Applications of Recurrent Neural Networks in Modeling the COVID-19 Pandemic

---

Kentaro Hayashi

Supervisor:
Professor Bruce Mellado

A research report submitted in partial fulfillment of the requirements for the degree of Master of Science in the field of e-Science

in the

School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg

7 March 2024

Declaration

I, Kentaro Hayashi, declare that this research report is my own, unaided work. It is being submitted for the degree of Master of Science in the field of e-Science at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.

Kentaro Hayashi

Kentaro Hayashi
7 March 2024

Abstract

This study attempted to introduce moving averages and a feature selection method to the forecasting model, with the aim of improving the fluctuating values and unstable accuracy of the risk index developed by the University of Witwatersrand and iThemba LABS and used by the Gauteng Department of Health. It was confirmed that the introduction of moving averages improved the fluctuation of the values, with the seven-day moving average being the most effective. For feature selection, Correlation-based Feature Selection(CFS), the simplest of the filter methods with low computational complexity, was introduced as it is not possible to spend as much time as possible on daily operations due to providing information timely. The introduction of CFS was found to enable efficient feature selection.

Acknowledgements

I am very grateful to Professor Bruce Mellado for his sound advice throughout the course of this research. I am also indebted to iThemba LABS of the National Research Foundation for their generous support, which made it possible for me to carry out this research. I also consider my colleagues in the data analytics team of DataConvergence and the Wits Institute for Collider Particle Physics, with whom I have worked on data collection and methodology, to be indispensable colleagues for life. Finally, I am deeply grateful to everyone I have met in South Africa for their kindness and acceptance of me from the distant Far East of Asia.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Discovered in Wuhan, China, in late December 2019, COVID-19, an infectious disease caused by a virus named Severe Acute Respiratory Syndrome-related Coronavirus type 2 (SARS-CoV-2), spread very quickly worldwide [1], and countries around the world implemented their own non-pharmaceutical interventions (NPIs) to prevent the spread of the disease within their countries [2]. While NPIs such as travel restrictions and curfews have been effective in reducing the spread of infection by reducing the movement of people, they have also resulted in economic stagnation [3] and increased stress among the people [4] [5]. As long as this pandemic continues, it has become clear that the government needs to adjust the NPIs in such a way that they cause as little stress as possible to economic activity and individuals.

South Africa also confirmed the first cases on 5 Mar. 2020 and started lockdown on 27 Mar., when the total number of infected people was still less than 1000; the timing was very early. The regulations were then adjusted according to the number of infected people [6], exceeding the first wave in early Sep. 2020 [7]. After that, the number of new infections remained stable at 1,000 to 2,000 per day, but in mid-Nov., a rapid increase in the number of positive cases led to the second wave of infections, which could be controlled in mid-Feb. 2021 by making the NPIs stricter again from mid-Dec. Figure 1.1 shows the daily number of cases in South Africa. However, as of April 2021, South Africa could face a third wave at any time, given that other European countries experienced the third wave soon after the second wave had subsided [8]. It was desirable to detect the rapid increase in the number of infected people as quickly as possible to help the government make decisions, such as tightening NPIs. In response to this state of affairs, the University of Witwatersrand and iThemba LABS COVID-19 modelling group developed an AI-based risk index using data on the number of infected cases and NPIs [9]. The risk index was applied by the Gauteng premier David Makhura's advisory council on COVID-19 and used in decision-making for the Gauteng COVID-19 policy [10]. This study addressed the challenges and improved the accuracy of the AI-based risk index.

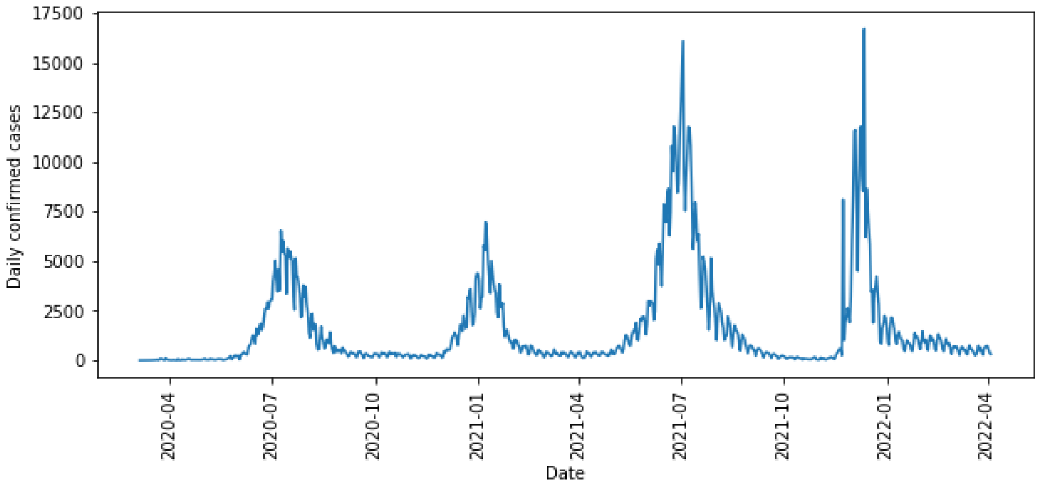


FIGURE 1.1: Change in daily infection in South Africa

1.2 Literature Review

Numerous studies have been conducted worldwide to predict the spread of COVID-19. These studies can be categorized into three main approaches: mathematical models, statistical models, and machine learning models. A notable example of a mathematical model is the Susceptible-Infected-Recovered-Deceased (SIRD) model, developed to simulate the spread of viral infections [11]. Statistical models include ARIMA and PROPHET, which are used for time series analysis [12]. Additionally, there have been many attempts to apply deep learning within the machine learning model category [13]. While each of these models has achieved certain levels of success, the high number of asymptomatic COVID-19 cases makes it difficult to determine the actual number of infections accurately. Consequently, a comprehensive predictive model has not yet been developed [14].

Since perfect prediction is difficult, The University of Witwatersrand and iThemba LABS developed a risk index using the Susceptible-Infected model (SI model) to detect situations where there is a sudden increase in the number of infected people from a steady state (called a wave) in order to help speed up decision-making [15]. However, the model is straightforward, and the fitting assumes a peak and quantifies the current situation in relation to it. As the model does not consider NPIs, there is room for defining a more accurate risk index by taking into account NPIs. This research aims to develop a risk index that is more accurate than previous studies by taking into account NPIs.

In order to incorporate NPIs into the model, quantification of NPIs is necessary. We have looked at previous studies and found quantification of data on regulation and changing mobility of people as a result of the regulation. Regarding regulations, The Oxford COVID-19 Government Response Tracker (OxCGRT) team has proposed a stringency index. The index is an aggregated and quantified version of the Policy indices of COVID-19 government responses [16]. This indicator has been used to study the impact of COVID-19 on many aspects of society [17], [18]. As South Africa has also been quantified on a daily basis, this study will also consider the use of these figures. In terms of mobility, Google has quantified their users' mobility, Parks, Transit stations, Workplaces and Residential[19]. Facebook offers two indicators as Facebook Movement Range: Change in Movement, which is a measure of how much people are moving, and Stay Put, which is a measure of how much people stay in a small area over the course of a day [20]. These two indicators of migration can be used in this study because daily data are available for each province in South Africa.

Using those NPIs, Stevenson & Hayashi et al. developed the COVID-19 risk index [9]. This risk index does not predict the actual number of confirmed but rather confirms the discrepancy between measured and predicted values. This index focuses on the fact that lulls are easy to predict due to the values being stable, and considers that a new wave may have occurred when the actual measured data deviates significantly from the prediction. Predictions are derived using LSTM RNNs with the Google Mobility Indices, Facebook Mobility Indices, and OxCGRT Stringency Index explained above as features and the number of daily positives as the objective number. LSTM RNNs are Recurrent Neural Networks (RNNs) with Long Short Term Memory added [21] [22], and Long Short Term Memory solves the vanishing gradient problem that RNNs have. Specifically, it solves this problem by introducing Long Short Term Memory cells and Short Short Term Memory cells in the RNN block [23]. In addition, as many of the NPIs used are published about once a week with a one-week delay, once a week, the data up to one week in advance is used to generate a forecast up to two weeks later, from the day of the forecast to one week later. The relative difference between that forecast value and the actual measured value is calculated and used as the risk index below,

$$Risk\ Index \quad = \quad \frac{Actual\ Daily\ Confirmed\ Cases - Predicted\ Daily\ Cases}{Predicted\ Daily\ Cases}. \quad (1.1)$$

Though the index was also adopted by the Gauteng Department of Health to formulate measures against COVID-19 [24], the index had two major areas for improvement. One is the problem of fluctuations. As can be seen from Figure 1.2, fluctuations occur during the week.

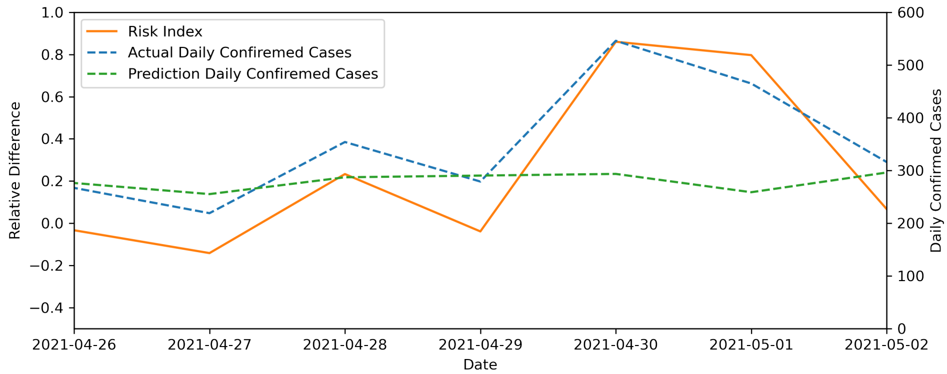


FIGURE 1.2: Example of the relative difference fluctuations

This is thought to be created mainly by differences in the operating conditions and the flow of people between weekdays and holidays at laboratories and hospitals that carry out PCR and other tests, etc. There is a concern that the values obtained when long-term forecasts are made using LSTM cannot represent fluctuations, and therefore, the errors in the forecast results increase. Moving averages have been used to counteract cyclical fluctuations of time series analysis in many studies, such as financial data [25]. It is expected that using this method will reduce fluctuations due to the day of the week.

Another improvement is in feature selection. In the previous study, all acquired NPIs were used as features for prediction [9]. It is generally known that not performing feature selection causes over-learning, which reduces prediction accuracy[26]. Feature selection methods are classified into three main categories: Filter methods, Wrapper methods and Embedded methods [27]. Filter Methods are methods that do not use machine learning models, but only data sets, and rely on the nature of the data. Therefore, they are effective for most types of machine learning models and are fast. The procedure is as follows:

1. Rank the features one by one according to the evaluation index.
2. Select and use the features with the highest rank.

Variables are calculated independently, so the relationship between variables is not considered. Wrapper Methods are methods that use machine learning models to evaluate combinations of the features. By actually using machine learning models for evaluation, it is possible to find out relationships between the features that were not found in Filter Methods, and to find the optimal combination of features for each model. To perform the Wrapper Methods, the following steps are repeated:

1. Select a combination of features.
2. Train the model using the features selected in 1.
3. Evaluate the performance.

Since Wrapper Methods evaluate the machine learning model after trying 1-3 for each combination of the features, the computational cost is very high compared to the Filter Method. Embedded Methods are methods in which feature selections are performed at model training time. The procedure is as follows:

1. Train a machine learning model.
2. Calculate the importance of the features.
3. Delete unimportant features.

Lasso models, feature importance in tree models, and linear models with regression coefficients are commonly used methods. Since the importance is evaluated during training, the computational cost is lower than that of the Wrapper Method, and this method can calculate relationships between the features that cannot be calculated by Filter methods. However, since the model is trained, the computational complexity is larger than that of Filter methods. In the case of this study, the risk index needs to be updated daily, and the model needs to be rebuilt periodically; therefore, the introduction of Filter Methods with the lowest computational complexity was considered to facilitate smooth operation.

Taking into consideration the two points of fracture reduction and feature selection, this study attempted to improve the risk index to be more accurate and reliable than the previous study.

1.3 Research Question

This research question aims to examine two aspects of the risk index used in the Ministry of Health of Gauteng, as developed in a previous study by Stevenson and Hayashi et al.[9].

- Validate whether the use of Moving Average reduces the fluctuation and improves forecast accuracy.
- Verify whether the introduction of a Filter method as a feature selection method can improve the accuracy of the prediction.

1.4 Research Aims and Objectives

1.4.1 Research Aims

This research aims to improve the accuracy of the risk index developed by Stevenson, Hayashi et al.[9]. For early detection of when the number of COVID-19 cases increases rapidly, which is used by the Gauteng State Government.

1.4.2 Research Objectives

- Confirm that the use of moving averages can reduce the fluctuation of the risk index for each day of the week.
- Confirm that the use of feature selection methods can increase forecasting accuracy.

1.5 Limitations

The study experimented with data from Gauteng. It was not discussed in this study whether the same could be said for other cantons and countries.

1.6 Overview

This study aimed to improve the accuracy of the Risk Index, an alert indicator developed by Stevenson and Hayashi et al.[9]. And adopted by the Gauteng State Ministry of Health[24] to detect early on when the number of COVID-19 cases begins to increase rapidly. We focused on the fact that the current risk index has a fluctuation problem and that no feature selection method is used, and We attempted to improve these two points.

Chapter 2

Data

Data on COVID-19 is available from many governments, research institutes and companies. The National Department of Health and NICD publish daily data on COVID-19 in South Africa, including the number of cases and deaths. In addition, since COVID-19 can be expected to spread through the number of infected people due to the activation of human movement, the actual number of infected people and data related to human movement are used to build a prediction model. For data on human mobility, I use the OxCGRT Stringency Index, Google Mobility Indices and Facebook Mobility Indices. A description of each feature is given below.

2.1 COVID-19 South African Data

The confirmed cases of COVID-19 in South Africa are published daily on the Internet by The National Department of Health [28] and The National Institute for Communicable Diseases (NICD) [29].

2.2 OxCGRT Stringency

The global prevalence of COVID-19 led to the development of several NPIs, the most comprehensive of which is the OxCGRT Stringency Index, which shows the policy and stringency of countries around the world. The index includes 20 indicators of government response, eight government response indicators such as containment and closure responses (e.g. school closures, movement restrictions), four economic policies such as income support and foreign aid for citizens, eight healthcare system-related policies such as COVID-19 testing regime, emergency investment in health and other services, vaccine-related policies. It consists of 20 indicators of government response, eight government response indicators such as school closures, movement restrictions and other containment and closure responses, four economic policies such as income support and foreign aid for citizens, and four healthcare system-related indicators such as COVID-19 testing regimes, emergency investment in health and other services, and vaccine-related policies.

The data from these 20 indicators are compiled in the following four common indices, which indicate the level of government response on a scale from 1 to 100 [30]:

1. the overall government response index, which indicates the extent to which the government's response varied across all indicators in the database.
2. containment and healthcare index: inspection policies and contact tracing, short-term investment in healthcare and vaccine investments, for example, in 'lockdowns' and closures.
3. economic support index records measures such as income support and debt forgiveness.
4. initial severity index: which records the severity of 'lockdown' type measures, mainly those that restrict people's activities.

No other index has defined policies and their severity in such detail for this pandemic.

2.3 Google Mobility Indices

Google Mobility Indecies are indicators that Google is releasing for a limited period of time while COVID-19 is being expanded. The indices are broken down by location and display the change in visits to places using the location history of mobile devices. Specifically, six places to visit are indexed: Retail and Recreation, Grocery and Pharmacy, Parks, Transit Stations, Workplaces and Residential. The numbers show the percentage of change compared with the baseline, which is the median value, for the corresponding day of the week, during the 5-week period 3 Jan.– 6 Feb. 2020 [31].

2.4 Facebook Mobility Indices

The Facebook movement data sets were developed to assist researchers and public health experts in monitoring and tracking how populations are responding to physical distancing measures [20]. The Facebook mobility report has two complimentary indicators to describe the change in movement over time: 'Change in Movement' and 'Stay Put'. Each of the indicators provides different perspectives on movement trends. The Facebook mobility report methodology divides up geographical areas into equal area tiles. The 'Change in Movement' indicator measures the number of tiles people are visiting in a day in a specific region as a comparison to a baseline defined as the average number of tiles visited daily in Feb. 2020. The 'Stay Put' indicator conversely measures how many people are staying within a single tile area for the whole day as a comparison to the February baseline. People who use Facebook on a mobile device have the choice of providing their precise location. Movement Range Trends are produced by aggregating this data.

2.5 Data acquisition timing

The table below summarises the timing of updates for each feature. It can be confirmed that the Google and Facebook Mobility Indices are available with a one-week delay, while other features are available daily. In line with the Google and Facebook Mobility Indices, a two-week forecast is made every Sunday using data that is one week behind.

TABLE 2.1: Data acquisition timing

	Data acquisition timing	Acquisition data
Confirmed cases	Every evening	Data of the day
Google Mobility Indices	Every Sunday	1 week old date
Facebook Mobility Indices	Every Sunday	1 week old date
OxCGRT Stringency Index	Every evening	Data of the day

Chapter 3

Research Methodology

3.1 Research design

This research uses an empirical approach. Using actual data from the COVID-19 in South Africa, which is the same as the risk index in the earlier Stevenson et al. study[9], a prediction model was developed based on data between the waves, which were the periods of sudden increases in the number of the infections, and the accuracy of the prediction results were compared with the results of the previous study to check for improved accuracy.

3.2 Methods

The project was divided into two parts. First, experiments were conducted to see how much the introduction of the moving average increased forecasting accuracy, stabilised the value of the risk index and made it more reliable. Then, it was checked how much the prediction accuracy could be increased by the feature selection method. Note that the same method for calculating the risk index is used as in the previous study by Fin et al.

3.2.1 Methodology of the previous study by Stevenson & Hayashi etl.

The previous study has already been described in the Literature review section, but a more in-depth explanation is provided here. The most important feature of this

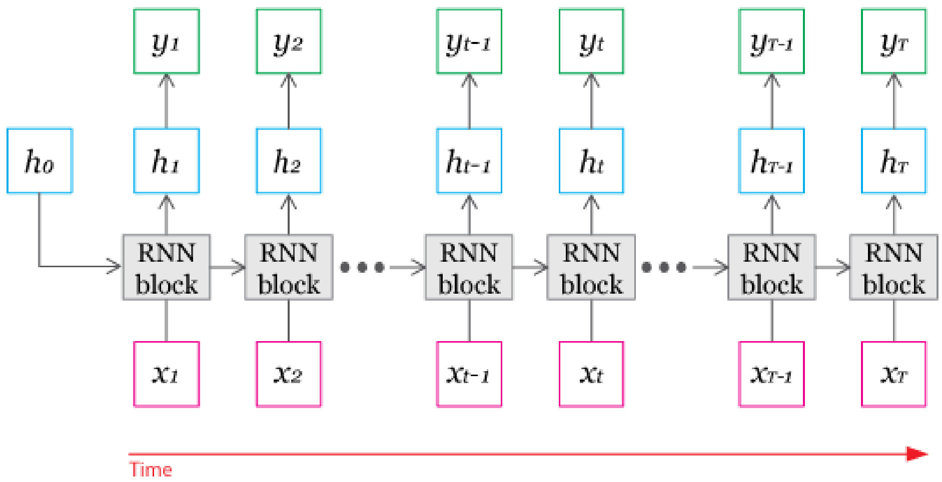


FIGURE 3.1: RNN structure

method is that it does not predict the increase of confirmed cases. It assumes that the number of new confirmed cases will stabilise in the absence of a rapid increase in the number of infections, and that these figures can be predicted. When the discrepancy between predicted and measured figures widens, it is taken as a sign that a new wave of increased infections is coming.

LSTM RNNs are used for the prediction model. RNNs have the structure shown in Fig. 3.1, where an input x_t is put in along with the time series and outputs the value y_t corresponding to the inputs through the hidden layer h_t. The hidden layer h_t is updated according to the RNN block of the previous time and the input x_t. The original RNN block looks like Figure 3.2 [21], and as tanh is used as the activation function in each RNN block, this value always takes values between 0 and 1. This means that as the value increases, the past weights are cancelled out. This phenomenon is called vanishing gradient, and the LSTM RNN solves this problem by adding long-term and short-term memory cells, as shown in Fig. 3.2. Furthermore, by introducing the forget gate vector, the LSTM RNN has evolved to be able to adjust the long-term memory cell. It is known that LSTM RNNs are generally more accurate in prediction than RNNs[22][23].

As explained in the Data chapter, the six Google Mobility Indices and two Facebook Mobility Indices are updated once a week with data up to one week old. A two-week forecast is made once a week to implement the daily risk index update.

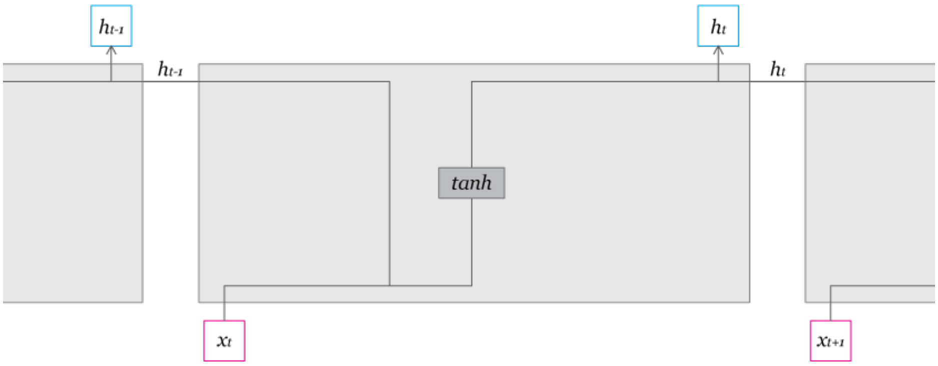


FIGURE 3.2: Detail of Simple RNN Block

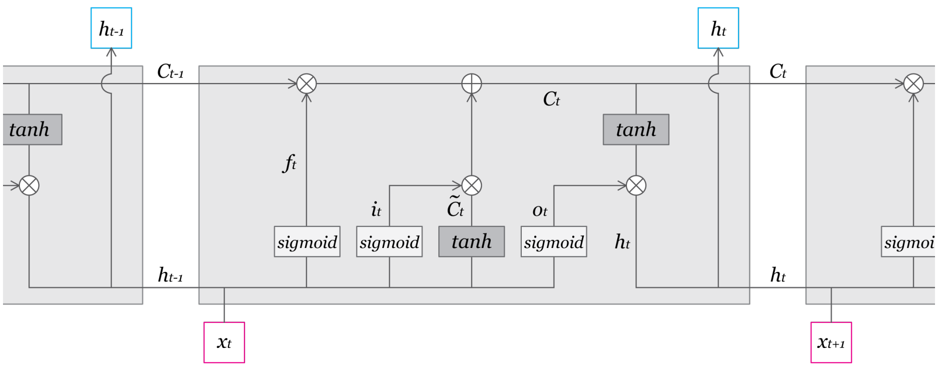


FIGURE 3.3: Detail of LSTM RNN Block

The validity of this risk index is discussed in more detail in $AppendixA$.

The problem with this method is that it is unable to cope with fluctuations during the week, resulting in a high degree of variability in the results. This leads to deviations between measured and predicted values, which can make it appear that there are signs of a wave. Furthermore, there is room for improvement in the fact that all feature values are used as they are without considering feature selection.

3.2.2 Introduction of moving averages

As already mentioned earlier, it has occurred in previous studies that the risk index is inevitably unstable in value. Moving averages, such as those used in finance, were used to remove this fluctuation in time series analysis. Moving averages can also be described as simple weighted averages using historical data. The use of a weighted average allows trends to be derived more clearly [32]. There are also two different ways of thinking about moving averages [33]. One is a backward moving average, and the other is a central moving average. The backward moving average is a method of averaging the values of each day going back n days from that day, while the central moving average is a method of calculating the average of the values of n days centred on that day. The ith position of the nth-order backward moving average, BM_n, and the ith position of the nth-order center moving average, CM_n, are given by

$$BM_n[x(i)] = \frac{1}{n}\sum_{j=0}^{n-1} x(i-j), \tag{3.1}$$

$$CM_n[x(i)] = \frac{1}{n}\sum_{j=-(n-1)/2}^{(n-1)/2} x(i+j) \text{ for } n \text{ odd}, \tag{3.2}$$

$$CM_n[x(i)] = \frac{1}{n-2}\sum_{j=-(n-2)/2}^{(n-2)/2} x(i+j) + \frac{1}{2}(x(i-n/2) + x(i+n/2)) \text{ for } n \text{ even}. \tag{3.3}$$

Since the central moving average is centered on the day when the window size is even, $1/2$ of the values at both ends are adopted and the average is taken.

In general, the central moving averages tend to be closer to the actual measured values than the backward moving averages because they are averaged around the measurement date, but they have the problem of the value resulting after $n/2$ days instead. In this study, the backward moving averages were implemented because they are necessary always to have the most up-to-date information in order to follow the objective of tracking the status of COVID-19. From a heuristic point of view, since the fluctuations in the risk index are thought to be caused by changes in the movement of people between weekends and weekdays and the opening hours of inspection stations, the use of a 7-day moving average is considered optimal. To verify whether this is actually the case, a moving average of 0 to 7 was used to check for fluctuations.

3.2.3 Introduction of a feature selection method

As mentioned earlier, the prior study has utilized all available NPIs as features to generate prediction models and a feature selection has not been considered. As discussed in the Literature Review section, feature selection can be expected to improve prediction accuracy. Feature selection methods can be divided into three main categories: filter methods, wrapper methods, and embedded methods [34]. The first, filter methods, are methods for selecting features based on statistical indicators such as correlation coefficients between explanatory and objective variables. The second, wrapper methods, are those in which a subset of features is input to a machine learning algorithm, and the required features are determined by comparing the performance of the model at that time. This method is time-consuming and machine-power-intensive, as the training process needs to be repeated many times with different combinations of features. Lastly, embedded methods are described where variable selection is carried out simultaneously as the machine learning algorithm learns [35]. Filter methods have the advantage of being computationally less expensive than wrapper methods because they only consider the association between features and class labels. Wrapper methods require time and machine power as they have to be trained many times with different combinations of features. In addition, embedded methods are limited in the number of machine learning models that can be adapted. For the LSTM used in this study, there are no embedded methods that can be adapted. From these perspectives, this study investigates how much accuracy can be improved by using a filter method.

One of the simplest methods, Correlation-based Feature Selection (CFS), was introduced. CFS is based on the hypothesis that the appropriate combination of features in Ghiselli's test theory [36] is correlated with the objective variable. For this data set, it is also hypothesised that stricter regulations will reduce the number of infected people and that increased movement of people will increase the probability of people coming into contact with each other, which will increase the number of infected people. In this study, the following equation showing standardised Pearson's correlation coefficients for all variables was used:

$$Merit_s = \frac{\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \tag{3.4}$$

where $\overline{r_{cf}}$ is the mean of the correlation coefficient between each feature and the objective variable, $\overline{r_{ff}}$ is the mean of the correlation coefficients between the respective features, and k is the number of the selected features. As the numerator is the correlation coefficient for the objective variable, the higher the subset of highly correlated features, the larger the value. For the denominator, which is the correlation coefficient for each feature, the higher the redundancy of the subset, the smaller the value [37].

3.3 Analysis

For the accuracy of the model, Root Mean Square Error (RMSE) is used and Naive forecast is used as the baseline for accuracy comparison. RMSE is the most common performance indicator for regression models and is expressed by the following formula,

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{3.5}$$

where y_i is actual confirmed cases and $\hat{y}_i$ is predicted cases. In addition, as the study required a two-week forecast, the models were evaluated on RMSE for the average forecast value over a two-week period. Furthermore, the standard deviation (σ) over two weeks was also evaluated to assess not only the accuracy but also the reduction in fluctuations. Comparisons were also made between the existing model by Stevenson & Hayashi et al. as the original. [9] and the simplest method, naive forecast.

3.4 Implementation Conditions

To improve accuracy and build a robust model, hyperparameter turning, the optimizer, and overfitting prevention measures were implemented.

3.4.1 Hyperparameter turning

We performed training using all combinations of layer count, unit count, window size, and batch size through the greedy method and adopted the combination with the highest accuracy. The conditions for each are as follows:

- Number of Layers: 1-4 layers
- Unit Size: 5, 25, 50, 100
- Window Size: 1, 3, 7
- Batch Size: 1, 2, 4, 8

3.4.2 Optimizer

We used Adam [38], which is the de facto standard optimizer.

3.4.3 Overfitting Prevention

Given the computational complexity due to the greedy method, Early Stopping [39] [40] was employed to reduce the computation time.

3.5 Limitations

In the case of South Africa, there were steady states between the first and second waves for each province and similar steady states after the second wave, so We believe that the detection of this third wave was successful. If the steady state between the waves was shorter, as in European countries, it is considered that further work is needed to adapt the method proposed in this study.

3.6 Ethical Considerations

All of the data in this research was statistical data and not the data of individual persons and so did not require ethical clearance. In addition, for Google and Facebook Mobility Indices, We dealt with data obtained and aggregated by each company using appropriate processing. Detailed policies can be found on the respective websites [41] [42].

Chapter 4

Results and Discussion

4.1 Moving Average

The figure and table below show the results when the moving average is introduced. The introduction of the moving average confirmed that the RMSE was lower compared to the original, and the RMSE and the σ were lowest at BA_7, as predicted by the heuristic. From these results, it was confirmed that the introduction of a 7-day moving average would stabilise the value of the risk index more than the current method. This result supports the hypothesis that the operation of inspection stations and the movement of people depends on the daily confirmed cases on different days of the week.

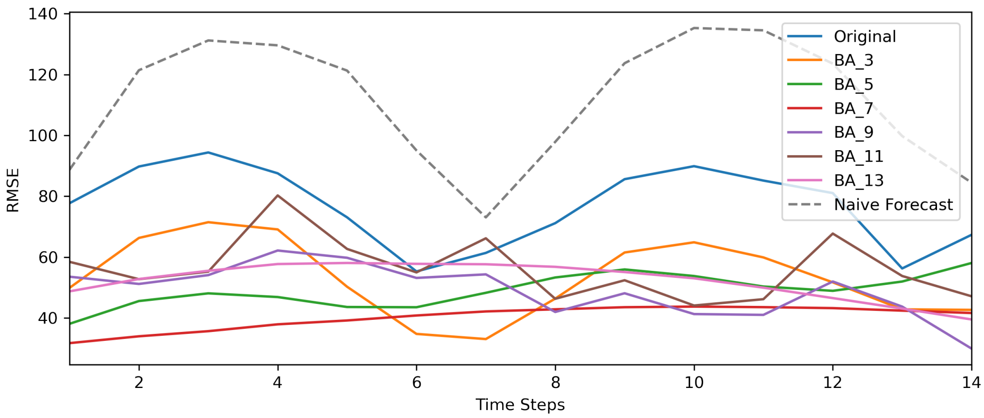


FIGURE 4.1: RMSE per timestep for each BA

TABLE 4.1: RMSE for each BA

	$RMSE(14daysAve.)$	σ
$Original$	76.8	13.0
BM_3	53.2	12.5
BM_5	49.1	5.4
BM_7	40.2	3.9
BM_9	49.1	8.6
BM_{11}	56.3	10.1
BM_{13}	52.3	5.9
$NaiveForecast$	111.4	20.8

4.2 Feature Section

To introduce CFS, the correlation coefficients between the objective variable, confirmed cases, and the explanatory variables, NPIs, were checked. As shown in FIGURE 4.2, the order of correlation coefficients is OxCGRT stringency, Google Grocery and Pharmacy, Google Residential, Google Retail and Recreation and Google Parks, Facebook Stay Put, Google Transit Stations, Facebook Change in Movement and Google Workplaces. The results of the filter method were evaluated using step backwards, an approach in which features are reduced sequentially from features with low feature correlation coefficients. Using these correlation coefficients, Merits were calculated for all 511 combinations of feature selection. (There are 512 combinations to the ninth power of two, but one combination has no more features whatsoever.) Based on the results of the calculations, three groups were formed: the top three Merit combinations that would be candidates for selection with the introduction of the CFS, a group of combinations ranked 254-256 that would be in the middle and a group of the bottom three combinations with the lowest Merit. Details of each group are given in TABLE 4.2. Predictive models were built for a subset of Merit ranks to check the 14-day average RMSE and σ.

The results are shown in FIGURE 4.3. When attention was paid to groups according to ranking, the accuracies were high in the top group as a whole, although there were some exceptions. In combinations where moving averages were introduced, it was observed that the accuracies tended to be higher in the top group and lower in the bottom group. The results suggest that CFS is effective in day-to-day operations where time is not available for tuning feature selection and that a certain improvement in accuracy can be expected by selecting features that will be at the top of the Merit ranking.

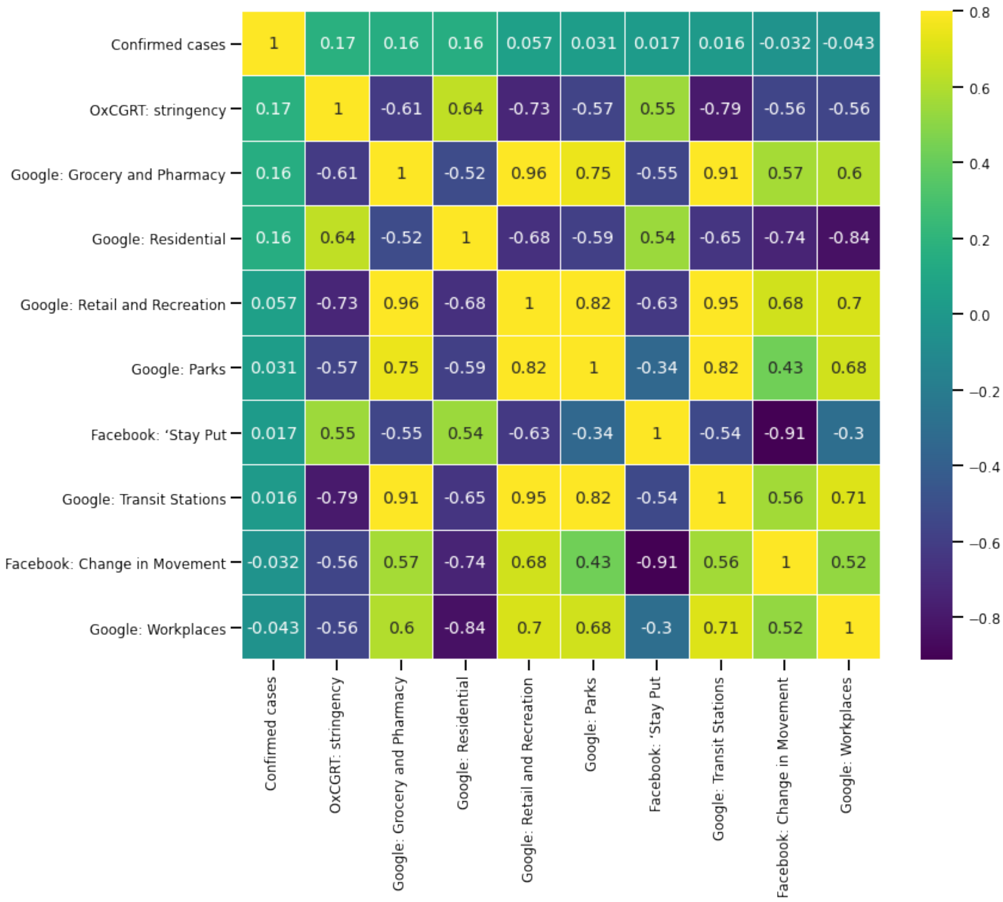
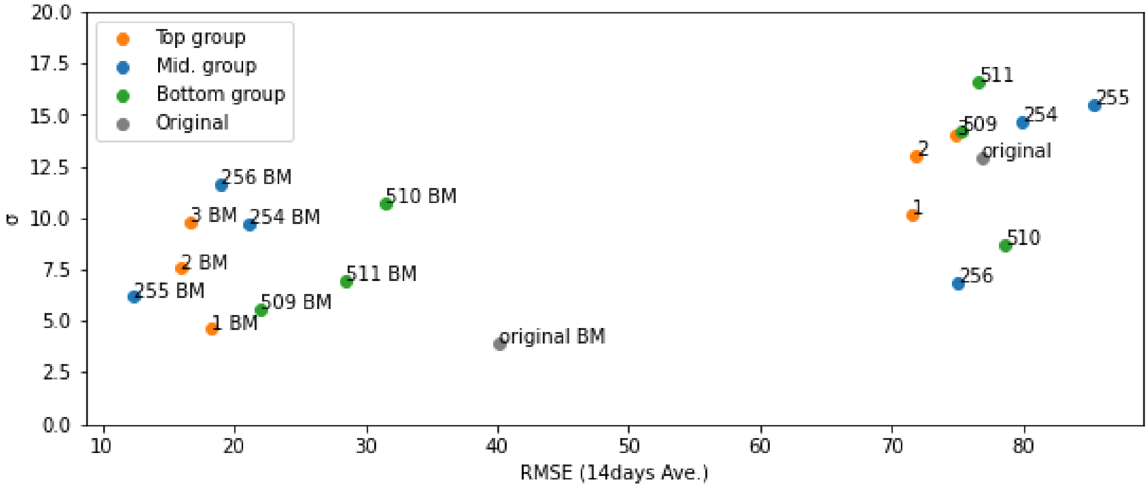


FIGURE 4.2: Correlation between explanatory and objective variables

TABLE 4.2: Group of Merits

Group	*Rank*	*Combination*	*Abs.Merit*
Top	1	$g2, s$	0.376
	2	$g2, g5, g6, s$	0.376
	3	$g1, g2, g6, s$	0.374
Original	211	$g1, g2, g3, g4, g5, g6, f1, f2, s$	0.157
Mid.	254	$g3, g6, f2$	0.137
	255	$g6, f2, s$	0.137
	256	$g3, g4, f1, f2, s$	0.137
Bottom	509	$g5, f2$	0.043
	510	$g5$	0.043
	511	$g5, f1, f2$	0.043

$g1$: Google Retail and Recreation
$g2$: Google Grocery and Pharmacy
$g3$: Google Parks
$g4$: Google Transit Stations
$g5$: Google Workplaces
$g6$: Google Residential
$gf1$: Facebook Change in Movement
$f2$: Facebook Stay Put
s: OxCGRT stringency



* Number: Rank of Merit, BM: BM_7

FIGURE 4.3: RMSE and σ for each feature subset

4.3 Summary

Backwards moving averages, which take into account the average of previous values, and the CFS, which calculates Merit based on correlation coefficients and ranks them accordingly, were introduced into the risk index by Stevenson and Hayashi et al. to see if their accuracy could be improved. It was verified that using a 7-day average for the moving averages improved forecasting accuracy and stability the most. For CFS-based feature selection, the accuracy was found to improve when the Merit value based on the correlation coefficient was high. However, the accuracy was also not maximised when Merit was maximum, but only a trend was observed.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this study, the accuracy of the risk index for wave detection in COVID-19, which has been utilised in the Gauteng department of health, was improved in two aspects: the introduction of moving averages and feature selection. For moving averages, the accuracy was improved by about 47.7% for the two-week RMSE average and by about 70.0% for the RMSE variability by introducing 7-day backward moving averages. For feature extraction, CFS, one of the simplest Filter methods, was introduced. The results were compared by classifying them into three groups based on the size of the merit: the group with the highest merit, the group with medium merit and the group with the lowest merit, and it was confirmed that the accuracy was higher in the group with the highest merit. These results confirmed the effectiveness of introducing 7-day backward moving averages and selecting a combination of features that maximises Merit by CFS in reducing fluctuations in the risk index and improving accuracy. Nevertheless, even the trend is still an improvement in accuracy, and it is considered reasonable to make predictions with the top Merit feature subset in daily operation.

5.2 Future Work

Fortunately, on 4 April 2022, President Ramaposa declared the National State of Disaster terminated, ending the COVID-19 pandemic in South Africa. Globally, many countries are returning to life as before the pandemic, and Google's mobility index has stopped updating since 17 October. In line with those changes, this risk index has also ceased to play a role. However, a situation similar to the worldwide explosion of COVID-19 could occur at any time. As the effectiveness for decision-making of the idea that a problem that deviates significantly from the forecast is a harbinger of a rapid increase in the number of infections was confirmed, although the risk index often makes it impossible actually to predict such a situation, further improvements to this risk index for the next possible infection are noted.

The main challenge with this risk index is that a certain lull period between waves is needed to create the model. If this lull is not long enough, the model's accuracy will suffer. In the case of Gauteng, the model was able to take data that was able to withstand predictions, but when the waves are close together, as in the case of the UK, the risk index cannot be calculated well. One reason for this is that only the model trend can be taken into account when forecasting 14-time steps. The forecasting model under consideration is Prophet [43], a forecasting model developed by Facebook, which is expressed by the following equation:

$$y(t) \quad = \quad g(t) + s(t) + h(t) + \epsilon t, \tag{5.1}$$

where $y(t)$ indicates the prediction, $g(t)$ the trend function, $s(t)$ the seasonal variation, $h(t)$ the holiday effect and ϵt the error term. A non-linear logistic trend or a non-linear trend can be used for $g(t)$, and in this case, because of the rapid rise and fall, it is better to use a non-linear logistic trend. This model is expected to eliminate the need for model building, especially by slicing out the lulls. Also worth mentioning is the section on holiday effects. The movement of people and the operating conditions of inspection stations on holidays are different from those on weekdays, and it wasn't easy to take this into account in the current model, but this can also be improved by using Prophet. Furthermore, the algorithm does not use deep learning and therefore has the advantage of being computationally smaller than the LSTM RNN. This is a significant advantage when considering day-to-day operations. In these respects, the introduction of Prophet is worth considering.

Appendix A

Validation of the Risk Index

A.1 Wave detection

The model built on the data in the stable state between the first and second waves is adapted to the data after the second wave has subsided and the risk index is observed. For the threshold of the risk index for each province, the distribution of the risk index is checked and 95% of the confidence interval is used as the threshold. When the risk index is exceeded for more than two consecutive days, a new wave can be regarded as likely to have occurred. Since the threshold is based on the 95% of the distribution of relative differences in the stable state between the first and second waves, there is a theoretical 5% chance that a value will exceed the threshold and only a 0.25% chance that the threshold will be exceeded on two consecutive days. As the forecasting accuracy should drop significantly when a new wave arrives, crossing the threshold two days in a row can be regarded as an indication of a new wave. The distribution of the risk index, which is a relative difference, showed a threshold value of approximately 1.0 for Gauteng.

A.2 The third wave detection

A forecasting model was created using the data from the first and second waves, and the evolution of the predicted values was observed every Monday. Figure A.1 shows the measured evolution of the risk index. If the red line, the threshold, is noted, the day on which the threshold was exceeded for two consecutive days is 6 May. At this point, the number of people actually infected with COVID-19 can be considered to be increasing rapidly. The government actually announced the third wave on 14 May[44]. This shows that this system worked.

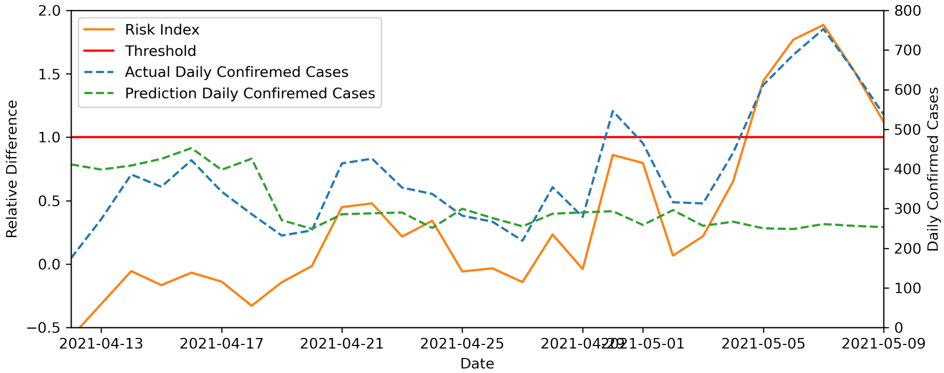


FIGURE A.1: Transition of the relative difference fluctuations

Bibliography

[1] Y.-C. Liu, R.-L. Kuo, and S.-R. Shih, "Covid-19: The first documented coronavirus pandemic in history", *Biomedical Journal*, vol. 43, no. 4, pp. 328–333, 2020, ISSN: 2319-4170. DOI: https://doi.org/10.1016/j.bj.2020.04.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2319417020300445.

[2] A. Atalan, "Is the lockdown important to prevent the covid-19 pandemic? effects on psychology, environment and economy-perspective", *Annals of Medicine and Surgery*, vol. 56, pp. 38–42, 2020, ISSN: 2049-0801. DOI: https://doi.org/10.1016/j.amsu.2020.06.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2049080120301485.

[3] M. König and A. Winkler, "Covid-19: Lockdowns, fatality rates and gdp growth", *Intereconomics: Review of European Economic Policy*, vol. 56, no. 1, pp. 32–39, 2021. DOI: 10.1007/s10272-021-0948-y. [Online]. Available: https://ideas.repec.org/a/spr/intere/v56y2021i1d10.1007_s10272-021-0948-y.html.

[4] A. Amerio, A. Brambilla, A. Morganti, *et al.*, "Covid-19 lockdown: Housing built environment's effects on mental health", *International Journal of Environmental Research and Public Health*, vol. 17, no. 16, 2020, ISSN: 1660-4601. DOI: 10.3390/ijerph17165973. [Online]. Available: https://www.mdpi.com/1660-4601/17/16/5973.

[5] C. Pieh, S. Budimir, and T. Probst, "The effect of age, gender, income, work, and physical activity on mental health during coronavirus disease (covid-19) lockdown in austria", *Journal of Psychosomatic Research*, vol. 136, p. 110 186, 2020, ISSN: 0022-3999. DOI: https://doi.org/10.1016/j.jpsychores.2020.110186. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022399920307480.

[6] *Covid-19 risk adjusted strategy*. [Online]. Available: https://sacoronavirus.co.za/covid-19-risk-adjusted-strategy/ (visited on 03/23/2023).

[7] N. Stiegler and J.-P. Bouchard, “South africa: Challenges and successes of the covid-19 lockdown”, *Annales Médico-psychologiques, revue psychiatrique*, vol. 178, no. 7, pp. 695–698, 2020, ISSN: 0003-4487. DOI: https://doi.org/10.1016/j.amp.2020.05.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003448720301785.

[8] *Covid-19 map - johns hopkins coronavirus resource center*. [Online]. Available: https://coronavirus.jhu.edu/map.html (visited on 03/23/2023).

[9] F. Stevenson, K. Hayashi, N. L. Bragazzi, *et al.*, “Development of an early alert system for an additional wave of covid-19 cases using a recurrent neural network with long short-term memory”, *International Journal of Environmental Research and Public Health*, vol. 18, no. 14, 2021, ISSN: 1660-4601. DOI: 10.3390/ijerph18147376. [Online]. Available: https://www.mdpi.com/1660-4601/18/14/7376.

[10] B. Live, *Risk of covid-19 third wave within next two weeks is low, say sa researchers*. [Online]. Available: https://www.businesslive.co.za/bd/national/health/2021-04-12-risk-of-covid-19-third-wave-within-next-two-weeks-is-low-say-sa-researchers/ (visited on 03/23/2023).

[11] D. Caccavo, “Chinese and italian covid-19 outbreaks can be correctly described by a modified sird model”, *medRxiv*, 2020. DOI: 10.1101/2020.03.19.20039388. eprint: https://www.medrxiv.org/content/early/2020/04/21/2020.03.19.20039388.full.pdf. [Online]. Available: https://www.medrxiv.org/content/early/2020/04/21/2020.03.19.20039388.

[12] C. B. Aditya Satrio, W. Darmawan, B. U. Nadia, and N. Hanafiah, “Time series analysis and forecasting of coronavirus disease in indonesia using arima model and prophet”, *Procedia Computer Science*, vol. 179, pp. 524–532, 2021, 5th International Conference on Computer Science and Computational Intelligence 2020, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.01.036. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921000417.

[13] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun, "Deep learning methods for forecasting covid-19 time-series data: A comparative study", *Chaos, Solitons Fractals*, vol. 140, p. 110 121, 2020, ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2020.110121. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S096007792030518X.

[14] V. Papastefanopoulos, P. Linardatos, and S. Kotsiantis, "Covid-19: A comparison of time series methods to forecast percentage of active cases per population", *Applied Sciences*, vol. 10, no. 11, 2020, ISSN: 2076-3417. DOI: 10.3390/app10113880. [Online]. Available: https://www.mdpi.com/2076-3417/10/11/3880.

[15] B. Lieberman, R. Gusinow, A. Asgary, *et al.*, "Big data- and artificial intelligence-based hot-spot analysis of covid-19: Gauteng, south africa, as a case study", *SSRN Electronic Journal*, Jan. 2021. DOI: 10.2139/ssrn.3803878.

[16] *University of oxford covid-19 government response tracker*. [Online]. Available: https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker (visited on 03/23/2023).

[17] S. Kumagai, T. Aoyama, E. Ino, and K. Watanabe, "Oxcgrt-based evaluation of anti-covid-19 measures taken by japanese prefectures", *Journal of Disaster Research*, vol. 16, no. 1, pp. 16–23, 2021. DOI: 10.20965/jdr.2021.p0016.

[18] P. Senthil Prakash, B. Hariharan, S. Kaliraj, R. Siva, and D. Vivek, "The impact of various policy factors implemented for controlling the spread of covid-19", *Materials Today: Proceedings*, 2021, ISSN: 2214-7853. DOI: https://doi.org/10.1016/j.matpr.2021.01.524. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214785321006155.

[19] *Covid-19 community mobility reports help overview*. [Online]. Available: https://support.google.com/covid19-mobility/answer/9824897 (visited on 03/23/2023).

[20] *Facebook research protecting privacy in facebook mobility data during the covid-19 response*. [Online]. Available: https://research.fb.com/blog/2020/06/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response/ (visited on 03/23/2023).

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] A. Graves, "Long short-term memory", in *Supervised Sequence Labelling with Recurrent Neural Networks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 37–45, ISBN: 978-3-642-24797-2. DOI: 10.1007/978-3-642-24797-2_4. [Online]. Available: https://doi.org/10.1007/978-3-642-24797-2_4.

[23] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm", in *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, vol. 2, 1999, 850–855 vol.2. DOI: 10.1049/cp:19991218.

[24] D. Maverick, *Wits develops artificial intelligence project with canadian university to tackle covid-19 in africa*. [Online]. Available: https://www.dailymaverick.co.za/article/2020-12-06-wits-develops-artificial-intelligence-project-with-canadian-university-to-tackle-covid-19-in-africa/ (visited on 03/23/2023).

[25] J. D. Hamilton, "Time series analysis", in Princeton University Press, 1994, pp. 48–52. [Online]. Available: https://www.worldcat.org/title/time-series-analysis/oclc/1194970663&referer=brief_results.

[26] A. Bommert, T. Welchowski, M. Schmid, and J. Rahnenführer, "Benchmark of filter methods for feature selection in high-dimensional gene expression survival data", *Briefings in Bioinformatics*, vol. 23, no. 1, Sep. 2021, bbab354, ISSN: 1477-4054. DOI: 10.1093/bib/bbab354. eprint: https://academic.oup.com/bib/article-pdf/23/1/bbab354/42229629/bbab354.pdf. [Online]. Available: https://doi.org/10.1093/bib/bbab354.

[27] J. Miao and L. Niu, "A survey on feature selection", *Procedia Computer Science*, vol. 91, pp. 919–926, 2016, Promoting Business Analytics and Quantitative Management of Technology: 4th International Conference on Information Technology and Quantitative Management (ITQM 2016), ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2016.07.111. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050916313047.

[28] *Covid-19 online resource and news portal*. [Online]. Available: https://sacoronavirus.co.za/ (visited on 03/23/2023).

[29] *Nicd covid-19 surveillance reports*. [Online]. Available: https://www.nicd.ac.za/diseases-a-z-index/covid-19/surveillance-reports/ (visited on 03/23/2023).

[30] *Codebook for the oxford covid-19 government response tracker*. [Online]. Available: https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/codebook.md (visited on 03/23/2023).

[31] *Covid-19 community mobility reports help understand the data*. [Online]. Available: https://support.google.com/covid19-mobility/answer/9825414 (visited on 03/23/2023).

[32] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, English, 2nd. Australia: OTexts, 2018.

[33] J. Alvarez-Ramirez, E. Rodriguez, and J. Carlos Echeverría, "Detrending fluctuation analysis based on moving average filtering", *Physica A: Statistical Mechanics and its Applications*, vol. 354, pp. 199–219, 2005, ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2005.02.020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437105001524.

[34] C. Lazar, J. Taminau, S. Meganck, *et al.*, "A survey on filter techniques for feature selection in gene expression microarray analysis", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 4, pp. 1106–1119, 2012. DOI: 10.1109/TCBB.2012.33.

[35] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective", *Neurocomputing*, vol. 300, pp. 70–79, 2018, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2017.11.077. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231218302911.

[36] E. E. Ghiselli, *Theory of psychological measurement / by E.E. Ghiselli.* (McGraw-Hill series in psychology), eng. New York, 1964.

[37] M. Hall, "Correlation-based feature selection for discrete and numeric class machine learning", *Proceedings of the 17th international conference on machine learning (ICML-2000)*, pp. 359–366, Jan. 2000.

[38] D. Kingma and J. Ba, "Adam: A method for stochastic optimization", in *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[39] L. Prechelt, "Early stopping — but when?", in *Neural Networks: Tricks of the Trade: Second Edition*, G. Montavon, G. B. Orr, and K.-R. Müller, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 53–67, ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_5. [Online]. Available: https://doi.org/10.1007/978-3-642-35289-8_5.

[40] X. Ying, "An overview of overfitting and its solutions", *Journal of Physics: Conference Series*, vol. 1168, no. 2, p. 022 022, 2019. DOI: 10.1088/1742-6596/1168/2/022022. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1168/2/022022.

[41] *Google privacy & term*. [Online]. Available: https://policies.google.com/privacy (visited on 03/23/2023).

[42] *Facebook data policy*. [Online]. Available: https://www.facebook.com/policy.php (visited on 03/23/2023).

[43] S. J. Taylor and B. Letham, "Forecasting at scale", *PeerJ Preprints*, vol. 5, e3190v2, Sep. 2017, ISSN: 2167-9843. DOI: 10.7287/peerj.preprints.3190v2. [Online]. Available: https://doi.org/10.7287/peerj.preprints.3190v2.

[44] EYEWITNESS, *Gauteng officially in covid third wave, announces premier makhura*. [Online]. Available: https://ewn.co.za/2021/05/14/gauteng-officially-in-covid-third-wave-announces-premier-makhura (visited on 03/23/2023).