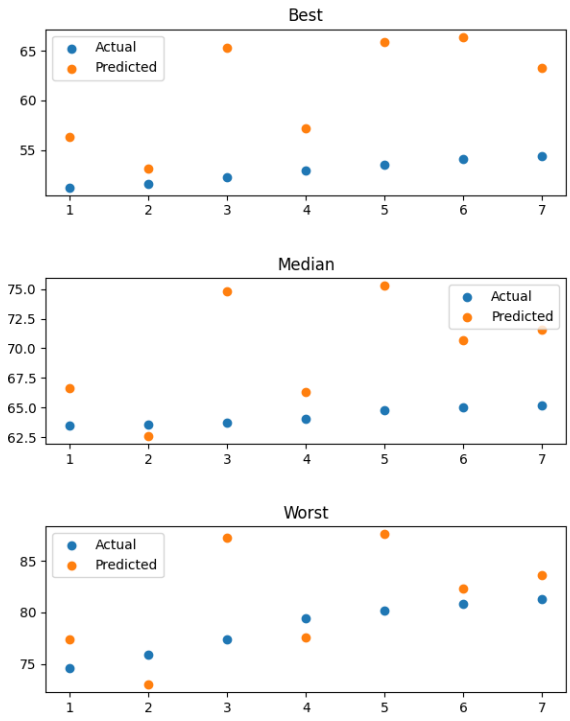

Fig. 5. Plots visualising the best, the median, and the worst predictions in the test set. The x-axis shows the number of days, and the y-axis shows the number of cases per 100000 population.

VI. CONCLUSION

In the article we evaluated MLP, LSTM, and GRU models for the task of COVID-19 incidence forecasting. We found that, when using relatively few data, simpler models perform better. From the noisy predictions of the final model we can also see that it is not particularly useful in short-term incidence forecasting. For future work we would recommend using more data, perhaps incorporating the statistics of past pandemic diseases.

REFERENCES

[1] Fatemeh Ahouz and Amin Golabpour. “Predicting the incidence of COVID-19 using data mining”. In: *BMC Public Health* 21.1 (June 2021), p. 1087.

[2] Vinay Kumar Reddy Chimmula and Lei Zhang. “Time series forecasting of COVID-19 transmission in Canada using LSTM networks”. In: *Chaos, Solitons & Fractals* 135 (2020), p. 109864. ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2020.109864. URL: https://www.sciencedirect.com/science/article/pii/S0960077920302642.

[3] Kyunghyun Cho et al. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014. DOI: 10.48550/ARXIV.1406.1078. URL: https://arxiv.org/abs/1406.1078.

[4] M.W Gardner and S.R Dorling. “Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences”. In: *Atmospheric*

TABLE II
THE AVERAGES AND MINIMUMS OF MEAN ABSOLUTE ERRORS OVER THE 10-FOLD CROSS VALIDATION FOR EACH TESTED MODEL CONFIGURATION USING THE FOLLOWING FEATURES: NUMBER OF CASES, DAY.

	MLP		LSTM		GRU	
	mean	best	mean	best	mean	best
L:1, H:16	0.104	0.007	0.193	0.019	0.212	0.011
L:1, H:32	0.261	0.025	0.327	0.016	0.137	0.009
L:1, H:64	0.117	0.020	12.706	0.012	0.066	0.007
L:2, H:16	0.244	0.022	0.399	0.024	0.271	0.011
L:2, H:32	0.159	0.027	0.250	0.014	0.257	0.019
L:2, H:64	0.226	0.012	3.255	0.016	0.219	0.005

TABLE III
THE AVERAGES AND MINIMUMS OF MEAN ABSOLUTE ERRORS OVER THE 10-FOLD CROSS VALIDATION FOR EACH TESTED MODEL CONFIGURATION USING THE FOLLOWING FEATURES: NUMBER OF CASES, DAY, DAY OF WEEK (SIN, COS), HOLIDAY, LOCKDOWN, AVERAGE TEMPERATURE.

	MLP		GRU	
	mean	min	mean	min
L:1, H:64	0.345	0.028	0.119	0.020
L:1, H:128	0.657	0.028	0.129	0.016

Environment 32.14 (1998), pp. 2627–2636. ISSN: 1352-2310. DOI: https://doi.org/10.1016/S1352-2310(97)00447-0. URL: https://www.sciencedirect.com/science/article/pii/S1352231097004470.

[5] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[6] Abolfazl Mollalo, Kiara M. Rivera, and Behzad Vahedi. "Artificial Neural Network Modeling of Novel Coronavirus (COVID-19) Incidence Rates across the Continental United States". In: *International Journal of Environmental Research and Public Health* 17.12 (June 2020), p. 4204. ISSN: 1660-4601. DOI: 10.3390/ijerph17124204. URL: http://dx.doi.org/10.3390/ijerph17124204.

[7] Sourabh Shastri et al. "Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study". In: *Chaos, Solitons & Fractals* 140 (2020), p. 110227. ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2020.110227. URL: https://www.sciencedirect.com/science/article/pii/S0960077920306238.

[8] Abdelhafid Zeroual et al. "Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study". In: *Chaos, Solitons & Fractals* 140 (2020), p. 110121. ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2020.110121. URL: https://www.sciencedirect.com/science/article/pii/S096007792030518X.