Forecasting COVID-19 New Cases in South Africa using ARIMA and LSTM Time-Series Analysis Methods

---

Sindisiwe Zulu (2030273)

Supervisor:
Professor. Ritesh Ajoodha

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

16 March 2023

Declaration

I, Sindisiwe Zulu (2030273), declare that this 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.

Sindisiwe Zulu (2030273)
16 March 2023

Abstract

The COVID-19 disease continues to threaten the lives of human beings both in South Africa and globally. Its impact has been felt in practically every facet of life, resulting in a new standard of living globally. In an effort to reduce the spread of the virus, a model that can accurately estimate new cases of COVID-19 in the future is required. The daily cases in South Africa were modeled in this study utilizing Auto-Regressive Integrated Moving Average (ARIMA) model and Long Short-Term Memory (LSTM) model. Forecasts for the last 15 days of data were generated using the models. Two model performance metrics, namely the MAE and RMSE were used to find the model that can best forecast future COVID-19 cases. The LSTM model was the superior model resulting in the smallest MAE and RMSE values. This study highlights the predictive capability of LSTM model, which can be used as a tool for forecasting the spread of COVID-19 to assist the health officials and government authorities in better preparing and being informed when managing the spread of the disease.

Keywords— COVID-19; Coronavirus; Time-series forecasting; ARIMA; LSTM; Simple Exponential Smoothing; Pandemic

Acknowledgements

I would like to give gratitude to Dr Helen Robertson for the guidance and encouragement during my time at the University of the Witwatersrand. I am most grateful to all my lecturers who lectured me for the coursework of this degree. I would like to give thanks to my supervisor Professor Ritesh Ajoodha for being available to guide and support me during the development of this research work.

I would also like to give thanks to my sponsors the National e-Science Postgraduate Teaching and Training Platform (NEPTTP), this study would not be possible without your support. My special thanks to Casey Sparkes for the kind help in making sure my fees were paid on time. I am also grateful to my family for their support throughout this journey of completing my Masters degree and my son Bohlale Zulu for motivating me to work harder. Lastly, I would like to thank Dr. Blessing Ogbuokiri and Dr. Taiwo Omomule for their guidance and sharing their knowledge on the subject of data science throughout this research.

Contents

List of Figures

List of Tables

Listings

Chapter 1

Introduction

The Coronavirus was initially discovered in Wuhan, China, and was later declared a pandemic on the 11^{th} of March 2022 [35]. Individuals infected with COVID-19 commonly report fever, dry cough, and fatigue. Some people may have respiratory illnesses, nasal congestion, sore throats, aches and pains, or diarrhea [20]. While other individuals may show symptoms of COVID-19, others may not develop any symptoms.

Some individuals infected with the virus may not require special medical treatment. Those at high risk of developing significant illness because of COVID-19 include the elderly and those suffering from conditions such as diabetes, heart disease and cancer ailments [35, 20]. The infected individuals can spread the virus during coughing or sneezing where COVID-19 virus droplets are found in the saliva or released through the nose.

On the 7^{th} of May 2021 the reported cases in South Africa was over 1,5 million [20]. The ongoing rise of the number of cases in South Africa prompted government initiatives such as the implementation of lockdown laws. These regulations were introduced to help control the rate at which the virus was spreading and assist the healthcare sector with preparing their facilities should more people need medical attention.

The COVID-19 is a global epidemic that demands all citizens to work together to fully halt its spread. Finding a predictive model to better understand the disease and accurately forecast future COVID-19 cases is crucial because it may be used to assist the government in establishing regulation to halt the spread of the disease.

Since the inception of the pandemic, academic researchers have published studies that attempts to model and forecast the spread of Coronavirus [9, 50, 51].

The COVID-19 daily new cases reported represents a time-series data. Several analysis techniques can be used to study or forecast the trend of a time-series data. Statistical techniques and Machine Learning approaches have proven beneficial for time-series data analysis, with some already yielding important insights into infectious disease transmission patterns [24, 31].

The scope of machine learning has been expanded to include prediction, and a significant amount of study has been carried out on this subject. In this research the time-series of COVID-19 in South Africa recorded between March 11, 2020 to June 18, 2022, which models the pattern of the disease was used. Two distinct techniques were applied in the study, each of which has the potential to forecast the future values of the disease. The Auto-Regressive Integrated Moving Average (ARIMA) model and the Long Short-Term Memory (LSTM) model were used.

1.1 Problem Statement

The human population across the globe is under threat as a result of Coronavirus. Government officials in different countries are working tirelessly to halt the virus from spreading uncontrollable. Medical experts and scientists have developed mathematical models and studies to better understand COVID-19 and be able to advise health authorities about the next COVID-19 outbreak apex, the duration it will last for and the population size that will be infected [43].

Forecasting the number of active cases can assist the government in enforcing preventive measures. Tight restrictions can drastically alleviate the spread of COVID-19 virus. The application of time-series forecasting models that employ single or multiple features to estimate the active cases can help healthcare authorities allocate healthcare resources fairly and to plan ahead of time for resources should a need arise.

1.2 Research Aim and Objectives

The aim of this study was to forecast the number of COVID-19 new cases in South Africa using ARIMA model and LSTM model. The following specific objectives were realized in an attempt to attain the aim:

- Analyze the COVID-19 data for South Africa.
- Review methods previously applied for forecasting COVID-19 cases.
- Develop ARIMA model to forecast COVID-19 new cases.
- Develop LSTM model to forecast COVID-19 new cases.
- Compare the performances of these models using two performance metrics and suggest the best model for the provided data.

1.3 Limitations

The limitation in this study included the data quality. The data was extracted from a publicly available source which reported the COVID-19 daily confirmed cases. The dataset when filtered for South Africa consisted of missing values. One could not distinguish whether the missing value indicated a zero or no confirmed cases reported for that particular day.
Another limitation regarding the dataset was the inconsistency of the data, with some abrupt fluctuations in the number of confirmed cases on specific days. This may be because some testing facilities or government departments would be closed or did not retain records of new cases on specific days(weekends). Such data inconsistency may negatively impact the model performances.

1.4 Research Contribution

This research will contribute to the current body of knowledge by:

1. Using off-the-shelf time-series models to produce forecasts of COVID-19 new cases. This is to show that these models are applicable on COVID-19 dataset

for South Africa and may be applied on COVID-19 datasets for other countries.

2. Conducting a comparison of the time-series models used for forecasting COVID-19 dataset.

1.5 Structure of the Research Study

The research study outline is presented as follows:

- Chapter two gives an overview of the Coronavirus and the literature review on the statistical and machine learning approaches used for the modeling and prediction of COVID-19 cases.
- Chapter three presents the methodologies and methods that were employed in carrying out the experiments. This chapter includes data collection procedure, research design and evaluation metrics. The models that were used in the experimentation are outlined. Also included in this chapter is the parameter settings and hyperparameter tuning of the models and model implementation.
- Chapter four presents the experimental results and discussion of the possible causes that resulted in such outcomes.
- Chapter five summarizes the research study and concludes with the recommendations to be considered for future studies.

Chapter 2

Related Work

2.1 Introduction

The high number of people infected with Corona virus during the pandemic led to a call for research-focused interventions to assist health officials in halting the Corona virus growth in communities. This call had been answered through the development of statistical methods, artificial intelligent systems and machine learning models that analyze and predict the COVID-19 cases.

Many research institutions have acknowledged a call for a more comprehensive understanding of the COVID-19 spread and have collated data on COVID-19 daily reports across the globe and on national levels to predict COVID-19 confirmed cases. The accessibility of this data may provide insights into understanding how COVID-19 spreads and ways to contain the spread. Institutions such as the Data Science for Social Impact research (DSFSI) [4] and OurWorldInData [16] have already adopted collecting COVID-19 datasets for public use.

In this chapter the general themes related to the topic covered in this study are introduced. The themes include, COVID-19 disease overview, input varaibles used to predict COVID-19 cases, successful predictive models for capturing COVID-19 cases and COVID-19 prediction in South Africa. By studying COVID-19 spread from the global scale and narrowing down to national level (SA), will allow us to identify methods that can be used on South African datasets to predict future COVID-19 cases.

2.2 Conceptual Framework for COVID-19

Figure 2.1 shows the conceptual framework that was developed for COVID-19 confirmed cases for this research. The conceptual framework correlates the factors associated with COVID-19 cases and how the virus spread as described in literature [33, 35, 13, 12]. There are various ways in which the virus can be transmitted from one person to another. This may include being in close contact with an infected person(eg.sneezing, talking) or being in a poorly ventilated place with infected individuals [33].

Persons infected with COVID-19 disease will develop symptoms(eg.coughing, sneezing) and would take a COVID-19 test at any healthcare facility [13, 12, 33, 35]. Other individuals do not develop symptoms of COVID-19 disease when infected and are referred to as being asymptomatic [53, 32, 10]. Through taking a COVID-19 test the results may come out positive or negative. When results are positive, a person can either isolate, monitor their symptoms at home while taking medication or consult to a healthcare professional.

Individuals who tests negative for COVID-19 are advised to protect themselves and those around them by vaccinating, wearing a musk, keeping a social distance, avoiding crowded places, taking regular COVID-19 test when one has been exposed to an infected person, monitoring ones health and disinfecting surfaces prior to using them.

Due to public health concerns, tested COVID-19 patients are likely to be experience the following phases, recovery, hospitalization, ICU admission due to immune or organ failure or death [13]. If an individual has undergone either phase, records of their COVID-19 status will be collated and used for tracking purposes in the communities and on the national level by health officials[33]. Testing and diagnosis methods used for COVID-19 include RT-PCRtest, antigen test, antibody test and Chest lung CT scan images [8].

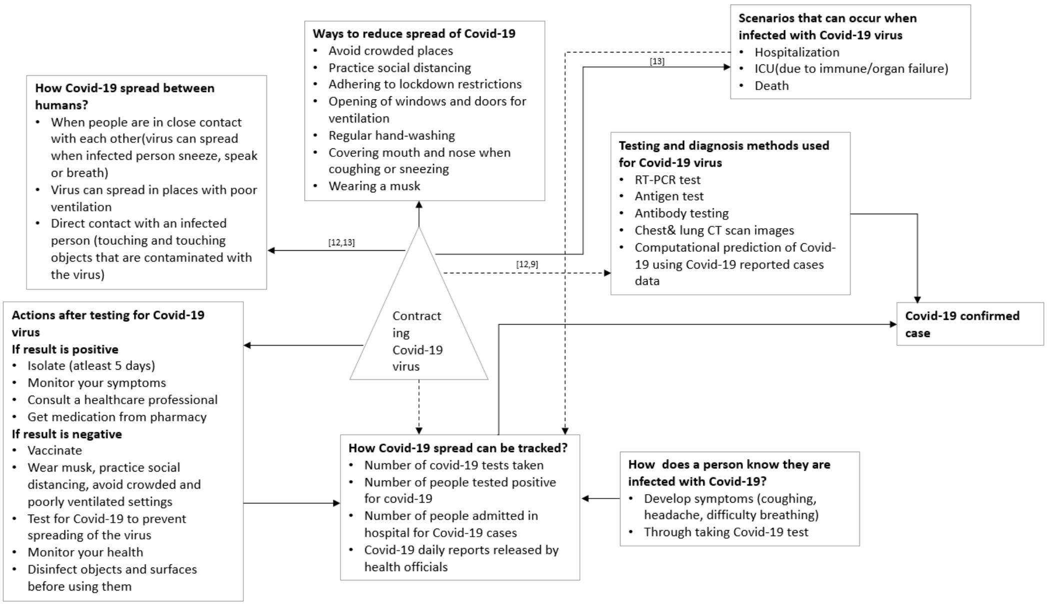

FIGURE 2.1: A Conceptual framework showing symptoms related to the spread of COVID-19 cases

2.3 Coronavirus Disease (COVID-19)

Coronavirus (COVID-19) is a contagious disease caused by a family of virus called SARS-CoV-2 virus [35]. COVID-19 spreads through tiny droplets and particles given off from the nose or mouth into the air when infected individuals sneezes, breathes, talks or coughs. The droplets when released, can accumulate on surface objects and cause contamination on the surface.

Once the surface object is contaminated with the tiny particles of the virus, a person who touches the surfaces is likely to get infected with the virus. It is for this reason that the wearing of mask, frequent hand washing, sanitizing and keeping a safe physical distance are important for prevention of COVID-19. The origin of SARS-CoV-2 is not well defined yet but the initial case was discovered in December 1st 2019 [35, 28]. More research studies are currently underway to better understand the evolution of coronavirus.

2.4 Input Variables used for Predicting COVID-19 Cases

Different variables from the Coronavirus data have been employed in the prediction of COVID-19 cases globally. The variables used include text data from web search engines, epidemiology data which include DNA sequence samples and clinical data which includes death cases, susceptible cases and confirmed cases. This section reviews the variables used in the study of Coronavirus prediction.

Clinical variables which included daily positive cases, number of recoveries, number of deceased individuals and number of active cases were used for prediction of COVID-19 in India [41]. Yulis et al. [40] used susceptible, exposed, infected, recovered, and death data in the modelling of the characterization of the COVID-19 pandemic. In the work of Muhammad et al. [29], forty one features from the epidemiology datasets were utilized in the prediction of Coronavirus in Mexico.

In the work of Da Silva et al. [15], confirmed cases and climatic features which included temperature and precipitation for Brazil and USA were used for predicting positive cases. Qin et al. [39] predicted the number of Coronavirus cases using text data as input variables extracted from a popular web search engine from China and suspected cases as input variables to their model. In the work of Soui et al. [47], two sets of datasets were used, one consisted of fifteen clinical symptoms and three demographic features while the other set consisted of eight clinical symptoms.

In the study by Sen et al. [45], the CT scan images of the chest were used to forecast Coronavirus. The human genome sequence data for COVID-19 patients and non-COVID-19 patients was used in the prediction of confirmed cases [6]. In the work of Yadaw et al. [54], clinical features from anonymised COVID-19 patients was used to predict the total number of deaths as a result of Coronavirus.

2.5 Successful Forecasting Models for Capturing COVID-19 Confirmed Cases

An exponential smoothing model ETS(MMN) was developed for forecasting COVID-19 confirmed cases and death cases on a global level [37]. A 10-day ahead forecasts were generated in the study. The model accuracy in forecasting the confirmed cases death cases was promising but could be improved with tuning of parameters. Furthermore, the model was implemented on aggregated data from different countries which contained a certain level of smoothing, as a result the model may not produce impressive results when applied on a country level data.

The Vector Autoregression Moving Average, Vector Autoregression, ARIMA and Holt-Winters' exponential smoothing models of time-series were used to predict COVID-19 pandemic [48]. The features used for forecasting included new cases, new death, new tests, positive rate, new vaccinations and hospital patients. Vector Autoregression model outperformed the univariate models. However the data used in the study was on a global level and forecasts done on country level data may not be accurate.

An ANN model was implemented to forecast one-day ahead and twenty-days ahead of new cases and death cases as a result of COVID-19 in Portugal and Brazil [11]. In order to compare the efficiency of the model, the model was trained using two different approaches. The total confirmed cases and death cases were used as the networks' input data in the first approach. In the second approach additional information about the use of masks by the public was used as input data together with the total confirmed cases and death cases.

The model resulted in satisfactory results for the first approach in the case of Brazil but higher errors were observed for Portugal. In the second approach, a reduction in errors for the death cases was observed. The information about the architecture of the ANN used was not presented in the study as a result it may be difficult to reproduce the results from the study.

ARIMA model was implemented to forecast the confirmed cases of COVID-19 in

Iraq [31]. The forecasts produced by the model copied the behavior of the original data. The study used one model which was not compared to other models to determine the effectiveness of the model. A hybrid model using an ANN and ARIMA model was developed to predict the daily cases of COVID-19 globally [5].

The purpose of using a hybrid model was to represent both the linearity and nonlinearity in the COVID-19 daily confirmed cases. ARIMA model outperformed the MLP model when the seasonal effects were not considered. MLP model performed better when seasonality was considered in the data. A limitation of the study included the use of the global data when training the model which may not capture accurate results when the model is used on a country level data.

Talkhi et al. [51] fitted nine models to COVID-19 data to find the best performing model for forecasting confirmed cases and death cases in Iran. The models used included Prophet, MLP, Holt-Winter, NNETAR, ARIMA, BSTS, TBATS, Hybrid and ELM network. MLP neural network performed best when forecasting the confirmed cases and Holt-Winter outperformed the other models when forecasting the death cases. The strength of this work included the use of diverse models to compare their predictive capability. A limitation of the study included the use of relatively small dataset for forecasting, which might negatively affect the predictions of neural networks.

The Exponential Smoothing (ES) model and Linear Regression (LR) were developed for forecasting the COVID-19 daily number of confirmed cases in India, France and Saudi Arabia [24]. ES outperformed LR and this may be as a result of the ability of ES to take into account the historic and present data during forecasting. The strength of the study included the use of diverse models.

2.6 COVID-19 Prediction in South Africa (SA)

Since the beginning of the pandemic, researchers across different countries have responded to the extraordinary public health issue by presenting prediction models and alternative plans to guide decision-making. Several mathematical models and machine learning models have been reported for modelling and predicting COVID-19 spread using South African dataset [34, 27, 18].

Gu et al. [19] investigated the spread of COVID-19 in SA using epidemiological model called eSEIRD. The model incorporated the unascertained cases, population activity and the significance of the intervention strategies. As a resulted of inadequate data, the model assumptions used in the study were derived from other countries which affected the predictions of COVID-19 cases in the study. The ascertained rate was the second limitation in this study, which assumed to have a constant distribution in a long term period whereas this might not hold true in actuality.

Niazkar and Niazkar [34] employed ANN models for predicting COVID-19 confirmed cases in seven countries which included United States of America, Iran, South Africa, Japan, Italy, China and Singapore. The study demonstrated the importance of taking into account the maximal incubation period when predicting the COVID-19 pandemic.The spreading rate of COVID-19 virus in South Africa was investigated using Bayesian inference technique [27]. The study showed a 60% reduction in the average spreading rate and an increase in community-level infections.

The analysis of COVID-19 dynamics in SA was modelled using Susceptible Infectious Recovered (SIR) model [18]. Furthermore, the evolution algorithm was employed to predict the long-term trend of the epidemic in SA. The SIR model used in the study was represented by the following equations:

$$\frac{dS(t)}{dt} = -\frac{\beta I(t)S(t)}{P}, \tag{2.1}$$

$$\frac{dI(t)}{dt} = \frac{\beta I(t)S(t)}{P} - \gamma I(t), \tag{2.2}$$

$$\frac{dW(t)}{dt} = \gamma I(t), \tag{2.3}$$

where β represents the effective contact rate and γ is the removal rate. The SIR model produced unsatisfactory prediction accuracy for SA data.

A study on short-term forecasting of COVID-19 cumulative and death cases in South Africa was presented [42]. In the study five non-linear growth models were employed which included, Richards, Gompertz, Weibull growth models, 3-parameter logistic and 4-parameter logistic. The models produced reliable and accurate forecast for 10 days ahead forecasting only. An additional limitation of the models included; using the models at specific phases of the outbreak which could produce different results during another pandemic wave.

The spread of COVID-19 under several epidemic intervention scenarios was modelled using an improved SEIR model [55]. Data for the following countries was used in the study, South Africa, Kenya, Algeria, Senegal, Nigeria and Egypt. A study that investigated clinical features and biochemical parameters associated with COVID-19 patients admitted in ICU found that levels of D-dimer and HCO3std were correlated with COVID-19 deaths [3].

TABLE 2.1: A review of key publication studies that used ARIMA model and LSTM model to predict COVID-19 cases

Author	Datasets type	Feature names	Models	Performance Measure scores
[1]	COVID-19 time series data	Confirmed cases, total number of cases, population density	LSTM	0.0579 (MAPE)
[7]	COVID-19 time series data	Confirmed cases and death cases	ARIMA	2.944 (MAPE), 3.113 (RMSE), 0.714 (R^2)
[17]	COVID-19 time series data	Total cases	ARIMA	-
[22]	COVID-19 time series data	Total Confirmed cases	LSTM	Epoch = 50: (MSE= 0.0003, RMSE= 0.0173, MAE= 0.0273), batch-size = 8: (MSE= 0.003, RMSE = 0.0173, MAE= 0.0273), lookback = 10: (MSE= 0.0002, RMSE = 0.0141, MAE= 0.0224)
[23]	COVID-19 time series data	Confirmed cases and date	ARIMA , Sigmoid fitting, SEIR, LSTM	-
[25]	COVID-19 time series data	Cumulative confirmed cases	ARIMA, LSTM and NARNN	SMAPE range: LSTM (0.16-2.55), NARNN (0.27-7.95), ARIMA (0.34-5.46)
[26]	COVID-19 time series data	Date, cum. confirmed cases, new cases, cum. deceased cases	ANN	3.981 (MAPE)
[46]	COVID-19 time series data	Confirmed cases and active cases	ARIMA	16.014 (MAPE)

Limitations of this study included limited dataset, patient dataset used was from one hospital, thus study findings may not be reliable to other regions. Cox et al. [14] estimated the COVID-19 cases at a point in time during the outbreak, based on mortality count and infection fatality rate. Limitations of the study included the use of infection fatality rate may not reflect the true estimate of COVID-19 mortality

rate. Secondly, the analysis cannot be used for predicting or forecasting COVID-19 spread.
Table 2.1 presents the current studies on the prediction of COVID-19 cases across different countries using LSTM model and ARIMA model. All the studies used COVID-19 data and the common features used for predictions included confirmed cases and cumulative cases. Different performance metrics were used to evaluate the performance of the models. The most common metrics used included MAPE and RMSE [1, 7, 22, 26, 46]. In this study, the performance metrics used included RMSE and MAE.

2.7 Conclusion

This chapter introduced the background of COVID-19 disease using a conceptual framework to describe how the disease is contracted and how the test data may be collected from individuals who have tested for COVID-19. The commonly used input variables for prediction purposes were presented in section 2.4. In section 2.5, methods that have produced promising results for predictions of COVID-19 cases were presented. The last section of this chapter presented research studies of COVID-19 predictions in South Africa.

Chapter 3

Research Methodology

3.1 Introduction

Although some statistical methods have yielded promising results in the study of infectious diseases, limitations are still observed in the generalization performance of the different methods when applied on COVID-19 data. Another limitation of the research methodology used in this research is the ARIMA model used. The ARIMA model used was a univariate model which implied that only one variable can be used to forecast the COVID-19 new cases. The use of a univariate model may result in poor model performance due to insufficient variables the model can use to learn the patterns from the data.

Therefore, these limitations motivated the application of different models which included the LSTM model to forecast COVID-19 new cases in South Africa. This chapter is structured as follows: first the experimental design of the study is outlined in section 3.2. Section 3.3 describes the data used in the study and how it was collected and prepared. The methods used in the study are outlined in section 3.4. Section 3.5 outlines the process and the implementation of the experiments. Lastly the evaluation methods used are presented in section 3.6.

3.2 Experimental Design

This section presents the research methodology used for this study. In order to achieve the aim and objectives of this research, the following research methodology was adopted as illustrated in Figure 3.1. Three experiments were conducted, which entailed the implementation of the three time series models. The models developed included Simple Exponential Smoothing model used as a baseline model, ARIMA model and LSTM model. The research methodology is further discussed in the sections that follow.

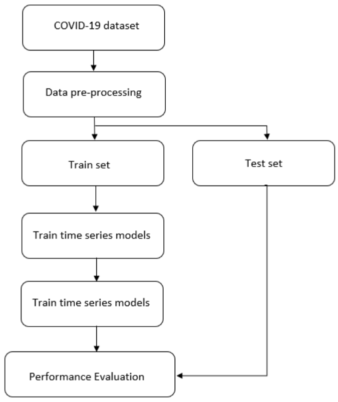

FIGURE 3.1: Research framework adopted in this research study

3.2.1 Software

The experiments were implemented on google colaboratory (Colab) platform. Colab enables the generation and execution of python code using the browser. The colab notebooks are stored on google drive to enable ease of sharing the notebook. The colab notebooks are jupyter notebooks. The python version used for the experiment was 3.7.15. The python libraries used for data analysis included pandas and numpy. The matplotlib libray was used for computing data visualizations. The ARIMA model and exponential smoothing models were computed using statsmodels package. Lastly the LSTM was implemented using Tensorflow packages.

3.3 Data

3.3.1 Experimental Data

The COVID-19 dataset used in this research study was extracted from Our World in Data [16], which is a publicly available online source. The data source providers collects and compiles the COVID-19 data from different sources globally and updates the data daily. The dataset included daily statistics on COVID-19 confirmed cases from different countries globally. The data also contained information about the COVID-19 test cases, confirmed fatalities, reproduction rate, immunizations, tests and positivity, and other characteristics.

3.3.2 Data Description

The complete time series COVID-19 dataset was downloaded as a csv file. For the purpose of this research, the COVID-19 new cases and the date columns were used. Figure 3.2 shows a subsample of the data with the field names. The columns represent different features which included location, date, new cases from the 11^{th} of March 2020 to the 18^{th} of June 2022. For the purpose of this study the location field was filtered for South Africa.

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed
0	ZAF	Africa	South Africa	2/7/2020	NaN	NaN	NaN	NaN	NaN	NaN
1	ZAF	Africa	South Africa	2/8/2020	NaN	NaN	NaN	NaN	NaN	NaN
2	ZAF	Africa	South Africa	2/9/2020	NaN	NaN	NaN	NaN	NaN	NaN
3	ZAF	Africa	South Africa	2/10/2020	NaN	NaN	NaN	NaN	NaN	NaN
4	ZAF	Africa	South Africa	2/11/2020	NaN	NaN	NaN	NaN	NaN	NaN

FIGURE 3.2: Sample of COVID-19 confirmed cases in filtered for South Africa

3.3.3 Data Pre-processing

Based on the objective of this study, a univariate time series data was used. This implied that only one time-dependent feature was used from the data available. The feature used for forecasting was the daily new cases which represented the number of COVID-19 confirmed cases reported daily. The remaining variables were excluded from the analysis. The date range used in the study was data from 11^{th} March 2020 to 18^{th} June 2022 and was converted to an index in the data. The data had 830 records as shown in table 3.1.

The data pre-processing step employed in the data included the imputation of missing data. Two records from the data had missing values. The pandas interpolate function [36] was used to impute missing values in the data. Table 3.2 presents a description of the feature used in the study. Figure 3.3 shows the plot of the time

TABLE 3.1: Properties of COVID-19 dataset for South Africa

Dataset	Records	Attributes	Feature names
COVID-19 dataset	830	2	Date, new_cases

series data after imputing the missing values.

TABLE 3.2: COVID-19 feature description

Feature name	Description
New cases	Number of COVID-19 confirmed cases recorded daily

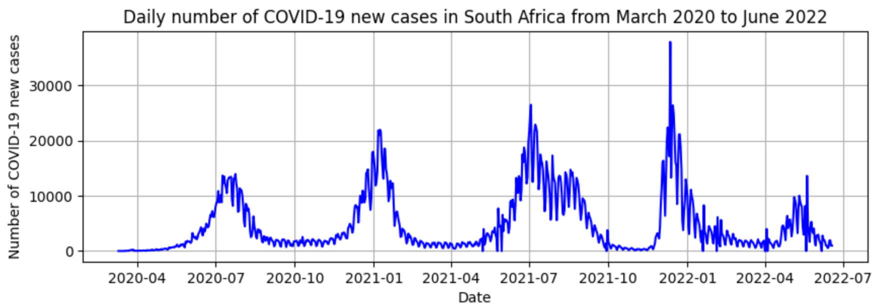

FIGURE 3.3: The COVID-19 daily reported cases in South Africa from March 2020 to June 2022

3.4 Time-Series Forecasting Models

Time-series is a sequence of observations equally spaced and recorded at fixed time intervals. The google stock price is an example of time series where the stock price is the observation collected at a fixed time interval such as minutes, hourly or daily. The objective in forecasting time-series data is to estimate how the pattern of the observations will move in the future. A time series data may exhibit the following patterns; trend, seasonal and cyclic.

A time-series data is said to have a trend when there is a long-term increase or decrease pattern in the data. This pattern can be observed by plotting the time-series data. A seasonal pattern exists as regular upward and downward fluctuations which occur as a result of seasonal factors such as during December festive season people tend to shop more frequently. Cyclic patterns exists when the data exhibits rise and dips that occurs at irregular intervals.

3.4.1 Simple Exponential Smoothing (SES)

Exponential smoothing methods was first proposed by R.G Brown, C.C Holts and P.R Winter [44]. In exponential smoothing technique, weighted averages from past values are used to generate forecasts. The weights are said to decay exponentially as past values gets older. The simple exponential smoothing is a forecasting technique for time-series data that does not have a clear trend or seasonal pattern.The smallest weights are assigned to the oldest values, which is described as follows:

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + ..., \tag{3.1}$$

where α is a smoothing parameter and lies in the range $0 \leq \alpha \leq 1$. The α parameter controls the rate at which the weights decrease.

3.4.2 ARIMA (Auto-Regressive Integrated Moving Average)

Auto-Regressive Integrated Moving Average represented by an acronym ARIMA, is a time-series model that is popularly applied in economics and statistics [44]. ARIMA is based on the intuition that future forecast are based on the information of the past observations. It is important to determine whether the time-series data is stationary before creating ARIMA models. A stationary time-series data is one whose characteristics are independent of the observation time [44].

Time-series data containing trends or seasonality are not stationary since they change the time series value at various points across time. The ARIMA model is characterized by three component models, AR, I and MA. These components are described as follows:

- (AR) Autoregressive model: In AR model, the future forecast y_t depends on the past observations referred to as lags. In mathematical terms, the future observations are based on a linear combination of the past observations of that particular variable and is represented as:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_k y_{t-k} + \epsilon_t, \tag{3.2}$$

  where ϵ_t describes the white noise in the data, y_{t-1} represents the first lag of the time series data, ϕ_1 is the coefficient of the first lag which is the slope of the line and c represents the intercept.

- (MA) Moving Average model: In MA model, the future forecasts depend on the past forecast errors. The model is represented as follows:

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q}, \tag{3.3}$$

  where the error terms represent the errors of the past forecasts.

- (I) Integration model: this model generates differencing in order to make the time series stationary.

The combination of these three models results in a non-seasonal ARIMA model denoted as ARIMA(p,d,q) model, where:

- p is the order of AR component.
- d is the number of differencing needed to make the series stationary.
- q is the order of MA component.

The partial autocorrelation (PACF) and autocorrelation (ACF) plots can be used to determine the p and q parameters of the ARIMA model. Alternatively, the parameters can be obtained through an Auto-ARIMA model which finds the optimal parameters with the lowest Akaike Information Criteria (AIC) score. The AIC is given by:

$$AIC = -2log(L) + 2(p + q + k), \tag{3.4}$$

where L describes the likelihood of the time series data, p is the AR component and q is the MA component and k value represents the intercept of the ARIMA model.

3.4.3 Machine Learning

Machine Learning is a sub-field of artificial intelligence (AI) that uses data and algorithms to learn in order to improve the performance or generate accurate predictions of a specific task. ML has been applied to different domains to make predictions, forecast and estimate the future outcome of an event. This is possible because the different ML algorithms automatically learned the relationship between the attributes and target in a dataset to deliver the trend and insight in the data.

Categories of ML methods

Machine learning approaches are categorised into supervised learning, Unsupervised learning and reinforcement learning. The methods are briefly discussed as follows:

- Supervised Learning: The objective in supervised learning is to learn the mapping that describe the relationship that exist between the inputs and outputs of the model. Labeled data is used to train algorithms and the output results in correctly classified data points or an accurate prediction of an outcome. Algorithms used in supervised learning include support vector machine (SVM), logistic regression, artificial neural networks (ANN), K-nearest neighbor (KNN) and random forest. The focus of this research in is supervised learning.

- Unsupervised Learning: In unsupervised learning unlabeled data is used to train the algorithm. Unsupervised learning is applicable for performing tasks which include exploratory data analysis, data clustering, dimensionality reduction in order to discover hidden patterns that may exist in the data. Algorithms used in unsupervised learning include k-means clustering, neural networks, anomaly detection algorithms.

- Reinforcement Learning: Reinforcement learning is a behavioral ML approach where the learner interacts with the environment to find the most suitable behavior or path. The algorithm is not trained on data but learns from experience through trial and error.

3.4.4 LSTM

The Long Short-Term Memory represented by the acronym LSTM is a type of a recurrent neural network (RNN) capable of learning both short-term dependencies and long-term dependencies. It is used to solve the exploding or vanishing gradients problem experienced in deep recurrent neural networks. LSTM was initially introduced by S. Hochreiter and J. Schmidhuber [21].

LSTM is appropriate for processing time series data and is applicable to a variety of tasks which include speech recognition and video analysis. The concept behind the LSTM network is to remember all the previous information seen and forget information not required. This is achieved through the different gated units in the LSTM structure. The building blocks of the LSTM is illustrated in figure 3.4.

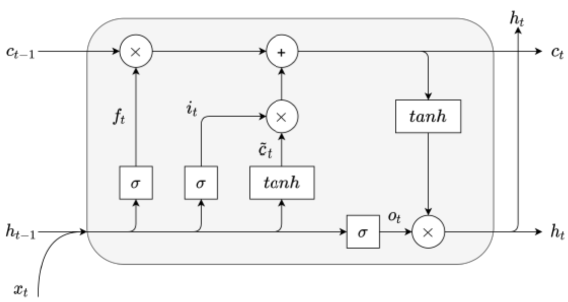

FIGURE 3.4: Graphical illustration of a LSTM unit

The fundamental elements of the LSTM unit includes the forget gate, the cell state, the input gate, and the output gate. Information is processed by the LSTM units through the following four steps:

1. **Forget gate**
   The first step is to decide which information to discard from the cell state. This is generated by the forget gate which computes the function f_t. This function considers the input vector x_t and the previous hidden state h_{t-1} and returns a value between 0 and 1. A value of 1 means the unit will keep the information and 0 means the information must be discarded. The f_t function is represented as:
   $$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3.5}$$

2. **Input gate**
   The next step is to choose the new information that will be kept in the cell state. The first process the input gate performs is deciding which information from the input vector and the previous hidden state must be stored and which information must be updated. The second process in this step is generated by the tanh layer which generates a vector of new information denoted by $\tilde{C}_t$ that could be added to the cell state. The functions that represent these two processes in the input gate are:
   $$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3.6}$$
   $$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3.7}$$

3. **Cell state**
   In the third step, the old cell state C_{t-1} is updated to the new cell state Ct based on the decisions generated in step 2. The updated state is computed through point wise multiplication of the old state and the information in the forget gate. The output is then added to the cell state. The updated cell state is represented as:
   $$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{3.8}$$

4. **Output gate**
   In the final step the output gate decides what information should be generated as output. The sigmoid layer determines which portions of the cell state will be used as output. This is followed by passing the cell state through the tanh layer (to force the values to be between -1 and 1). The sigmoid function output is then multiplied by the tanh output, which is the new information kept in the new hidden state. The functions that represents this step are:

$$\sigma_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{3.9}$$

$$h_t = o_t * tanh(C_t) \tag{3.10}$$

The key unit of the LSTM is the cell state. The cell state processes and regulates the information flow through the different gates.

3.5 Experimental Procedure

This section outlines the experimental processes followed in developing each model. The section also describes the parameters and hyper-parameter tuning employed during model development.

3.5.1 Simple Exponential Smoothing (SES)

SES model was used as a baseline model. It was implemented using the stats-models library [49]. The results produced by the baseline model were used to compare and evaluate the performance of the other models. The model took one parameter, which controlled the amount of influence from past observations. The parameter settings are listed in table 3.3. When developing the model, it was noted that the SimpleExpSmoothing model from the stats-model library lacked analytical expressions. As a result, the prediction interval estimations were generated using simulation while assuming the data represented a Gaussian distribution.

TABLE 3.3: Parameters used in the development of the SES model

Parameters	Description	Setting
endog	Time series data the model is trained on	Train data
initialization method	Method used to initialize the recursions	estimated

3.5.2 ARIMA

Prior to training the ARIMA model and Simple Exponential Smoothing model, the stationarity of the time-series data was tested using the Augmented Dickey Fuller Test (ADF). The ADF test begins with testing the null hypothesis which states that the provided time-series has a unit root, indicating that it is non-stationary [2]. If the null hypothesis is rejected, the time-series is stationary.

The Augmented Dickey Fuller test is based on a statistic value that produces a negative number, and rejection of the hypothesis is determined by that negative number [2]. The more negative the number is compared to some critical value at different level of confidence, the more strongly the hypothesis that there is a unit root is rejected at some level of confidence. The ADF test produces the following outputs, p-value, test statistics and critical values.

Since the null hypothesis presumes the presence of a unit root, the p-value returned by the test must be smaller than the significance level (0.05) in order to reject the null hypothesis [30]. The results of the ADF test obtained in this study are presented in table 3.4. The t-statistic is lesser than the critical values at all the significance levels and the p-value is less than 0.05. As a result, the null hypothesis was rejected and proved that the time-series had no unit root thus the series was stationary.

TABLE 3.4: Summary of the ADF test results for stationarity of Covid-19 new cases data in South Africa

t-statistics	Level of significance	Critical value	p value
-3.5928	1% critical value	-3.4386	0.0059
	5% critical value	-2.8651	
	10% critical value	-2.5687	

After testing for stationarity in the data, the ARIMA model was developed. The pmdarima library [38] was used to implement the ARIMA model. The key advantage of employing Auto-ARIMA function is that it first execute numerous tests to determine whether or not the time-series is stationary. It also applies a grid search approach based on Hyndman-Khandakar algorithm [44] to identify the optimal parameters p, d, and q of the ARIMA model. The Auto-ARIMA model that was used to determine the optimal ARIMA model is presented in listing A.2. The description of the parameters that were tuned is presented in table 3.5 and the final parameters selected by the Auto-ARIMA model are shown in Table 3.6.

TABLE 3.5: ARIMA model parameters tuned in the experiment

Parameters	Description	Setting
y	Time series data to which the ARIMA model should be applied	Training data.
start p	The initial value of p, defined as the order of the AR model	0
start q	The initial value of q, defined as the order of the MA model	0
d	The first-differencing order	None
test	Determines the type of unit root to employ to test stationarity of the data	adf
seasonal	Identify whether a seasonal ARIMA model must be used	False
suppress warnings	Determines whether warnings generated from statemodels should be suppressed or not	True
error action	Used to manage errors generated when ARIMA model can not be fit	Ignore
stepwise	Determines whether to apply the stepwise algorithm in order to get the optimal parameters of the model	True
trace	Determines whether to print the status of the model fitting	True

TABLE 3.6: ARIMA model parameters obtained from Auto-ARIMA function using python

Seasonal	Trend	p	d	q
no	no	5	0	1

3.5.3 LSTM

The Tensorflow library [52] was used in python to develop the LSTM model. Prior to developing the LTSM model, data pre-processing was implemented. This included scaling the data and generating the input vectors and their corresponding output vectors. Finally the LSTM model structure was selected with suitable parameters.

3.5.3.1 Data Pre-processing for LSTM Model

The dataset was split into training and testing sets where the first 800 records where assigned as train data and the last 30 records were further split into validation set and test set. From the 30 records, the first 15 rows were allocated as test data and the last 15 records were validation data.

The train and test sets were normalized using normalization. Normalization step entailed scaling the values to fall within a small specified range of [0, 1] as most machine learning algorithms do not perform well on unscaled data. Normalization was performed using the min-max normalization method as follows:

$$V'_F = \frac{v_F - min_F}{max_F - min_F}, \tag{3.11}$$

where subscript F represents features, V'_F is the new value of F, v_F represents the current value of F, max_F denotes the maximum feature value in the dataset, min_F represents minimum feature value in the dataset.

In order for the LSTM model to learn from past values, the univariate series was transformed into input vector and output vector. This was achieved by creating time lags which indicated the number of past observations the model must consider in order to predict the next day value of new cases which was the model output. In this study, the time lag were selected to be 10 days.

3.5.3.2 Parameter Tuning

The LSTM structure consisted of two layers and one dense layer. The input to the network was expressed in terms of the number of time lags as defined in the previous section and the number of feature which was one. Two parameters were tuned in the experiment: the number of layers and the learning rate. The LSTM model is presented in listing A.3.

3.6 Evaluation Methods

Forecast accuracy can be measured by looking at how well a model performs on new data that was not utilized when developing the model. The performance of the models used in this study was evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The performance metrics are defined by the following equations:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}, \tag{3.12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(\hat{y}_i - y_i)|, \tag{3.13}$$

where y_i denotes the actual value, $\hat{y}_i$ is the predicted value and n is the total number of data points. The forecast measures show how well the forecasting model can predict the previous values of the data. The output of these measures is a numerical value, where the most accurate prediction model is represented by the performance measure with the lowest value. These two commonly used metrics are easy to compute. The metrics are computed for each model and the results are compared against each other to determine the best performing model.

3.7 Conclusion

Chapter three introduced the research methodology used in this study. Section 3.3 discussed the process followed for collecting COVID-19 dataset for SA and the pre-processing steps applied on the data. In section 3.4, the time-series analysis methods were discussed. In section 3.5 the experimental procedure was discussed. This section outlined how each method was developed, the parameters used and the parameters that were tuned. The last section of this chapter introduced the performance metrics used to evaluate the performance of the models. The experimentation results are presented in chapter four.

Chapter 4

Results and Discussion

4.1 Introduction

Chapter 3 outlined the research methodology employed in this study. The main aim of this study was to conduct a comparative analysis of the performance of ARIMA model and LSTM model in order to determine the model that produced the most accurate predictions when forecasting COVID-19 new cases in South Africa (SA). The prediction accuracy of the models were measured using RMSE and MAE performance metrics respectively. Chapter 4 presents the results obtained from the experiments and the prediction accuracy results. In section 4.2, the results are presented in the form of figures and tables. In section 4.3, a comparison of the results is discussed including the potential causes of the performances of the models.

4.2 Results

4.2.1 Baseline Model: Simple Exponential Smoothing

Figure 4.1 presents the forecast results generated from simple exponential smoothing model. The black line represented the training data of COVID-19 new cases and the green line represented the predicted values of new cases for the last 15 days. The model produced constant predictions for all 15 days. The 95% prediction interval shown in figure 4.1 was more wider.

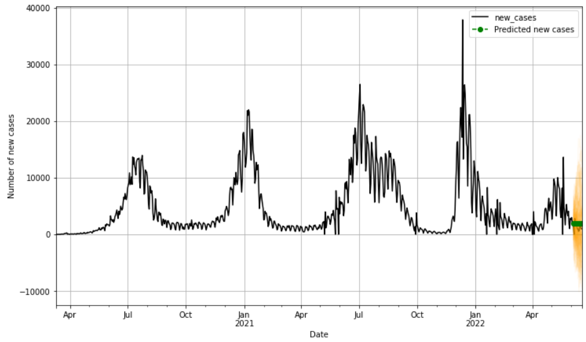

FIGURE 4.1: Forecasts of COVID-19 daily new cases for 15 days ahead using Simple Exponential Smoothing model

4.2.2 ARIMA

Figure 4.2 shows the forecast results obtained from the ARIMA model. The figure depicts the COVID-19 new cases in South Africa, where the blue line represented the actual observations used for training the model, the red line represented the test data and the green line represented the predicted values. The predictions were generated for the last 15 days. The generated predictions followed a similar trend as the training data of gradually increasing and decreasing with oscillation. It is evident as shown in figure 4.2 that the predicted number of new cases was below 15000.

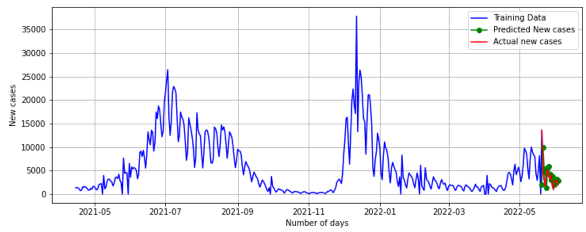

FIGURE 4.2: Forecasts of COVID-19 daily new cases for 15 days ahead based on ARIMA model

Figure 4.3 shows the predicted new cases for the last 15 days with 95% prediction interval. The 95% prediction intervals were represented in an orange shading and the predicted observations were represented by the green line are shown in figure 4.3. The prediction intervals addressed the uncertainty of the forecasts generated. The intervals provided a range upon which the true values may lie within the interval.

The predicted values were within the 95% prediction intervals. Comparing the actual new cases around June 2021 and predicted new cases around June 2022, it was observed that the number of predicted new cases in June 2022 were higher than the actual new cases in 2021.

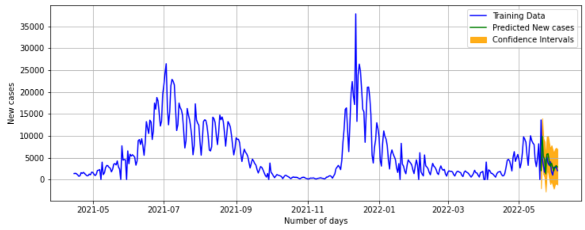

FIGURE 4.3: 95% prediction intervals for the COVID-19 daily new cases based on a ARIMA model

The forecast errors between the predicted value and the actual value are presented in table 4.1. The predicted values were slightly higher than the actual observations for some records. The forecast errors were high for the first 5 days predicted and the error slightly decreased towards the last days. The results as shown in figure 4.3 and table 4.1 shows that the number of new cases of COVID-19 in South Africa can reach a maximum of 9921 new cases within the period of June 2022.

TABLE 4.1: The forecast errors computed as the difference between the true observations and the predicted observations produced by the ARIMA model

Actual values	**Predicted values**	**Forecast Error**
13613	2057	-11556
5019	9921	4902
3220	5325	2105
1662	4639	2977
4227	1292	-2935
5284	5773	489
3800	5907	2107
3274	4262	988
4074	3015	-1059
1774	3755	1981
1004	2769	1765
2806	1969	-837
2647	2928	281
2970	3170	200
2028	2899	871

4.2.3 LSTM

The predicted results of new cases for the last 15 days using LSTM model are presented in figure 4.4. The blue line represented the test data and the orange line represented the predicted observations. The predicted results did not correspond with the trend of the actual observations. Table 4.2 shows the forecast errors obtained from the LSTM model.

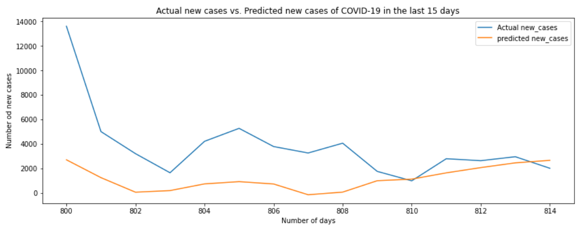

FIGURE 4.4: Forecasting results of LSTM model

TABLE 4.2: Comparison between the actual observations and the predicted observations of COVID-19 new cases generated from LSTM model

Actual values	Predicted values	Forecast Error
13613	5311	-8302
5019	5572	553
3220	5337	2117
1662	4811	3149
4227	4234	7
5284	3850	-1434
3800	3647	-153
3274	3537	263
4074	3459	-615
1774	3389	1615
1004	3327	2323
2806	3276	470
2647	3237	590
2970	3207	237
2028	3184	1156

TABLE 4.3: Performance evaluation results for the three models

Model	MAE	RMSE
Simple Exponential Smoothing	748	923
ARIMA	2337	3604
LSTM	1532	2531

4.3 Discussion

Table 4.3 presents the MAE and RMSE results for the three models employed in this study. Based on the performance metric results, the ARIMA model and the LSTM model failed to outperformed the Simple Exponential Smoothing model which was used as a baseline model. However, the LSTM model outperformed the ARIMA model where the LSTM model produced the smallest MAE value of 1532 and RMSE of 2531.

The predicted values for the baseline model and the ARIMA model were within the 95% prediction intervals. The prediction intervals for the ARIMA model were significantly wider. Since the prediction intervals accounted for the uncertainty of the forecasts, the true number of COVID-19 new cases could potentially be higher than the reported numbers. The ARIMA model selected by the Auto-ARIMA function may have been less accurate and a manual process of selecting the parameters of the ARIMA model may have been required in order to obtain the best model within the class of ARIMA models.

The LSTM model was the best model in this study. The results of this study were consistent with the findings from [23, 25], which found LSTM models more superior as compared to other models used for COVID-19 prediction. However, the LSTM model failed to forecast the number of COVID-19 new cases with significant accuracy as illustrated in table 4.2. Table 4.1 and table 4.2 presents the forecast errors which is the difference between the estimated values and the true values.

Based on table 4.1, the ARIMA model over-estimated the forecasts values for the first eleven days. This is reflected by the high value of forecast error for each estimated value. In the last four days of the predicted values for ARIMA model, the forecast error started reducing significantly. The significant decrease observed in the last days may imply the effectiveness of using ARIMA model for long-term time-series data as stated in the study by singh et al. [46].

It is a challenge to comprehend why a neural network model behaves as it does. This is because of the parameter space being too broad. However, what is better understood about the network model is that it trains on the data, attempts to generalize the training, and then predicts an output. Due to the internal structure of the LSTM model which can filter out unwanted information from its' memory and add new information, the model is able to estimate better regardless of the size of the data.

A wider range of parameter tuning and more data for COVID-19 may be required to better improve the performance of the LSTM model. In the case of COVID-19, the accuracy generated by the models was crucial since the model may be used to plan for the upsurge number of new cases in the future. If the models predicted higher values than normal, more resources may be used which in reality the high expenses may have not been necessary.

The same is true for when the models predicted lower values of new cases as compared to the actual observations, a smaller budget may be set whereas in reality more cases could be reported and more resources required. The predicted values generated using LSTM may suggest that the number of COVID-19 new cases in South Africa may initially increase, the number of new cases will eventually decrease. The decrease in the number of new cases may be a result of a number of reasons.

Firstly, the decrease in the number of new cases may imply that people are no longer testing for COVID-19 virus and they prefer recovering from their respective homes. Secondly, the decrease in numbers may imply that people are adhering to the lockdown rules imposed by the government and the government must continue with enforcing the lockdown rules in order to stop the spread of COVID-19 virus.

The experimental results obtained from this study affirmed that although more data may be required to develop models that perform better and are more accurate, the LSTM model can be used to model and predict COVID-19 cases in South Africa.

4.4 Conclusion

This chapter presented the experimental results of the methods described in chapter three. The first section presented the results of the baseline model. The Simple Exponential Smoothing model was used as a baseline model. This was followed by the experimental results of the ARIMA model. Lastly the LSTM results were presented. The results of the ARIMA model and LSTM model were compared against each other and based on the results, the LSTM model outperformed the ARIMA model.

Chapter 5

Conclusions and Future Work

5.1 Conclusion

The purpose of this study was to forecast COVID-19 new cases in South Africa using two distinct time-series approaches which includes the ARIMA model and the LSTM model. The COVID-19 dataset was collected from a publicly accessible source and filtered for South Africa [16]. The data was pre-processed by filling in missing values. The two models were then developed and compared to the baseline model (Simple Exponential Smoothing model).

After fitting the models on the training data, the models were used to forecast the number of COVID-19 new cases in the future (15 days ahead). The top performing model was determined by comparison using the two performance indicators MAE and RMSE. The LSTM model outperformed the ARIMA model, yielding the lowest MAE and RMSE values. It can be concluded that the LSTM model was the most effective time series technique for forecasting future COVID-19 instances in South Africa.

5.2 Future Work

In order to address the limitations and problems encountered when performing experiments in this study, the following recommendations should be considered for future work.

1. More COVID-19 dataset should be collected for the experiments.

2. Include more time series models to extend the comparison.
3. Use more features from the COVID-19 dataset to build more robust models that can better model the COVID-19 disease.

References

[1] S Aarathi et al. "Predicting the number of new cases of COVID-19 in India using Survival Analysis and LSTM". In: *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. IEEE. 2021, pp. 1–4.

[2] KP Ajewole, SO Adejuwon, and VG Jemilohun. "Test for stationarity on inflation rates in Nigeria using augmented dickey fuller test and Phillips-persons test". In: *J. Math* 16 (2020), pp. 11–14.

[3] Brian W Allwood et al. "Predicting COVID-19 outcomes from clinical and laboratory parameters in an intensive care facility during the second wave of the pandemic in South Africa". In: *IJID Regions* 3 (2022), pp. 242–247.

[4] Mohammad Arashi et al. "Evaluating prediction of COVID-19 at provincial level of South Africa: a statistical perspective". In: *Environmental Science and Pollution Research* (2021), pp. 1–14.

[5] Lucas Rabelo de Araújo Morais and Gecynalda Soares da Silva Gomes. "Forecasting daily Covid-19 cases in the world with a hybrid ARIMA and neural network model". In: *Applied Soft Computing* 126 (2022), p. 109315.

[6] Hilal Arslan. "Machine learning methods for covid-19 prediction using human genomic data". In: *Multidisciplinary digital publishing institute proceedings* 74.1 (2021), p. 20.

[7] Lutfi Bayyurt and Burcu Bayyurt. "Forecasting of COVID-19 cases and deaths using ARIMA models". In: *medrxiv* (2020).

[8] Martina Bergant and Ario de Marco. "Diagnostics and Monitoring of COVID-19 Infection–Current Understanding". In: (2020).

[9] SI Busari and TK Samson. "Modelling and forecasting new cases of Covid-19 in Nigeria: Comparison of regression, ARIMA and machine learning models". In: *Scientific African* (2022), e01404.

[10] Oyungerel Byambasuren et al. "Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis". In: *Official Journal of the Association of Medical Microbiology and Infectious Disease Canada* 5.4 (2020), pp. 223–234.

[11] Kathleen CM de Carvalho, João Paulo Vicente, and João Paulo Teixeira. "COVID-19 Time Series Forecasting–Twenty Days Ahead". In: *Procedia computer science* 196 (2022), pp. 1021–1027.

[12] *Centers for Disease Control and Prevention*. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html. Accessed: 2022-20-02.

[13] Marco Ciotti et al. "The COVID-19 pandemic". In: *Critical reviews in clinical laboratory sciences* 57.6 (2020), pp. 365–388.

[14] Laura Cox and Clarence Suh Yah. "Estimating actual COVID-19 case numbers using cumulative death count-A method of measuring effectiveness of lockdown of non-essential activities: a South African case study". In: *The Pan African Medical Journal* 35.Suppl 2 (2020).

[15] Ramon Gomes Da Silva et al. "Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables". In: *Chaos, Solitons & Fractals* 139 (2020), p. 110027.

[16] Our World in Data et al. "Coronavirus Pandemic (COVID-19)". In: (2021).

[17] Tania Dehesh, HA Mardani-Fard, and Paria Dehesh. "Forecasting of covid-19 confirmed cases in different countries with arima models". In: *MedRxiv* (2020).

[18] Wei Ding, Qing-Guo Wang, and Jin-Xi Zhang. "Analysis and prediction of COVID-19 epidemic in South Africa". In: *ISA transactions* (2021).

[19] Xuelin Gu et al. "COVID-19 PREDICTION IN SOUTH AFRICA: ESTIMATING THE UNASCERTAINED CASES-THE HIDDEN PART OF THE EPIDEMIOLOGICAL ICEBERG." In: *medRxiv* (2021).

[20] Minister of Health. "COVID-19 Statistics in South Africa". In: (2021).

[21] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[22] Clarissa Angelita Indriyani, Claudia Rachel Wijaya, and Nunung Nurul Qomariyah. "Forecasting COVID-19 Total Daily Cases in Indonesia Using LSTM Networks". In: *2022 5th International Conference on Computing and Informatics (ICCI)*. IEEE. 2022, pp. 385–391.

[23] Aishwarya Jakka et al. "Forecasting COVID-19 cases in India using machine learning models". In: *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*. IEEE. 2020, pp. 466–471.

[24] Badriya Khayyat, Fouzi Harrou, and Ying Sun. "Predicting COVID-19 Spread using Simple Time-Series Statistical Models". In: *2021 International Conference on ICT for Smart Society (ICISS)*. IEEE. 2021, pp. 1–5.

[25] İsmail Kırbaş et al. "Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches". In: *Chaos, Solitons & Fractals* 138 (2020), p. 110015.

[26] Pratima Kumari and Durga Toshniwal. "Real-time estimation of COVID-19 cases using machine learning and mathematical models-The case of India". In: *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*. IEEE. 2020, pp. 369–374.

[27] Rendani Mbuvha and Tshilidzi Marwala. "Bayesian inference of COVID-19 spreading rates in South Africa". In: *PloS one* 15.8 (2020), e0237126.

[28] Johns Hopkins Medicine. "Coronavirus". In: (2021).

[29] LJ Muhammad et al. "Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset". In: *SN computer science* 2.1 (2021), pp. 1–13.

[30] Rizwan Mushtaq. "Augmented dickey fuller test". In: (2011).

[31] Hadeel I Mustafa and Noor Y Fareed. "Covid-19 cases in Iraq; forecasting incidents using box-Jenkins Arima model". In: *2020 2nd Al-Noor International Conference for Science and Technology (NICST)*. IEEE. 2020, pp. 22–26.

[32] MWM Mustafa. "Audiological profile of asymptomatic Covid-19 PCR-positive cases". In: *American Journal of Otolaryngology* 41.3 (2020), p. 102483.

[33] *National Institute for Communicable Diseases*. https://www.nicd.ac.za/diseases-a-z-index/disease-index-covid-19/surveillance-reports/. Accessed: 2022-20-02.

[34] Hamid Reza Niazkar and Majid Niazkar. “Application of artificial neural networks to predict the COVID-19 outbreak”. In: *Global health research and policy* 5.1 (2020), pp. 1–11.

[35] World Health Organization et al. “WHO-convened global study of origins of SARS-CoV-2: China part”. In: (2021).

[36] *Pandas interpolation function*. URL: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.interpolate.html (visited on 10/10/2022).

[37] Fotios Petropoulos, Spyros Makridakis, and Neophytos Stylianou. “COVID-19: Forecasting confirmed cases and deaths with a simple time series model”. In: *International journal of forecasting* (2020).

[38] *pmdarima: ARIMA estimators for Python - pmdarima 2.0.1 documentation*. URL: https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html (visited on 09/20/2022).

[39] Lei Qin et al. “Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index”. In: *International journal of environmental research and public health* 17.7 (2020), p. 2365.

[40] Yullis Quintero et al. “Machine learning models for the prediction of the SEIRD variables for the COVID-19 pandemic based on a deep dependence analysis of variables”. In: *Computers in Biology and Medicine* 134 (2021), p. 104500.

[41] Smita Rath, Alakananda Tripathy, and Alok Ranjan Tripathy. “Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model”. In: *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14.5 (2020), pp. 1467–1474.

[42] Tarylee Reddy et al. “Short-term real-time prediction of total number of reported COVID-19 cases and deaths in South Africa: a data driven approach”. In: *BMC medical research methodology* 21.1 (2021), pp. 1–11.

[43] Matheus Henrique Dal Molin Ribeiro et al. “Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil”. In: *Chaos, Solitons & Fractals* 135 (2020), p. 109853.

[44] George Athanasopoulos Rob J Hyndman. *Forecasting: Principles and Practice*. 2021.

[45] Shibaprasad Sen et al. "A bi-stage feature selection approach for COVID-19 prediction using chest CT images". In: *Applied Intelligence* 51.12 (2021), pp. 8985–9000.

[46] Sarbhan Singh et al. "Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models". In: *The Journal of Infection in Developing Countries* 14.09 (2020), pp. 971–976.

[47] Makram Soui et al. "NSGA-II as feature selection technique and AdaBoost classifier for COVID-19 prediction using patient's symptoms". In: *Nonlinear dynamics* 106.2 (2021), pp. 1453–1475.

[48] K Sreehari et al. "A Comparative Study between Univariate and Multivariate Time Series Models for COVID-19 Forecasting". In: *2021 International Conference on Computational Performance Evaluation (ComPE)*. IEEE. 2021, pp. 485–491.

[49] *Statsmodel SimpleExpSmoothing*. URL: https://www.statsmodels.org/dev/generated/statsmodels.tsa.holtwinters.SimpleExpSmoothing.html (visited on 10/15/2022).

[50] Jian Sun. "Forecasting COVID-19 pandemic in Alberta, Canada using modified ARIMA models". In: *Computer Methods and Programs in Biomedicine Update* 1 (2021), p. 100029.

[51] Nasrin Talkhi et al. "Modeling and forecasting number of confirmed and death caused COVID-19 in IRAN: A comparison of time series forecasting methods". In: *Biomedical Signal Processing and Control* 66 (2021), p. 102494.

[52] *Tensorflow*. URL: https://www.tensorflow.org/ (visited on 10/05/2022).

[53] Rui Wang et al. "Decoding asymptomatic COVID-19 infection and transmission". In: *The journal of physical chemistry letters* 11.23 (2020), pp. 10007–10015.

[54] Arjun S Yadaw et al. "Clinical features of COVID-19 mortality: development and validation of a clinical prediction model". In: *The Lancet Digital Health* 2.10 (2020), e516–e525.

[55] Zebin Zhao et al. "Prediction of the COVID-19 spread in African countries and implications for prevention and control: A case study in South Africa, Egypt, Algeria, Nigeria, Senegal and Kenya". In: *Science of the Total Environment* 729 (2020), p. 138959.

Appendix A

Time-series models used in the study

A.1 Baseline Model: Simple Exponential Smoothing Model

Listsing A.1 presents the Simple Exponential Smoothing model that was used in the study as a baseline model.

LISTING A.1: Simple Exponential Smoothing model

```
from statsmodels.tsa.api import ExponentialSmoothing
from statsmodels.tsa.api import SimpleExpSmoothing
from statsmodels.tsa.exponential_smoothing.ets import ETSModel

# define Simple exponential smoothing model
fit = SimpleExpSmoothing(train,
initialization_method="estimated").fit()
simulations = fit.simulate(14, repetitions=100)
```

A.2 Auto-Regressive Integrated Moving Average (ARIMA) Model

Listsing A.2 presents the Auto-ARIMA model that was used for finding the optimal ARIMA model.

LISTING A.2: Auto-ARIMA model

```
from statsmodels.tsa.arima_model import ARIMA
```

```
import pmdarima as pm

# define ARIMA model
model = pm.auto_arima(train['new_cases'], start_p=0, start_q=0,
                      d=None,
                      test='adf',
                      seasonal=False,
                      suppress_warnings=True,
                      error_action='ignore',
                      stepwise=True,
                      trace=True)

model.summary()
```

A.3 Long Short-Term Memory (LSTM) model

Listing A.3 presents the LSTM model that was used in the study.

LISTING A.3: LSTM model architecture used in the study

```
import tensorflow as tf
from tensorflow.keras import Sequential
from tensorflow.keras.layers import LSTM,
Dense, Dropout, GRU, Flatten

# define model
model = Sequential()
model.add(LSTM(20, activation='relu', return_sequences=True,
input_shape=(n_steps, n_features)))
model.add(LSTM(30, activation='relu'))
model.add(Dense(1))
model.summary()

adam = tf.keras.optimizers.Adam(lr=0.001)
model.compile(optimizer ='adam', loss = 'mse')
```

```
history = model.fit(X_train, y_train,
epochs = 50, batch_size = 1, shuffle=False)
```