PHUONG DANG

220436263

P78 – DATA ANALYTICS AND DEEP LEARNING IN BETTER UNDERSTANDING COVID 19

PHUONG DANG

Abstract

Corona virus 2019 (covid-19) is an epidemic that has developed from Wuhan, China and has spread to 213 different countries around the world. According to the World Health Organization (WHO), on July of 2020, it was proclaimed that the covid-19 virus infected more than 10,509,505 people around the world, and this led to 559,694 deaths. Governments and hospitals have been struggling to deal with the number of confirmed cases each day and are looking at ways to make this easier. This will allow both the government and hospitals to prepare for affected people in terms of treatments and payment plans.

The virus is transmitted when two people are directly in contact with each other or when the infected person expels small droplets by sneezing or coughing. It is important to always wash your hands or to avoid other people who have caught the virus, to decrease the transmission rate of the virus. Is has been seen that transmission rates often occurs when areas are experience high wind speed or large population densities. Wind speed is the speed of the weather related to air movement from one area to the next area. Population densities can be defined as the concentration of individuals within a species in a specific geographic location. These two factors has been shown to contribute to 94% of how fast covid-19 transmits from one person to another which in turns increases daily cases.

Machine learning and deep learning algorithms have been previously been used in order to try and predict future daily confirmed cases. A lot of research paper proposed to apply and improve on existing deep learning algorithm such as the gated recurrent units (GRU) model or the Long Short-Term Memory (LSTM) model. The dataset that will be used to train and test our models will be from three different sources: the “European Centre for Disease Prevention and Control” website, “Kaggle - Population by Country” website and the “Air Quality Historical Data platform” website.

After extensive testing with hyperparameter tuning, the models were able to achieve results which are then used to compare whether adding in wind speed and population density will have a impact to the accuracy. For LSTM without the population density and wind speed factors data, it achieved a accuracy score of 0.4138 and for GRU it gave a accuracy of 0.5862. However, with the factors included the accuracy for both models increased to 0.6934. The research concluded that including wind speed and population densities does indeed play an important role to covid-19 transmission and helps with the accuracy of the forecasting of daily cases.

Contents

1. Introduction

Corona virus 2019 (covid-19) is an epidemic that was originated from Wuhan, China and has spread to 213 different countries around the world. According to the World Health Organization (WHO), on July of 2020, it was proclaimed that the covid-19 virus infected more than 10,509,505 people around the world, which led to 559,694 deaths. With these statistics, Covid-19 has become one of the worst pandemics that has been seen in the modern era (1,9). This pandemic continues to challenge the medical systems around the world. This includes sudden increases in demands for hospital beds and shortages of medical resources. Therefore, it is crucial to find ways to help make clinical decisions and decide the most effective way to distribute healthcare resources (2).

The virus has been known to remain in the air for multiple hours and it is known that wind speed can affect the spread and the direction of the virus while it travels through the environment (3). With this information, it proves that the wind speed is an important factor in terms on how fast the virus travels one from section to another. This could increase the speed that the virus spread and would affect the daily confirmed cases of each country. As the virus started spreading across the world, data shows that places with more people are more likely to have larger number of cases. This is due to the fact that social distancing becomes more difficult when there are a lot more people cramped into a small area. As shown in a study, even if population density is the only factor considered, it can still provide a high explanatory power in the variation of covid-19 transmission (4). Monitoring wind speed and population density would help hospitals prepare for beds and medical equipment each day.

It has been shown that daily confirmed covid-19 cases usually follow specific patterns. Based on these patterns, there has been different methods being thought up in order to find and forecast such infective diseases. Since the spread of this virus is exhibited as a non-linear nature, researchers have been trying to design specific non-linear systems to describe the transmission rate of the virus (5). Deep learning algorithms have been a proven method used in the past to analyze and predict different outbreak data patterns. This has helped governments plan early and reduce the number of people getting infected (6). Researchers have commonly suggested using Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), which are the two strongest candidates when it comes to evaluating and forecasting the number of confirmed cases of covid-19 (7).

This paper will first aim to review an overview of covid-19, the factors that can cause increase transmission rate, review deep learning algorithms specifically LSTM and GRU and establish key performance indicators to which we could expect the techniques to achieve. 31/12/2019 to 14/12/2020

Areas covered will be the research design and methodology to give more information about the models and dataset used. The paper will then discuss the artefact that was developed, and the approach taken to ensure the best accuracy was developed for each model. The paper will last outline the research questions to be answered, experiments conducted and the resulting metrics. This will then be concluded with a discussion, threat to validity and conclusion.

2. Literature Review

2.1 Covid-19

Covid-19 is a deadly widespread virus that has affected thousands of people around the world. Its origin was from Wuhan City, China in December 2019. The virus slowly spread undetected to every country and became a pandemic (1,8). It became one of the worst pandemics in modern era when it infected more than 10 million people across the world within 7 months (1). According to the WHO, covid-19 is defined as a collection of viruses that causes a range of symptoms in humans from the common cold to much more severe conditions. The WHO has released many guides on how to identify if an individual is infected by the virus, how to remain unaffected and what kind of precautions should be taken while going into the community. There are also guides on when to go to the hospital. During the early days, it was suggested that everybody should avoid unnecessary travels, distancing from infected people, wash hands regularly and if experiencing any symptoms to wear a mask. If a person showed any symptoms of illness, that person would be taken into hospital for treatment as soon as possible (9).

2.1.1 Covid-19 Transmissions

The virus is believed to be transmitted from human to human when two people are directly in contact with each other or when the infected person expels small droplets by sneezing or coughing (10). covid-19 can also be spread if an infected person has touched a surface or an edible item that is then touched or eaten by a non-infected person (9). It is important to always wash your hands or to avoid other people who have caught the virus, to decrease the transmission rate of the virus.

2.1.2 Covid-19 Incubation period

The incubation period could be defined as the time between when a person catches the virus and when the symptoms start to appear (10). According to WHO, the virus has an incubation period of 2-14 days and therefore it is important to stay isolated for 14 days (11). However, recent data has suggested that the incubation period can be as long as 30 days and therefore being in contact with somebody during that period can also lead to a healthy person catching the virus (12).

2.1.3 Covid-19 Symptoms

The main symptoms that are experienced by patients includes fever, dry cough and fatigue (13). At the early stages of contracting the virus, many patients have shown symptoms such as the following: headache, languidness and unstable walking. This is believed to be caused by non-specific manifestations caused by covid-19 (14). In addition to these symptoms, there can also be diarrhoea, hearing problems, loss of sense of smell and chest pains. Anybody who have these symptoms should be isolating to reduce the possibility of spreading the virus. The WHO released some measures to help avoid the infection such as covering the face with a mask, avoiding any physical contact and enforcing lockdown.

2.2 Factors

There has been a lot of literature that talks about different factors that could affect the spread of the covid-19 virus. However, the major factors that is being focused on in this review are population density and wind speed. This is because based on a researched conducted by Coskun and co, it was found that population density and wind intensity explained 94% of the variance in the virus spread (15). Another research by Cao and co, shows that Wuhan and New York were the hardest hit cities in the respective country due to having the highest population density in their city. This research also showed that Wind speed had the highest correlation when it comes to the virus spread in these two cities when comparing it to other factors (16). These two factors need to be included in models that tries to forecast positive cases due it being such an important factor on why covid-19 transmission can increase or decrease.

2.2.1 Wind Speed

The coronavirus can be found in the air for many hours after it has been expelled by the body, which implies that the transmission of the virus can be influenced by wind speed (17). Since the virus holds this trait, the wind speed is accountable for the spread of the virus by accelerating the traveling time from one place to another. Researchers has shown that there is a increase of 0.113 times amount of cases for every 1km/h rise in wind speed in Pakistan (18). The same conclusion was reported in China, when 1 unit of wind speed increases, there is an increase of 2.28 units of confirmed cases (19). When a experiment was conducted in Latin American and Caribbean countries, Wind speed showed a positive correlation with covid-19 infection rate. The regression analysis shows that when the speed of the wind is increased by 1km/h, the log count of covid-19 was 0.074 times in Punjab, 0.042 times in Sindh and 0.082 times in Khyber Pakhtunkhwa (20). Another paper shows that when wind speed is increased by 1% there is also an increase in 11.21% in confirmed cases in Africa (21). Based on these findings, it shows that wind speed is an important correlation factor when it comes to covid-19 transmission. It is important that wind speed is one of the factors considered when trying to predict future covid-19 cases.

2.2.2 Population Density

Population density is defined as the concentration amount of people within a specific geographic location. Population density is considered a factor because the more people there is in a proximity, the higher the chances of infecting a healthy person with the virus. Researchers did a research where they picked different parts of Turkey with different population numbers to test the correlation of population density and covid-19 transmission. After using the regression model to test the experiment with many cities in Turkey, it was concluded that the virus spread, and transmission rate increases as the population density increases (15). Another group of researchers researched the effects of population density on the spread of covid-19 in Algeria. They found that population density had a very strong correlation relationship with covid-19 transmission, which can explain 50.50% of the transmission rate (22). This research clearly proves that when trying to forecast daily covid-19 cases, population density will be an important factor to include.

2.3 Deep Learning

Deep Learning (DL) is a subfield of machine learning that is giving promising results in time series data analysis and forecasting. DL models can learn the dependencies and structures by finding its trends and seasonality in the data. Neural networks are one of the most used algorithms used to forecast disease's daily cases one day in advance by training the past 'k' days historical medical measurements (23). It is important to only take features of data that are highly dimensional, without errors and noise (24).

2.3.1 LSTM

LSTM is a more advance model based on the neural network algorithms. They were able to overcome some of the limitations of other algorithms by using the hidden layer units known as memory cells (25). These memory cells have the self-connections that store and control the network temporal state via three different gates called the input, output and forget gates. The function of the input gate and output gates are to be used to control the flow of how the memory cells inputs and outputs into the rest of the network. These gates allow or denies input values, which can be kept for an amount of time depending on the weights and input of the data. The forget gate tells the cell states which information to forget in the matrix.

One of the more famous use of LSTM during the covid-19 pandemic is when a group of Canadian researchers used the LSTM algorithm to predict the exact date of when the number of daily confirmed cases will hit zero in Canada. Even though they did not get the exact day, they were able to get close enough for people to take notice of deep learning and LSTM (5). Another experiment is the forecast of India's daily cases using the LSTM model. In the experiment they were able to produce a predictive model that was very close to the actual numbers of covid-19 (36). This was a prime example how a non-linear model like LSTM can be a very effective tool to forecast and predict the number of covid-19 daily cases. This information can then be used to help with medical decisions and resources distribution.

2.3.2 GRU

GRU is another advance model that is based on the neural network algorithm and a newer variation of LSTM. GRU aims to fix the problem LSTM had with the vanishing gradients by only having two gates instead of three. The two gates are the "update gate" which determines how much of the past information is passed onto the next cell and the "reset gate" determines how much of the past data is unnecessary and forgotten (26). GRU can only control the information inside the unit because it has no extra memory cells to maintain information.

A group of researchers did a comparison of different models that consisted of GRU as one of their preferred deep learning algorithms. Their aim was to predict the amount of covid-19 confirmed cases and see which algorithm performed the best. They performed different algorithms for 10 different countries and concluded that LSTM performed the best and GRU was a close second (27). In that experiment, GRU was the best performing model for China. Another experiment done in the UK found that the GRU model best validates the data with the lowest percentage error (37). This proves that both GRU and LSTM are the two strongest candidates to be used when performing such predictions.

2.4 Key performance metrics

Of key importance for covid-19 confirmed cases algorithm on time series data is to ensure the prediction model is as close to the real number as much as possible.

2.4.1 Error Rate

Error rate can be described by the number of errors that are made by the model. This refers to the frequency of errors occurred when testing the model. It can be calculated by dividing the total number of errors to the total number of data tested. As the error rate increases the reliability of the model decreases.

2.4.2 Mean Squared Error

Mean squared error (MSE) is a measurement of how close a regression line is to the true values. It does this by taking the distances from the predicted values to the true values. The smaller the mean squared error, the closer the model is to finding the line of best fit.

2.4.3 Accuracy

Accuracy is the ratio of how much of the prediction is matching the true values. This is calculated by taking the number of correct predictions over the total amount (28). This is to help researchers how much of proportion of the model was correctly identified. The higher the number the more reliable the model is.

3.Research Design and Methodology

Research design for this topic must cover 3 main areas, identifying which factors is to be trained with the confirmed cases, models used to forecast the daily cases and benchmarking performance of the prediction models. This will be completed looking to produce an algorithm capable of forecast the daily cases by combining the covid-19 confirmed cases and death dataset, population density dataset and wind speed dataset. The objective of the research is to find an optimized algorithm capable of forecasting covid-19 confirmed cases and using the models to find whether the wind speed factor and population density factor improves the accuracy of the model. The problem is solution drive by nature and will therefore require a quantitative approach to measure the effectiveness of the algorithm. A total of 3 different datasets from 4 different countries will be used in this model which include United States of America, United Kingdom, India and Indonesia.

3.1Dataset

Datasets are used to train and test the model in order to try and forecast the correct number of people that will be infected for the day.

The dataset is a combination of the "European Centre for Disease Prevention and Control" website, "Kaggle - Population by Country" website and the "Air Quality Historical Data platform" website.

The European Centre for Disease Prevention and Control has collected data from countries all over the world and has recorded the total number of deaths and the total number of confirmed cases for each day caused by covid-19. This dataset has been designed in order to see the confirmed cases and deaths of covid-19 in each country. This data was collected from 31/12/2019 to 14/12/2020.

The Kaggle dataset was designed in order to see the population density for each country. They show the land area of the country and the total population of the country. They then calculate the population density based on these two numbers. The data contains number from 2020 for each country of the world.

The third data is taken from the Air Quality Historical Data Platform. This platform is managed by the World Air Quality Index organization by working with many other companies such as WHO, World Meteorological Organization and many other organizations. WAQI's aim is to provide historical Air Quality Data to relevant institutions and organizations working in the area of environmental awareness. The data shows the measured Wind speed minimum, maximum, median and variance for each date that was collected in each country for each day.

The dataset contains 6 different features:

F1: The date of which the data was collected

F2: The amount of confirmed covid-19 cases for the day

F3: The amount of confirmed covid-19 deaths for the day

F4: The country or Territory that the data was collected from

F5: The population density of the country in 2020.

F6: The continent that the data was collected from.

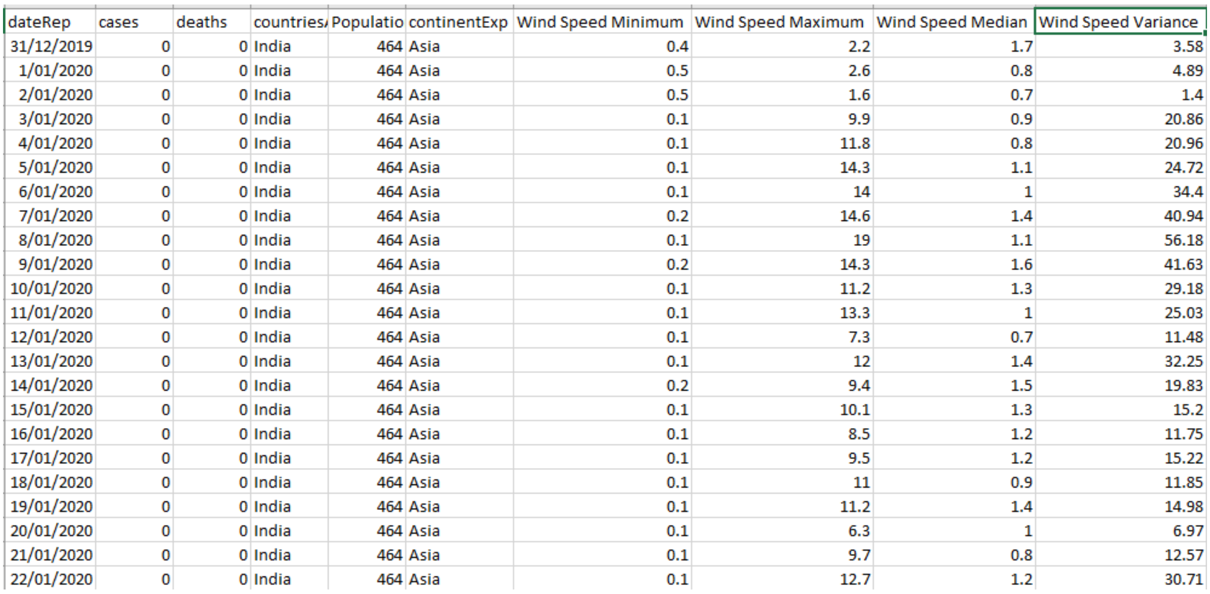
F7: The minimum wind speed collected in that country on that date

F8: The maximum wind speed collected in that country on that date

F9: The median wind speed collected in that country on that date

F10: The variance wind speed collected in that country on that date

An example of what the dataset looks like is shown below as well as a description of each of the column.



dateRep	cases	deaths	countries	Populatio	continentExp	Wind Speed Minimum	Wind Speed Maximum	Wind Speed Median	Wind Speed Variance
31/12/2019	0	0	India	464	Asia	0.4	2.2	1.7	3.58
1/01/2020	0	0	India	464	Asia	0.5	2.6	0.8	4.89
2/01/2020	0	0	India	464	Asia	0.5	1.6	0.7	1.4
3/01/2020	0	0	India	464	Asia	0.1	9.9	0.9	20.86
4/01/2020	0	0	India	464	Asia	0.1	11.8	0.8	20.96
5/01/2020	0	0	India	464	Asia	0.1	14.3	1.1	24.72
6/01/2020	0	0	India	464	Asia	0.1	14	1	34.4
7/01/2020	0	0	India	464	Asia	0.2	14.6	1.4	40.94
8/01/2020	0	0	India	464	Asia	0.1	19	1.1	56.18
9/01/2020	0	0	India	464	Asia	0.2	14.3	1.6	41.63
10/01/2020	0	0	India	464	Asia	0.1	11.2	1.3	29.18
11/01/2020	0	0	India	464	Asia	0.1	13.3	1	25.03
12/01/2020	0	0	India	464	Asia	0.1	7.3	0.7	11.48
13/01/2020	0	0	India	464	Asia	0.1	12	1.4	32.25
14/01/2020	0	0	India	464	Asia	0.2	9.4	1.5	19.83
15/01/2020	0	0	India	464	Asia	0.1	10.1	1.3	15.2
16/01/2020	0	0	India	464	Asia	0.1	8.5	1.2	11.75
17/01/2020	0	0	India	464	Asia	0.1	9.5	1.2	15.22
18/01/2020	0	0	India	464	Asia	0.1	11	0.9	11.85
19/01/2020	0	0	India	464	Asia	0.1	11.2	1.4	14.98
20/01/2020	0	0	India	464	Asia	0.1	6.3	1	6.97
21/01/2020	0	0	India	464	Asia	0.1	9.7	0.8	12.57
22/01/2020	0	0	India	464	Asia	0.1	12.7	1.2	30.71

Figure 1: Sample of the combined dataset that was taken from the three resources mentioned above

3.2 Model

The primary algorithm that is used when approaching this problem is using the Recurrent Neural Networks (RNN) model. The models uses a forward feeding network that allows nodes along a temporal sequence to be connected directly from a graph. RNNs feed the information forward from the previous cell in its internal memory to process variable length sequences of inputs.

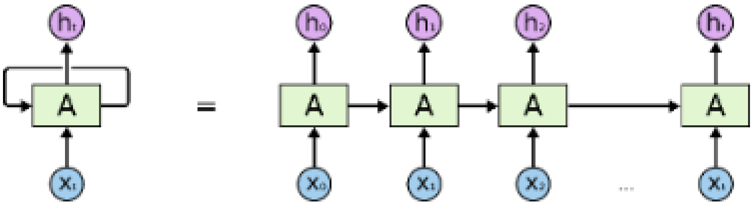


Figure 2: An example of how Recurrent neural network functions and how it feeds the information forward. (30)

The RNN models can be explained by logic and reasoning being carried forward in a stored state to build a picture for the proceeding information. As shown in Figure 1, if Hi requires the relevant information from Xi, RNNs are required pass information forward to feed the relevant Hi with necessary information. When making a decision, it considers the current input and also what it has learnt from the previous node. This makes RNN a really short-term model as it only considers information from the previous node and not all the nodes trained previously (31). Vanishing gradients is the main issue that occurs when designing at RNN model. Vanishing gradients occurs when the values of a gradient become too small and the model stops learning or it takes too much time to get a result. This is where models such as LSTM and GRU have been designed in order to overcome this problem by adding in time delays and feedback loops (32).

3.2.1 Long Short-Term Memory (LSTM)

LSTM is an updated extension of the traditional RNN by adding feedback connections. These connections allow the model to process multiple datapoints at the same time which allows the model to make comparisons to previous datapoints and go through training based on a sequence of the data (33). LSTM learns and manages the memory at each input by using memory cells and gate units. The LSTM cells consists of an forget gate, an input gate, an output gate and a memory cell. The forget gate is used to select useless information from previous nodes to forget. The input gate selects information from the current cell and shows the relevant information while filtering out the unnecessary information. The output gates have the final say on what information is transferred to the next cell (34). Since LSTM is one of the more advance versions of RNN, it will be used in order to run the dataset and forecasting daily covid-19 cases.

3.2.2 Gated Recurrent Units (GRU)

GRU is another version of RNN which is a simpler and a better version of LSTM. The reason why it is simpler and more efficient is because it requires fewer parameters to be updated and calculated in order to train the model. Unlike LSTM, GRU consist of 2 gates called the reset gate and the update gate. The reset gate is designed to filter out all the irrelevant information and the update gate determines what information are being transferred to the next cell. The GRU model structure that is implemented in this study will follow the encoder-decoder model with extra layers in order to try and improve the performance (35). Since GRU is known as a more efficient model then LSTM, it will be used to see if the performance is truly better in terms of accuracy.

4. Artefact development Approach

The main objective of the experiment will be to implement a series of prediction and forecasting models based on time-series data to check the confirmed cases in countries such as USA, UK, Indonesia and India. The two models that was chosen was LSTM and GRU. The three different metrics for model performance evaluation are Error rates, Mean Squared Error and Accuracy. A minimum-viable artefact was created which can produce results from where optimization and hyperparameter tuning can be performed. The next step would be to refine the model to a point where results are competitive and aimed to use a method that has not been previously researched.

4.1 Design

4.1.1Dataset

Since the dataset is only 130 rows long, we will use the whole dataset for research purposes. The three datasets will be combined in order to make it easier for the model to train and test.

4.1.2 Pre-processing

With pre-processing the data, it is crucial to find the optimal steps for input data and targets to feed into the neural network. This will be done by specifying a design to accommodate time series. Since the dataset is a multivariate dataset, it is necessary to normalize the data. Using every sample of the dataset, the first step is to tidy the dataset, normalize the input features and the last step is to split the data into training and test sets. This will help in predicting and testing the dataset. As the LSTM and GRU models will be used, transforming is needed from 2 dimensions to 3 dimensions.

4.1.3Classifier Construction

As previously discussed, the experiment will be using an RNN model for classifier construction to give a good baseline result for our dataset. The models are constructed using both LSTM and GRU to see which out of the two produces better results in term of accuracy. Both uses a gated model approach in order to regulate the flow of information and addresses the vanishing gradient problem that has caused standard RNN models to be unreliable for long sequences. Both will be constructed using the encoder-decoder approach which uses fixed-length inputs to fixed length outputs by using the sequence-to-sequence model. The gates will be able to identify important information and remove all the unnecessary information, which will be passed down the chain of sequences to make a prediction. LSTM and GRU are generally considered the most suitable for time series-based problems and at processing long sequences.

4.1.4 Training

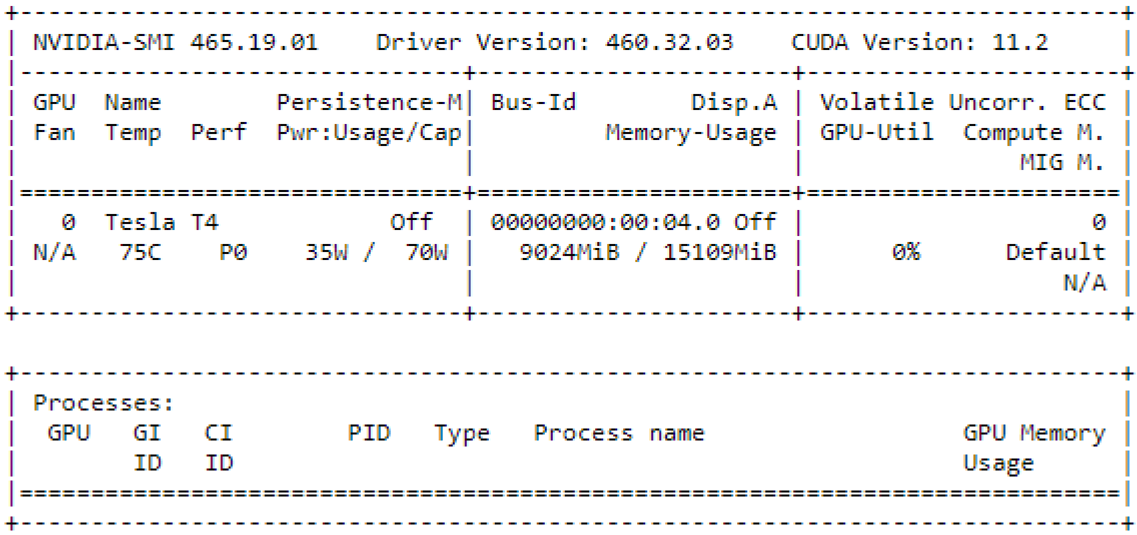
The standard of 25 epochs and a batch size of 32 will be trained initially and will be tweaked in order to see if the accuracy improves without causing underfitting or overfitting. It is important to optimize both the epochs and size in order to get the best results possible

4.1.5 Analysis of the model

Results will be analyzed using a training vs validation loss plot. Using this plot, we are able to find out at which point of the iterations the model will stop improving and the validation metrics begin to degrade as the model starts to overfit. This will give a good indication on where improvements can be made in the model.

4.1.6 Environment

All experiments and research was carried out using the Google Colab Pro GPU environment



```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 465.19.01    Driver Version: 460.32.03    CUDA Version: 11.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla T4            Off  | 00000000:00:04.0 Off |                    0 |
| N/A   75C    P0    35W /  70W |   9024MiB / 15109MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
+-----------------------------------------------------------------------------+
```

Figure 3: This figure shows the research software and hardware environment where the models was constructed and tested. All models was constructed and tested using the Google Colab Compute environment as this should allow for the most efficient processing of the models.

4.3 Experiments

After establishing a template model that shows a stable base of results, there are several experiments being conducted that to try and improve the metrics of the models. Hyperparameters tuning is very important within the models which should have a significant impact of predicted performance and metrics.

4.3.1 Pre-processing

In terms of testing, the evaluation of the effects of changing the number of data points carried forward into the LSTM/GRU models as memory units and how far forward the model will predict into the future will be tested. This will be crucial in determining the size of the data blocked by the time step unit at one time. There will also be a experiment to optimize the number of time steps forward the model will forecast. Time steps ranging from 2 to 20 will be experimented for both models to find the optimal number.

4.3.2 Designing the model

It will be important to optimize the number of neurons used to define the dimensionality of the output space. It is generally regarded that if the problem being addressed is simple and by time of nature, only a small number of neurons is required for each layer to accurately forecast daily cases. Too many neurons will not provide any additional benefits due to the LSTM/GRU model already being complexed. The number of neurons will be tested using trail and error to see which produces the best output.

It is also important to choose a type of activation to help the network learn the complex patterns in the data. The three most used activation is 'Tanh, 'sigmoid' or 'linear' and these three will be tested to see which gives the best output. This will help define the output of the nodes that are created by the LSTM and GRU models.

4.3.3 Training

The initial training of the model will start the epoch at 25 and the batch size at 32. Epoch is used to monitor the learning performance by plotting their values against the error of the model. By optimizing this number, it helps the model represent the sample will less errors. A high epoch will cause the model to overfit, and a low epoch value will cause the model to underfit. Batch size is one of the most important hyperparameter because it helps influences the dynamics of the learning algorithm which in terms controls the accuracy of the estimation of the error gradient when training RNNs. Epoch and batch sizes will be determined using a trial-and-error analysis to ensure that it does not overfit the model.

Selecting the best optimizer is crucial because it is used to change the attributes of the neural network such as the different weights and different learning rates to reduce the losses. This in turns solves problems by minimizing the function. The initial optimizer chosen will be the adaptive moment estimation (Adam), which is the most used because it is straightforward to implement and efficient at computing the dataset. It is also suited to problems with many parameters. The following optimizers will also be tested: stochastic gradient descent (SGD), root mean squared propagation (RMSProp) and Adadelta.

The learning rate is the most important hyperparameter when designing the neural network models because it controls how quickly the model adapts to the problem. This allows the model to change in response to the estimated error each time the model weights are updated. If learning rate is set to low, it could cause the training step of the model to get stuck.

The loss function will help determine the error for the current state of the model that is estimated repeatedly. This helps the weight continue to be updated and reduce the loss of the proceeding evaluation.

5. Empirical Evaluation

5.1 Research Questions

In determining if these models can be used to provide accurate forecasting of daily confirmed cases in different countries, a number of questions was created to help frame the direction of the research and a number of elements that can be tested to tune the model to the most optimal performance. There was a total of 3 questions that was kept in mind when conducting this experiment:

1. If wind speed and population density data improve the accuracy of the model. Given that these two factors are considered the most influential towards daily cases numbers, does including these two data give an improvement to predicting case numbers. The prediction is that the models would increase accuracy when implementing these two factors in the dataset.
2. Which hyperparameters provides the results as well as the best time of execution. This will be determined by analyzing each input of the model which will be tuned to best suit the problem at hand.
3. Which model out of LSTM and GRU will produce the best result, and by what means are we to determine which out of the two is superior.

5.2 Pre-Processing

A lot of different values and parameters were trialed in this model and the best values was chosen to give the best accuracy for the model. The whole dataset was used, and the average of each city was used when there was a blank or N/A. In terms of splitting the data, the training data pool needed to be large enough so that both the training and the test datasets are well presented with the problem at hand while maintaining computational efficiency. For this dataset, the optimal split between training and testing was found to be 80/20.

5.3 Designing the model

The optimal number for past data being carried forward appears to be 10. It was found while running this experiment, the higher the number of data being carried forward, the lower the accuracy. Given that our data requires forward observation in order to predict ahead of time, the optimal number for forward layers is 20. The optimal number of neurons was 2, this is probably because our dataset is considered very small. After testing out different types of activations, the optimal activation for this model is the hyperbolic tangent (tanh). Tanh helps defines the type of prediction that the model can make.

5.4 Training

The most optimal optimizer for the model was the Adam optimizer. Adam is usually the most effective and most popular when it comes to these types of problems. This is because Adam combines the best properties of the adaptive gradient algorithm (AdaGrad) and RMSProp algorithms to provide the model with an optimization algorithm that can handle sparse gradients on noisey problems.

The most used value for learning rates is 0.0001. This value has proven to be the most optimal learning rate for the problem at hand. It is important to choose the right learning rate because it allows the training process to be stable and leads the training to converge at an optimal level. During the testing phase, 0.0001 struck a balance so that the rate wasn't too large or small.

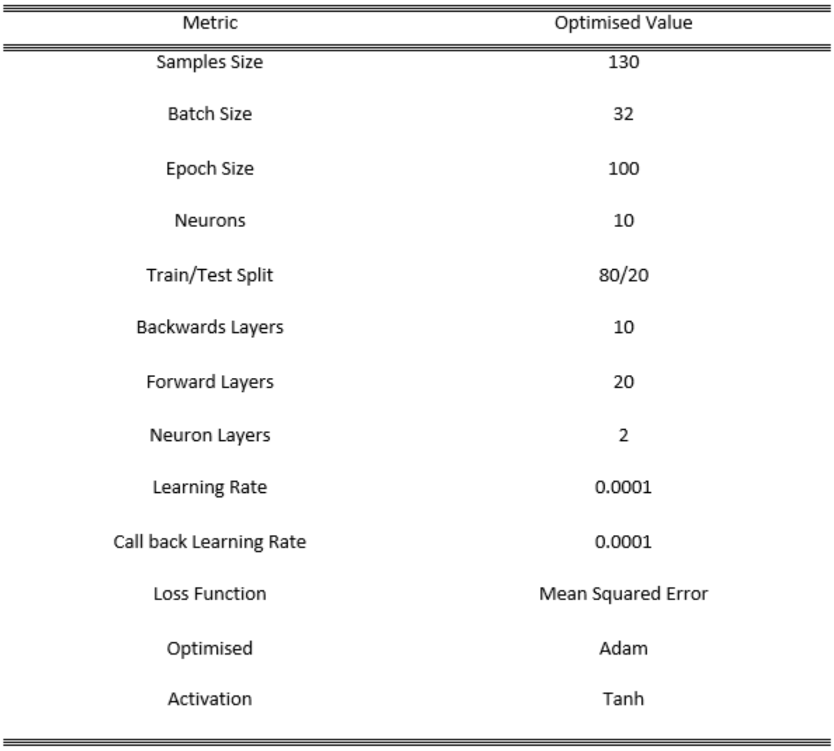
The loss function that was included in the model and found to be the most optimized is the MeanSquaredError loss function. The loss functions help the model calculates the averages of the squared differences between the predicted values and the actual values. A larger result will conclude that the model makes big errors in most of the data. The Mean Squared Error value for each trained model was low and therefore it was concluded that the model was not making errors during training and testing.

After testing several epochs, it seemed that 100 was selected as the best fit before the model underfits and after the model overfits. This means that the training data will pass through the model exactly 100 times and it was seen that anything above that will cause the accuracy to decrease. The number of samples processed before the model updates was found have a optimal number of 32. After conducting multiple test, it was found that anything above 32 will affect accuracy in a negative manner.

5.5 Finalized Model

Several experiments were conducted to conclude whether wind speed and population density factors plays an important role in the accuracy of the model. Therefore, for each LSTM and GRU model, the model was trained twice, once with the factors and once without the factors. This will help conclude whether adding in these two factors will improve the accuracy.

The table below shows the optimized hyperparameters conducted in order to produce a minimal viable product. The model was stable and provides a good base for further development in the future.



Metric	Optimised Value
Samples Size	130
Batch Size	32
Epoch Size	100
Neurons	10
Train/Test Split	80/20
Backwards Layers	10
Forward Layers	20
Neuron Layers	2
Learning Rate	0.0001
Call back Learning Rate	0.0001
Loss Function	Mean Squared Error
Optimised	Adam
Activation	Tanh

Figure 4: The optimized hyperparameters as concluded from extensive testing. These values shows the highest accuracy out of all the hyperparameters tested.

5.5.1 LSTM without factors

- Accuracy: 0.5174
- Error Rate: 0.4826
- Mean Squared Error: 0.059222

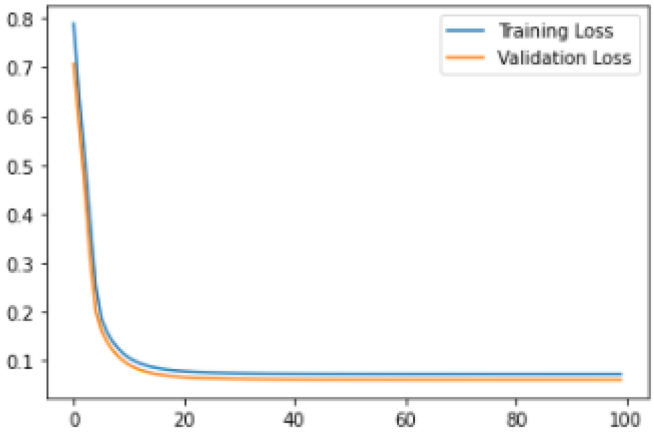


Figure 5: LSTM model without factors. This shows that the loss converges decreases sharply and then flat lines.

5.5.2 GRU without factor

- Accuracy: 0.5504
- Error Rate: 0.4496
- Mean Squared Error: 0.060094

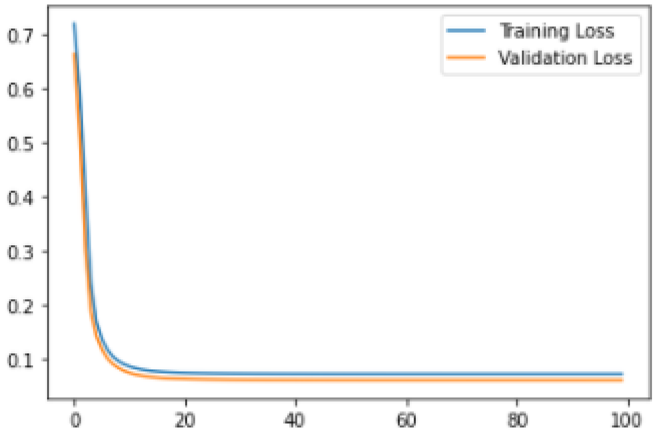


Figure 6: GRU model without factors. This shows that the loss converges decreases sharply and then flat lines.

5.5.3 LSTM with factors

- Accuracy: 0.693400
- Error Rate: 0.306600
- Mean Squared Error: 0.148720

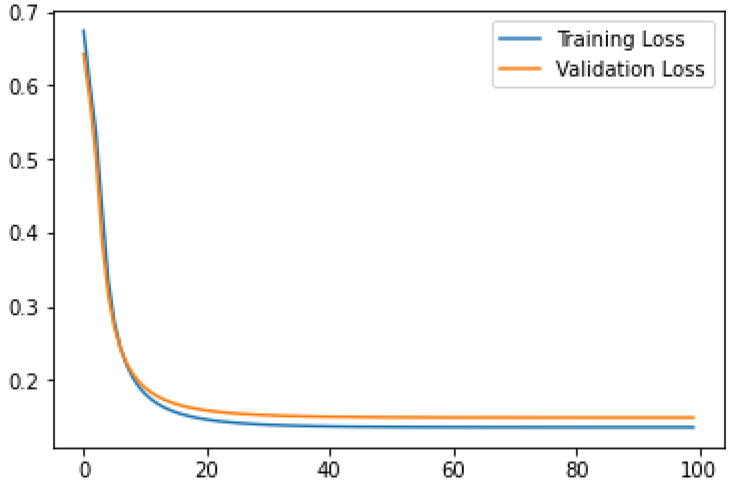


Figure 7: LSTM model with wind speed and population density factors. This shows that the loss converges decreases sharply and then flat lines.

5.5.4 GRU with factors

- Accuracy: 0.693400
- Error Rate: 0.306600
- Mean Squared Error: 0.146006

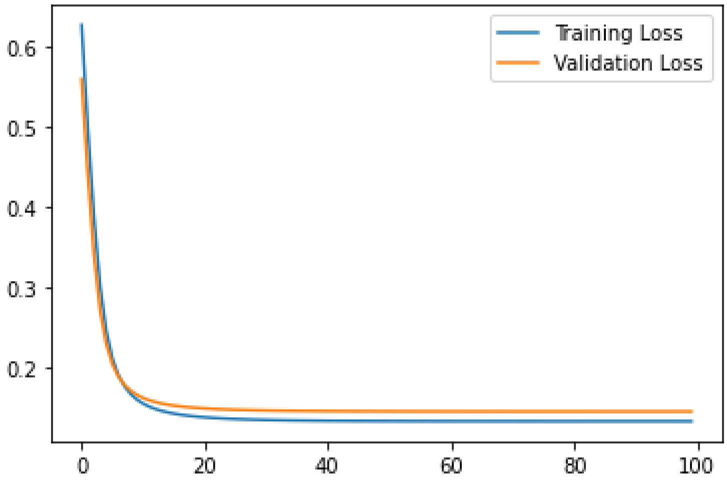


Figure 8: GRU model with wind speed and population density factors. This shows that the loss converges decreases sharply and then flat lines.

6. Discussion

From the results above, it can be clearly seen that with the inclusion of the factors, the models improve significantly. Without the factors, the LSTM model only showed an accuracy of 0.4138. This shows a poor performance by the model and should not be considered when trying to forecast future cases. When using GRU without the factors it showed an improved accuracy of 0.5862. This shows that the GRU model is far superior then LSTM when it comes to accuracy when training the models without factors. However, by adding the factors both LSTM model and GRU model showed the same accuracy of 0.6934. This showed a huge improvement from the models without the wind speed and population factors. This shows that wind speed and population factors have an impact on confirmed cases and therefore should definitely be considered when creating these models. The GRU model is slightly better than LSTM despite having the same accuracy. This is because their mean squared error is slightly lower and therefore shows that the GRU model is slightly better.

6.1 Comparison with Literature

The LSTM and GRU model in different literature reviews claims that the accuracy of their models is more than 97%. However, when this experiment was conducted, the accuracy got nowhere near that. This may be fixed with further refinements of the model, however, without the structure of the model to compare with, this may prove difficult.

6.2 Potential Issues

Some potential issues that may occurs could be that not everybody that has the virus would go and get tested therefore the daily confirmed cases data will not be accuracy. Another factor that could affect the daily cases Is the delay in results when everybody gets tested, delays can be caused by how fast the testing can be done or even government intervention. Overall, there is many factors that could change the number of daily cases and the more we are aware of this the better we can forecast future cases more accurately.

7. Threats to Validity

7.1 Threats to Construct Validity

The performance metrics that was chosen in for the current analysis relates to threats to construct validity. In this study, 3 different evaluation metrics were selected: accuracy, error rate and mean squared error. However, there are other measures such as root mean square relative error (RMSRE) and others that can be used to evaluate the time series forecasters. However, the 3 metrics used was the most recommended used for the problem at hand.

7.2 Threats of Internal Validity

The treats are primarily concerned with the internal variables that may affect the results of the models. The main internal threat is the possibility of faults during the implementation of our process. We used the best hyperparameters possible by using a error and trial approach, however, there may be a better way to find the best hyperparameter values therefore increasing our accuracy.

7.3 Threats to External Validity

This threat relates to the possibility of generalizing the current findings. The experiments were conducted using the covid-19 dataset from 31/12/2019 to 14/12/2020. The performance of the models using in this study depends on how the dataset is split into train and test data. Different results would be generated by using different timelines of the data.

8. Future Works

As people's understanding about the covid-19 virus continues to increase, there will be a deeper understanding of what factors affects the virus's transmission rate. Using this new information, there might be other factors besides wind speed and population density that can increase or decrease the transmission rate. Knowing these factors is key to better understanding covid-19 transmission rate. This will then be used to build a better model that would be more accurate in terms of forecasting daily covid-19 cases.

One noticeable theme that occurs during the LSTM and GRU training model is after a certain number of epochs, the training stalls and shows no changes in the remaining of the training. A future study can try and pinpoint the area of the model where there may be errors or inconsistency in the code. Doing this, future studies can work on researching and fine tuning the model so that the models can perform at its peak.

With these future studies in mind, there is no doubt that the model can be improved and give a better accuracy. There is no doubt that these models will be able to help government and hospitals prepare for daily covid-19 cases.

9.Conclusion

In the first 7 months of the discovery of covid-19, it has contaminated about 10 million people and caused 500,000 deaths across the world. This statistic shows that covid-19 is one of the deadliest diseases in the modern era. Since 94% of the covid-19 spread can be explained by population density and wind speed, it is important to consider these two factors in the prediction model. The main contribution of this review is to study several forecasting models as well as factors that could contribute to the transmission rate. LSTM and GRU algorithm look like the most appropriate candidate in forecasting daily cases on several domains which may assist the government and hospitals in executing appropriate strategies. DL techniques are promising and becoming more mature which makes them more attractive to assist with containing the covid-19 pandemic.

The research paper that was provided above shows a comprehensive assessment of deep learning model technologies applied to predicting covid-19 transmission rate, the results shows that by applying an encoder-decoder LSTM and GRU model we can forecast daily cases in different countries. Without the factors, LSTM and GRU had a accuracy of 0.4138 and 0.5862, while with the factors the models showed a accuracy of 0.6934. The research proves that with wind speed information and population density information, it makes the model accuracy higher. It also shows that the GRU model is indeed a better model then the LSTM model with higher accuracy and lower mean squared errors.

It is strongly encouraged for researchers to revisit this study in future months, considering that the understanding of covid-19 transmission will increase and therefore could change in terms of different factors and better models to represent it.

10.References

(1) Organisation WHO (2020). *Coronavirus disease 2019 (covid-19) situation report*. [online] Available at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200606-covid-19-sitrep-138.pdf?sfvrsn=c8abfb17_4. [Accessed 26 Jul. 2021].

(2) Zoabi, Y., Deri-Rozov, S. and Shomron, N. (2021). Machine learning-based prediction of covid-19 diagnosis based on symptoms. *npj Digital Medicine*, [online] 4(1), pp.1–5. Available at: https://www.nature.com/articles/s41746-020-00372-6 [Accessed 24 Jul. 2021].

(3) Chan, J.F.W., Yuan, S., Kok, K.H., To, K.K.W., Chu, H., Yang, J., Xing, F., Liu, J., Yip, C.C.Y., Poon, R.W.S. and Tsoi, H.W., 2020. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The lancet*, *395*(10223), pp.514-523.

(4) Wong, D.W. and Li, Y., 2020. Spreading of covid-19: Density matters. *Plos one*, *15*(12), p.e0242398.

(5) Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D.Y., Chen, L. and Wang, M., 2020. Presumed asymptomatic carrier transmission of covid-19. *Jama*, *323*(14), pp.1406-1407.

(6) Chimmula, V.K.R. and Zhang, L., 2020. Time series forecasting of covid-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals*, *135*, p.109864.

(7) Bandyopadhyay, S.K. and Dutta, S., 2020. Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release. *MedRxiv*.

(8) Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y. and Yu, T., 2020. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The lancet*, *395*(10223), pp.507-513.

(9) Painuli, D., Mishra, D., Bhardwaj, S. and Aggarwal, M., 2021. Forecast and prediction of covid-19 using machine learning. In *Data Science for covid-19* (pp. 381-397). Academic Press.

(10) Riou, J. and Althaus, C.L., 2020. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance*, *25*(4), p.2000058.

(11) Lei, S., Jiang, F., Su, W., Chen, C., Chen, J., Mei, W., Zhan, L.Y., Jia, Y., Zhang, L., Liu, D. and Xia, Z.Y., 2020. Clinical characteristics and outcomes of patients undergoing surgeries during the incubation period of covid-19 infection. *EClinicalMedicine*, *21*, p.100331.

(12) Lauer, S.A., Grantz, K.H., Bi, Q., Jones, F.K., Zheng, Q., Meredith, H.R., Azman, A.S., Reich, N.G. and Lessler, J., 2020. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, *172*(9), pp.577-582.

(13) Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X. and Cheng, Z., 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet*, *395*(10223), pp.497-506.

(14) Mao, L., Wang, M., Chen, S., He, Q., Chang, J., Hong, C., Zhou, Y., Wang, D., Li, Y., Jin, H. and Hu, B., 2020. Neurological manifestations of hospitalized patients with covid-19 in Wuhan, China: a retrospective case series study. *MedRxiv*.

(15) Coşkun, H., Yıldırım, N. and Gündüz, S., 2021. The spread of covid-19 virus through population density and wind in Turkey cities. *Science of the Total Environment*, *751*, p.141663.

(16) (2) Cao, W., Chen, C., Li, M., Nie, R., Lu, Q., Song, D., Li, S., Yang, T., Liu, Y., Du, B. and Wang, X., 2021. Important factors affecting covid-19 transmission and fatality in metropolises. *Public health*, *190*, p.e21.

(17) Rendana, M., 2020. Impact of the wind conditions on covid-19 pandemic: A new insight for direction of the spread of the virus. *Urban climate*, *34*, p.100680.

(18) Ali, Q., Raza, A., Saghir, S. and Khan, M.T.I., 2021. Impact of wind speed and air pollution on covid-19 transmission in Pakistan. *International Journal of Environmental Science and Technology*, *18*(5), pp.1287-1298

(19) Yuan, J., Yun, H., Lan, W., Wang, W., Sullivan, S.G., Jia, S. and Bittles, A.H., 2006. A climatologic investigation of the SARS-CoV outbreak in Beijing, China. *American journal of infection control*, *34*(4), pp.234-236.

(20) Ali, Q., Raza, A., Saghir, S. and Khan, M.T.I., 2021. Impact of wind speed and air pollution on covid-19 transmission in Pakistan. *International Journal of Environmental Science and Technology*, *18*(5), pp.1287-1298.

(21) Adekunle, I.A., Tella, S.A., Oyesiku, K.O. and Oseni, I.O., 2020. Spatio-temporal analysis of meteorological factors in abating the spread of covid-19 in Africa. *Heliyon*, *6*(8), p.e04749.

(22) Kadi, N. and Khelfaoui, M., 2020. Population density, a factor in the spread of covid-19 in Algeria: statistic study. *Bulletin of the National Research Centre*, *44*(1), pp.1-7.

(23) Lafta, R., Zhang, J., Tao, X., Li, Y., Abbas, W., Luo, Y., Chen, F. and Tseng, V.S., 2017, May. A fast Fourier transform-coupled machine learning-based ensemble model for disease risk prediction using a real-life dataset. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 654-670). Springer, Cham

(24) Devaraj, J., Elavarasan, R.M., Pugazhendhi, R., Shafiullah, G.M., Ganesan, S., Jeysree, A.K., Khan, I.A. and Hossain, E., 2021. Forecasting of covid-19 cases using deep learning models: Is it reliable and practically significant?. *Results in Physics*, *21*, p.103817.

(25) Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural computation*, *9*(8), pp.1735-1780.

(26) Chung, J., Gulcehre, C., Cho, K. and Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

(27) Shahid, F., Zameer, A. and Muneeb, M., 2020. Predictions for covid-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons & Fractals*, *140*, p.110212.

(28) Yin, M., Wortman Vaughan, J. and Wallach, H., 2019, May. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1-12).

(29) Kuan-Cheok, L.E.I. and Zhang, X.D., 2018, December. An approach on discretizing time series using recurrent neural network. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 2522-2526). IEEE.

(30) Olah, C., 2015. Understanding lstm networks.

(31) (Y. Bengio, P. Simard, and P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Transactions on Neural Networks, 5 (1994), pp. 157–166.)

(32) (Y. Liu, Z. Su, H. Li, and Y. Zhang, An lstm based classification method for time series trend forecasting, in 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2019, pp. 402–406.)

(33) (S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural Computation, 9 (1997), pp. 1735–1780.)

(34) (J. Wang, X. Wang, J. Li, and H. Wang, A prediction model of cnn-tlstm for usd/cny exchange rate prediction, IEEE Access, (2021), pp. 1–1)

(35) (N. Elsayed, A. S. Maida, and M. Bayoumi, Gated recurrent neural networks empirical utilization for time series classification, in 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2019, pp. 1207–1210.).

(36) ArunKumar, K.E., Kalaga, D.V., Kumar, C.M.S., Kawaji, M. and Brenza, T.M., 2021. Forecasting of covid-19 using deep layer recurrent neural networks (rnns) with gated recurrent units (grus) and long short-term memory (lstm) cells. *Chaos, Solitons & Fractals*, *146*, p.110861.

(37) Nabi, K.N., Tahmid, M.T., Rafi, A., Kader, M.E. and Haider, M.A., 2021. Forecasting COVID-19 cases: A comparative analysis between Recurrent and Convolutional Neural Networks. *Results in Physics*, *24*, p.104137.