ZAMBIA INFORMATION COMMUNICATION TECHNOLOGY (ICT) JOURNAL

Volume 6 (Issue 1) (2022) Pages 7-17

Challenges of using Data Mining Techniques to Analyze and Forecast COVID-19 Pandemic in Zambia

James Sakala[a] and Douglas Kunda[b]

a. Department of Computer Science, Mulungushi University, Kabwe, Zambia. Email: jamasjsakala@gmail.com
b. Department of Computer Science, ZCAS University, Lusaka, Zambia. Email: douglas.kunda@zcasu.edu.zm

Abstract

COVID-19 is a highly infectious respiratory disease that belongs to the SARS group of viruses that has presented a global challenge to almost everyone world-wide. During the early stages of the pandemic in Zambia, a major challenge was the limited data and datasets for COVID-19. This challenge restricted research, especially in data mining. The challenge of data and datasets is currently improving. This paper presents the challenges of using data mining techniques and models to analyze and forecast the COVID-19 pandemic in Zambia. The analysis initially presents the methodology used for creating a dataset that focuses on the pandemic at provincial scope and uses the Zambia National Public Health Institute (ZNPHI) and Ministry of Health Zambia daily situation reports. The analysis of the pandemic at country level used the COVID-19 datasets from the Humanitarian Data Exchange (HDX) and the European Center for Disease Prevention and Control (ECDC). The study finally discusses the development and evaluation of the forecasting model. The forecasting model is based on the COVID_SEIRD Python package. To evaluate the forecasting model, the research utilized a combination of correlation and the max-function from basic statistics. The analysis focuses on finding the provincial area with the most COVID-19 cases in Zambia, while the forecasting process manages to forecast the trend of the pandemic for recoveries and fatalities.

Keywords: Coronavirus, COVID-19, Health, Data Mining, Analysis, Forecasting, SEIRD, Python, COVID_SEIRD, Zambia.

1. Introduction

The COVID-19 (SARS-CoV-2) pandemic is one of the major health challenges the world is facing today. The virus was first reported by officials in Wuhan City, China in December 2019 [1]. By the end of February 2020 China was considered the epicenter of the disease [2]. Outside mainland China, Italy was one of the first countries with a high number of COVID-19 cases. The first COVID-19 case in Italy was reported on 31st January 2020 and by April 2020 COVID-19 had claimed over 100,000 lives in that country [3]. Today, almost every country worldwide has reported multiple infections and fatalities from the pandemic [1].

COVID-19 is a highly infectious respiratory disease that belongs to the SARS group of viruses [4]. The disease is caused by the SARS-CoV-2 virus. COVID-19 spreads when an infected individual comes into close contact with other individuals, such that the virus enters the uninfected individual's respiratory system. After the virus enters an individual's respiratory system, the virus incubates for approximately up to 14 days [5] before signs of symptoms begin to show. Symptoms of COVID-19 include fever, cough, fatigue and various difficulties in breathing such as shortness of breath. The disease primarily spreads through contact and infects individuals through the mouth, eyes and respiratory systems [6]. The disease takes an average of up to 5 days for infected individuals to move from infection to fatality (Wilson et al. 2020) While recovery from the disease takes a minimum of 14 days [7].

Since the beginning of the COVID-19 pandemic, data mining has been one of the most widely used tools in the fight against the pandemic [8]. Using data mining and machine learning, scientists and researchers in countries such as China, Italy and Spain have been able to analyze and forecast the COVID-19 pandemic, which has enabled them to flatten the curve in those countries [9]. One of the most widely used models in most of this research were the SIR based models [10]. In developing countries, especially those in Africa such as Zambia, one of the major challenges in using data mining and machine learning in the fight against the pandemic is that COVID-19 cases data collection has not been consistent. The lack of consistent data has resulted in uncertainty and limited information on both the current and future of the COVID-19 pandemic [11]. To overcome this lack of datasets, researchers attempted to utilize datasets from different countries to predict the pandemic for other countries. Most such research ultimately resulted in predictions that were largely off the mark. This was particularly true for developing countries, especially those in Africa [12].

The objectives of this study are a) to assess challenges associated with dataset for Covid-19 in Zambia; b) to develop a dataset for COVID-19 cases (infected, recovered, deaths susceptible and exposed) for Zambia; c) to develop and evaluate a forecast model for COVID-19 cases (infected, recovered, deaths susceptible and exposed) for Zambia.

2. Related works

2.1 Compartmental Model for Covid-19

Machine learning and data mining describe various technologies and techniques used to learn and extract useful patterns and information from datasets. They have been used to tackle various challenges such as fraud detection in banks, as well as pandemic modeling and forecasting [13]. Researchers have used various data mining tools and machine learning techniques and tools such as dashboards and models to help in the fight against the COVID-19 pandemic. One of the first and most popular models that emerged during the COVID-19 pandemic was the compartmental models [14]. The compartmental models are a group of machine learning models that are aimed at forecasting the spread of a pandemic based on epidemiology model developed in 1927 by Kermack and McKendrick [15]. Based on 3 differential equations, the Kermack-McKendrick model was designed to model and predict spread of a disease and predict the number of susceptible, infected and recovered individuals [16]. The compartmental models are a collection of epidemiological models that are used to predict and model the spread of a disease in a population. The models generally break up the population under consideration into segments called compartments, and over time individuals move between these segments. In the compartmental models, each segment of the population is typically represented by an equation. The most basic and foundation model in the compartmental models is the SIR model [16].

The SIR model primarily divides the population in a pandemic into 3 main groups (Susceptible, Infected and Recovered or Removed) and then forecasts the size of these groups over time. Susceptible refers to the segment of the population who do not have the disease but have a chance of contracting it while Infected and Recovered refer to segments of the population who are infected or have recovered from the disease, respectively. The SIR model forms the foundation of almost all other models in the compartmental models. Other advanced models generally extend the SIR model by adding other factors to it. One of the more advanced models in the compartmental models is the SEIRD model. The SEIRD segments the population under consideration into segments. In the SEIRD model, the population is segmented into 5 main segments, each of which is represented by a variable. The 5 important variables in the SEIRD model are the infection-rate (β), recovery/removal rate (γ) and, recovery-rate (γ) and, incubation-rate(σ), mortality rate (μ) and the population (N) , mortality rate (μ) and the population (N) and the population (N) [17], [18].

One of the earliest and most promising researches that showed the potential of the compartmental model was the research by [19]. In this work, the researchers were able to provide estimates of case fatality and recovery ratios with a high degree of confidence (about 90%) as the pandemic evolved in its early stages. The researchers in this work utilized the publicly available epidemiological data for Hubei, China from January 11 to February 10, 2020. Additionally, this research also provided an estimated reproduction number (R0), and the per day infection mortality and recovery rates [19].

Another example of one of the earlier research that showed the potential of the compartmental model was the research by [20]. In this research, the researchers' extended the basic compartmental model to create an advanced model collectively termed SIDARTHE. The model considers eight stages of infection: susceptible (S), infected (I), diagnosed (D), ailing (A), recognized (R), threatened (T), healed (H) and extinct (E). The model utilized case data from Italy, and it demonstrated that restrictive social-distancing measures needed to be combined with widespread testing and contact tracing to end the COVID-19 pandemic. Using these measures coupled up with other health measures, Italy was able to impede the effects of the pandemic, which at one point saw it being the epicenter of the pandemic in Europe [20].

Research based on the compartmental models showed a lot of potential for use by developed countries. However, results in developing countries using the compartmental models did not yield the desired results. For example, using publicly available health data on the COVID-19 pandemic in Iran, [21] discovered that the accuracy and suitability of the basic SIR model in predicting the COVID-19 cases in Iran was very low. [21] cites work by other researchers such as [22], who also attempted to utilize the compartmental model to forecast the COBID-19 pandemic with very low accuracy. In his research work, [22] attempted to forecast the COVID-19 in India, South-Korea and Iran using the basic SIR model. This research utilized publicly available health statistics and its results showed that the SIR model was only able to predict the pandemic in those countries with a very low degree of accuracy.

2.3.1 Hybrid Model

[23] presents a prediction model for COVID-9 that combines multiple techniques for calculating the reproduction number. This model termed the "Hybrid Model" utilizes artificial intelligence supported by multiple ideas and concepts including NLP, LSTM and an ISI model as shown in Fig 1.

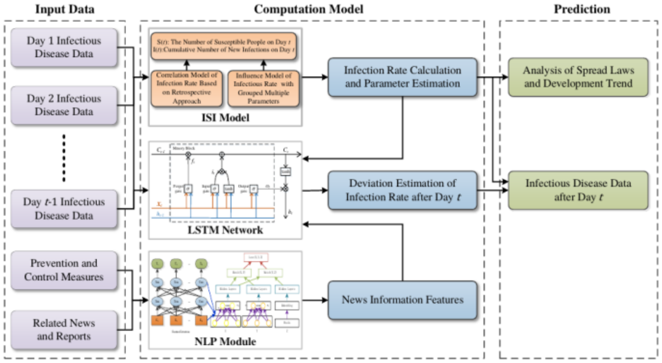

Input Data
Computation Model
Prediction
Day 1 Infectious Disease Data
Day 2 Infectious Disease Data
Day t-1 Infectious Disease Data
Prevention and Control Measures
Related News and Reports
S(t): The Number of Susceptible People on Day t
I(t):Cumulative Number of New Infections on Day t
Correlation Model of Infection Rate Based on Retrospective Approach
Influence Model of Infectious Rate with Grouped Multiple Parameters
ISI Model
LSTM Network
NLP Module
Infection Rate Calculation and Parameter Estimation
Deviation Estimation of Infection Rate after Day t
News Information Features
Analysis of Spread Laws and Development Trend
Infectious Disease Data after Day t

Fig 1 Hybrid Model [23]

This research utilizes the COVID-19 datasets from the from Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) repository. At its core, the model used in this research is a loosely coupled combination of an artificial intelligence (AI) based system composed of a Long

Short Term Memory (LSTM) [24], Natural language Processing (NLP) [25] and Improved Synthetic intelligence model (ISI) [26].

The model works by taking input data from the dataset, and it outputs another dataset. The output dataset is in the form of a time-series based values that consist of predicted cumulative values for susceptible, infected, exposed and recovered cases. Apart from this, the model is also able to predict the value of the reproduction number for COVID-19. This model was specifically designed to be used to predict COVID-19 cases for the country China and is one of the more highly complex models. Most of the complexity in this model arises mostly from the various ideas and concepts and techniques which it combines. A major advantage of this model is the fact that it produces some of the most fairly successful results. A second advantage is that it utilizes an openly available dataset for COVID-19 maintained by the Johns Hopkins University [27]. This data repository has emerged as one most used data repository for COVID-19 datasets by researchers world-wide, and it is also one of the most up-to-date repositories. One of the major main disadvantages of this model is that it is one of the most complex models, and it requires a lot of computing resources and expertise, a large and accurate dataset, and it was designed to be used in a country with a large population and a lot of infections [23].

2.3.2 Regression Based Model

In his work, [28] outlines a COVID-19 prediction model that combines epidemiological models with traditional data mining techniques to predict the COVID-19 pandemic for India. In this work, the researcher utilizes a model that is made up of 2 other models to predict COVID-19 cases for India. The first model is the SEIR model from the compartmental model in epidemiology. The second model is composed of linear and polynomial regressions techniques. The 2 models are used to produce 2 predictions, which are in the form of time-series datasets. This research primarily utilizes the COVID-19 datasets from the JHU CSSE repository and the output datasets produced represent cumulative values for infected, recovered and fatal cases.

This model was specifically designed for and used to predict the COVID-19 cases for India and is one of the simpler models. A major advantage of this model is the fact that it produces some of the most fairly successful results. A second advantage is that it also utilizes an openly available dataset for COVID-19 maintained by the Johns Hopkins University. This data repository has emerged as one most used data repository for COVID-19 datasets by researchers world-wide, and was one of the first repositories created for COVID-19 [29]. It is also one of the most up-to-date repositories. A major disadvantage of this model lies in its use of regression which while being suitable for the Indian COVID-19 dataset which is consistent might not be suitable for non-consistent and fluctuating datasets. This fact can be seen in Figure 2 which shows line plots of SEIR, Regression and actual cases for India from the research work [28].

2.3.3 Improved SIR Model

In his work, [30] presents research that aims to predict the COVID-19 pandemic in Africa. This work focuses on South Africa, Egypt, Algeria, Nigeria, Senegal and Kenya. This work proposes a COVID-19 prediction model based on the SEIR model. The biggest modification to SEIR in this model is the method used to estimate the parameters. Using a technique known as Metropolis-Hasting (MH), [30] model was used to predict COVID-19 statistics for South Africa, Egypt, Algeria, Nigeria, Senegal and Kenya.

This model takes as input a time-series based dataset,, and it utilizes the COVID-19 datasets from the JHU CSSE repository. Similar to other work, this work also produces as output a times-series based values for factors such as infections, recovered and exposed cases.

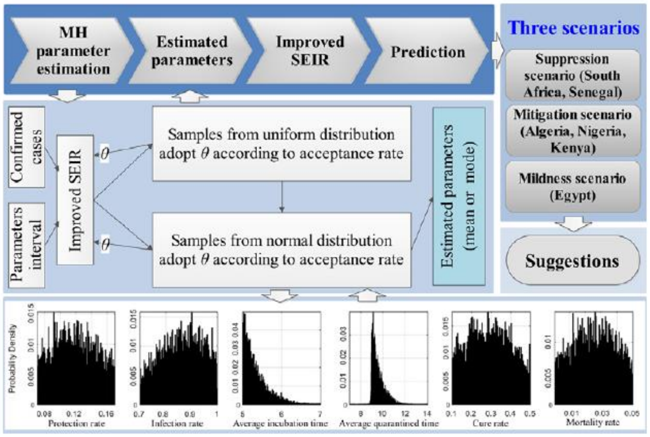

MH parameter estimation
Estimated parameters
Improved SEIR
Prediction
Three scenarios
Suppression scenario (South Africa, Senegal)
Mitigation scenario (Algeria, Nigeria, Kenya)
Mildness scenario (Egypt)
Suggestions
Confirmed cases
Parameters interval
Improved SEIR
θ
Samples from uniform distribution adopt θ according to acceptance rate
Samples from normal distribution adopt θ according to acceptance rate
Estimated parameters (mean or mode)
Probability Density
Protection rate
Infection rate
Average incubation time
Average quarantined time
Cure rate
Mortality rate

Fig 2 Improved SIR Model

A major disadvantage of this model lies in its moderately simple design that utilizes the traditional SEIR model in combination with the more advanced Metropolis-Hasting (MH) parameter-estimation technique. This fact can be seen in Figure 3 which shows a design overview of the model. Another advantage with Zhao's work lies in the way his model presents the results using traditional visualization in the form of bar-graphs and line-graphs [30].

2.3.4 Web News & Social Media Based Prediction Models

In their research, [31] present a COVID-19 prediction model fundamentally based on data and statistics from web news and social media. This work aims to predict the COVID-19 pandemic on a world scale, and it utilizes various sources for its data. Some sources of the data, statistics and metrics utilized by this model include keywords, number of likes, posts, tweets, views and even surveys.

This work by [31] is one of the earliest works in COVID-19 predictions, and it was fairly successful. The first disadvantage of this work was that it mostly utilizes anonymous and unverified values and metrics. A second disadvantage was that the accuracy of predictions decreased when the model was applied on a smaller scale such as at a country or provincial level, as such, the model used is not suitable at country level [31], [32], [33].

Covid-19 Modeling in Zambia

In Zambia, one of the earliest research that used data mining and machine learning to assist in the fight against the COVID-19 was that done by [34]. In their research, Phiri and his colleague attempt to assess the spread of the COVID-19 pandemic using environmental and social economic factors. One of the main similarities between the work by Phiri and colleagues and the one presented in this paper is that both utilize the same data source, the Ministry of Health daily COVID-19 reports. The first similarity between the work by Phiri and colleagues and the work presented in this paper is that both utilize the Ministry of Health daily COVID-19 reports as one of their main sources of data. A second similarity is that the work presented in this paper also utilizes a few similar tools and techniques to those utilized by Phiri and his colleague in their work. This research however, additionally utilizes a classification tree approach to achieve its goals [34].

One of the most recent researches aimed at forecasting the COVID-19 pandemic in Zambia is that done by [35]. In this research, an attempt was made to forecast the third wave of the COVID-10 pandemic in Zambia using the classical SIR model. The research utilizes the same data sources as this paper (ZNPHI and MOH). The results of this research are yet to be compared to the actual data and published [35].

[36] applied a number of existing classifiers available in the Waikato Environment for Knowledge Analysis (WEKA) machine learning library to model Covid-19 infections in Zambia. The classifiers used were ZeroR, OneR, J48 decision tree, Multilayer Perceptron, Naïve Bayes, Random Forest, Support Vector Machine, K Nearest Neighbor and Logistic Regression. WEKA is an open source collection of machine learning algorithms for data mining and contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization [37]. The results indicate that these classifiers performed well and was able provide insight and better understanding of epidemic spread within Zambia. In their development of a cross platform contact tracing mobile application, [38] uses deep neural networks to determine contacts in proximity to a Covid-19 positive case. The deep learning model has been evaluated against analytic models and machine learning models. They discovered that deep neural network model performed better than analytic and traditional machine learning models during testing.

3.0 Methodology

The methodology for this study consisted of five steps. The first step was creation of the dataset and the second step was analysis and validation of the dataset. Third step was an experiment aimed at creating a model using the SEIRD model. Fourth step was an experiment that combined the SEIRD model with other data mining techniques. The third and fourth steps were both aimed to arriving at a model that would produce value of cases that are close to the ECDC and HDX dataset. The final step was evaluation of the forecasting model.

3.1 Creation of the dataset

The study used four data sources to create the dataset for this research namely the Zambia National Public Health Institute (ZNPHI), European Center for Disease Prevention and Control, GitHub, and Johns Hopkins University Center for Systems Science and Engineering. The daily and periodic COVID-19 situation reports published by ZNPHI were used to create a COVID-19 dataset that focuses on the pandemic for Zambia at provincial scope. The dataset created was used to analyze the COVID-19 pandemic at provincial level for Zambia. The time series data on COVID-19 from the European Center for Disease Prevention and Control (ECDC) and Humanitarian Emergency Response Africa (HERA) were used to analyze the pandemic at country level. These datasets include datasets for many countries. This study however only focuses on the dataset for Zambia. The analysis focuses on infections, fatalities and recoveries [39], [40]. The geospatial shape files for Zambia obtained from GitHub repository were used in the analysis of the pandemic at provincial and country levels. The shapes are used to create heat-maps during analysis [41].

The global time series data on COVID-19 from JHU CSSE was used to forecast future cases (infections, fatalities and recoveries) for Zambia. The JHU CSSE COVID-19 dataset includes datasets for many countries. This study however, only focuses on the dataset for Zambia [42], [43]. The study utilized Python 3, Tesseract-OCR, Camelot and Excalibur, GImageReader, Bash Scripting, pdfgrep, pdftk and wget to extract data from the above sources. The creation of the dataset involved extraction, loading and transformation (ELT). The extraction process primarily involved using Optical-Character-Recognition tools and utilities to extract data from the daily and periodic situation reports by the Zambia National Health Institute. The Transformation process involved reducing the scope of all the data to provincial level, while the loading process involved saving the data. The ultimate dataset was in the form of a single XLSX spreadsheet file composed on 4 sheets. The first 3 sheets contain infected, recovered and fatal case values, while the third sheet contains notes about the dataset.

3.2 Analysis of the dataset.

The analysis process utilized classical data analysis tools and techniques. The primary tools and techniques utilized are pie-charts, bar-graphs, line-graphs, and geospatial heat-maps and data tables.

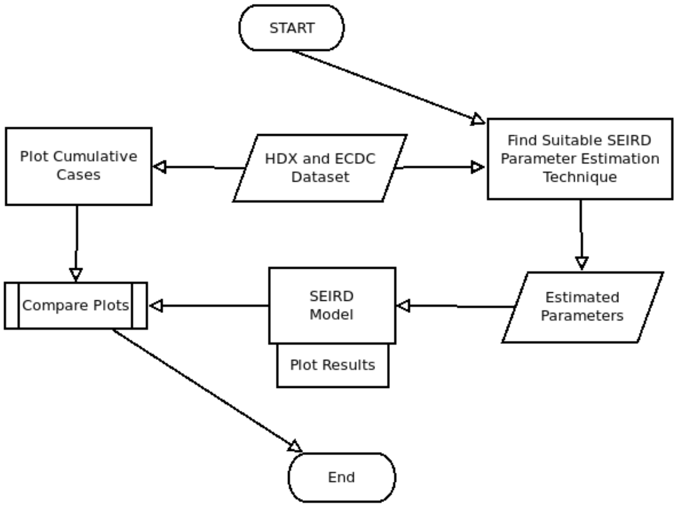


Fig 3. Creating a model using the SEIRD

3.3 Experiment 1 (Creating a model using the SEIRD)

The first experiment was aimed at creating a model using the SEIRD. The first step during this experiment focused on finding a suitable parameter-estimation technique (see Fig 3). The three main techniques that were used were the least-square technique, the variable reversal mathematical technique and the exponential-curve fitting technique. For each of the techniques, all its results were used with the SEIRD model, and a line graph plotted.

The variable-reversal technique attempts to estimate the values of beta and gamma by moving the variables beta and gamma to the LHS of the equation was implemented using the equation in Fig 4. This experiment focused on estimating those variables on a monthly basis. The least-square technique was then used to estimate the values of beta and gamma by finding the value of a slope between 2 arbitrary points on a plot of the actual data was implemented. This experiment focused on estimating those variables by picking 2 arbitrary points that represent 2 days between the data.

3.4 Experiment 2 (SEIRD model with data mining techniques)

The second experiment aimed at using an existing implementation to create a forecasting model for COVID-19 for Zambia. Two python packages that were considered for this purpose were CovsirPhy [44] and covid_seird [45]. CovsirPhy is a package that implements the SIR epidemiology model while covid_seird is a package that implements the SEIRD model (Fig 5). The "covid_seird" package implements the SEIRD model and includes tried and tested parameter estimation techniques built into the library. The covid-seird package was selected for this experiment.

$$\frac{dS}{dt} = -\beta \cdot I \cdot \frac{S}{N}$$
$$\frac{dE}{dt} = \beta \cdot I \cdot \frac{S}{N} - \delta \cdot E$$
$$\frac{dI}{dt} = \delta \cdot E - (1-\alpha) \cdot \gamma \cdot I - \alpha \cdot \rho \cdot I$$
$$\frac{dR}{dt} = (1-\alpha) \cdot \gamma \cdot I$$
$$\frac{dD}{dt} = \alpha \cdot \rho \cdot I$$

Fig 4 SEIRD Equations used by the Forecasting Model [46]

The experimentation process begins by calling the "country_covid_seird.CountryCovidSeird()" function from the "covid_seird" with the country code "ZM" for Zambia. The function returns an object for Zambia. The "best_fit()" method is then called on the Zambia object to utilize the "best fit curve" method to create a best fit curve for Zambia using the JHU CSSE dataset as shown in Fig 5. After calculating the parameters using the best-fit technique, The plot_simulation is called which plots a line plot of the results of the SEIRD model. After plotting a SEIRD model, the experiment process proceeds to create a line plot of infected, recovered and fatal cases from the simulation.

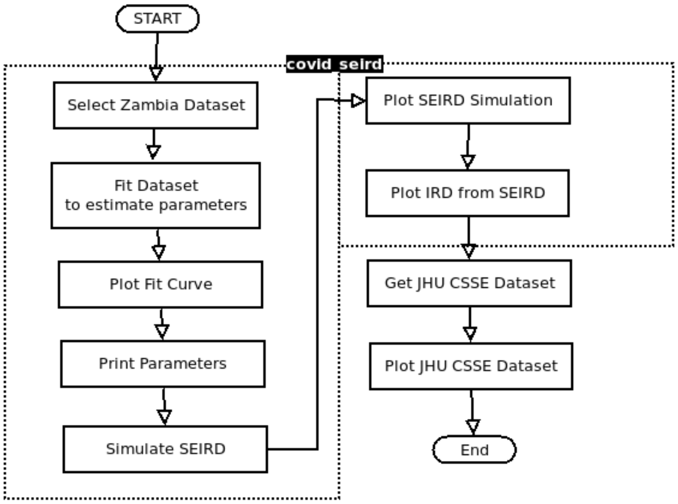


Fig 5 Experiment 2 process

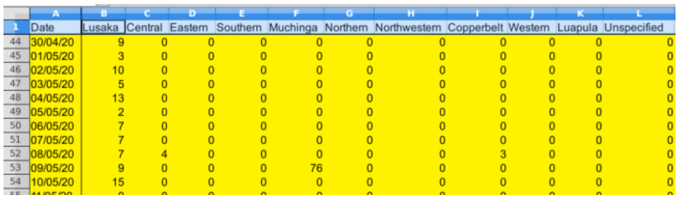
3.5 Model evaluation

To evaluate the forecasting model, the values it produced were compared with data from the JHU CSSE datasets. The comparison utilized visualization to accomplish this task. The visualization involved comparing the values for the infected, recovered and fatal cases from the period March 18, 2020 to June 2021 with those generated by the forecasting model. A confusion matrix was used to assess the performance of the forecasting model.

4.0 Results and discussion

4.1 Dataset creation

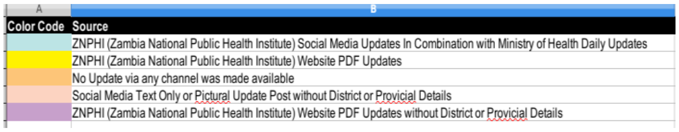
The extraction loading and transformation process involved downloading reports from the ZNPHI website and transcribing them to create a day-wise tally of the COVID-19 cases in Zambia. Due to gaps in the situation reports published by ZNPHI, their social media situation update posts in partnership with MOH are combined with the situation reports. The extraction of data from the situation reports and posts was done using optical character recognition tools to minimize errors and time for the process. The results of the ELT process are in the form of an Microsoft excel sheet file in which infected, recovered and fatal case statistics have their own sheet and each sheet contains a tally for each province.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Date	Lusaka	Central	Eastern	Southern	Muchinga	Northern	Northwestern	Copperbelt	Western	Luapula	Unspecified
44	30/04/20	9	0	0	0	0	0	0	0	0	0	0
45	01/05/20	3	0	0	0	0	0	0	0	0	0	0
46	02/05/20	10	0	0	0	0	0	0	0	0	0	0
47	03/05/20	5	0	0	0	0	0	0	0	0	0	0
48	04/05/20	13	0	0	0	0	0	0	0	0	0	0
49	05/05/20	2	0	0	0	0	0	0	0	0	0	0
50	06/05/20	7	0	0	0	0	0	0	0	0	0	0
51	07/05/20	7	0	0	0	0	0	0	0	0	0	0
52	08/05/20	7	4	0	0	0	0	0	3	0	0	0
53	09/05/20	9	0	0	0	76	0	0	0	0	0	0
54	10/05/20	15	0	0	0	0	0	0	0	0	0	0

Fig 6 Results of ELT process

The dataset created from the ELT process had challenges and limitations, for example some data had mixed or missing scope. Some of the daily and periodic reports contained data that had different scopes. Some cases were reported at district level while others were at provincial level. This scope challenge was overcome by aggregating the district to provincial level. Small subset of the reports contained case data that did not have area (district or province) scope. This data was put into an “unspecified” scope label as shown in Fig 6. A minimal number of days did not have any situation published or posted by either ZNPHI or the Ministry of Health via any channel. These days were outlined in the “NOTES" sheet of the Excel Sheet file for the created dataset as shown in Fig 7.



A	B
Color Code	Source
🟦	ZNPHI (Zambia National Public Health Institute) Social Media Updates In Combination with Ministry of Health Daily Updates
🟨	ZNPHI (Zambia National Public Health Institute) Website PDF Updates
🟧	No Update via any channel was made available
🟫	Social Media Text Only or Pictural Update Post without District or Provicial Details
🟪	ZNPHI (Zambia National Public Health Institute) Website PDF Updates without District or Provicial Details

Fig 7 NOTES

4.2 Analysis of dataset

Fig 8 shows a line graph created from the HDX and ECDC datasets. The line graph shows monthly infected and recovered totals for Zambia from January 2020 to 8 January, 8 2021. From this line graph it can be seen that the COVID-19 pandemic in Zambia arrived around March 2020 but within 2 months the cases began to rise and 4 months later the cases appear to reach an all-time high (around August 2020). Following this spike, the figures drop again in November before spiking again around January 2021.

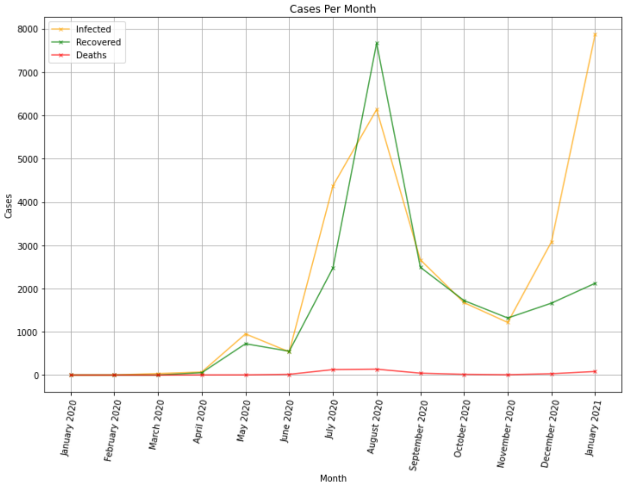


Fig 8 Monthly fatalities and recoveries

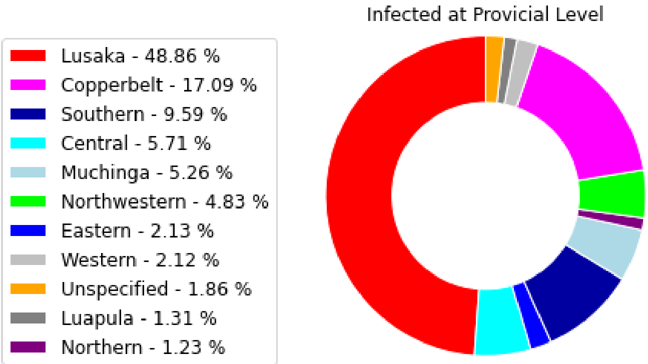
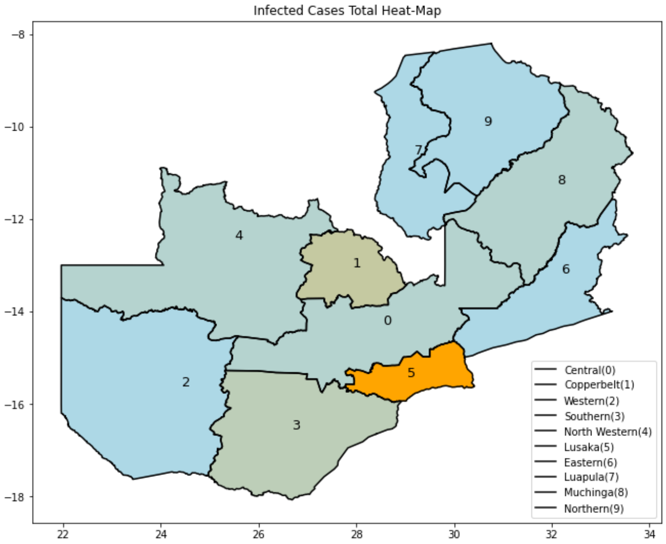


Fig 9 Infected Cases Heat-Map (Jan 2020 – Jan 2021)

Fig 8 also shows a line graph of total monthly fatalities and recoveries created from the HDX and ECDC datasets for Zambia. An important point that can be seen from this graph is that the fatalities for Zambia due to COVID-19 generally seem to be very low while the recovery figures seem to always be close to the infections.



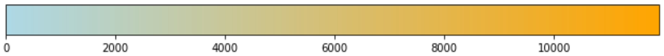


Fig 10 Infected Cases Heat-Map (Jan 2020 – Jan 2021)

Fig 9 shows a pie-chart and Fig 10 shows a heat-map of infections per province for the time period January 2020 to December 2020 created from the HDX and ECDC datasets. As can be seen from both the pie-chart and heat-map, the majority of infections for Zambia appear to be in the urban areas (Lusaka and the Copperbelt provinces in particular). As can be seen from Fig 8 and Fig 9, the same areas that have high infections also have matching high recoveries and low fatalities.

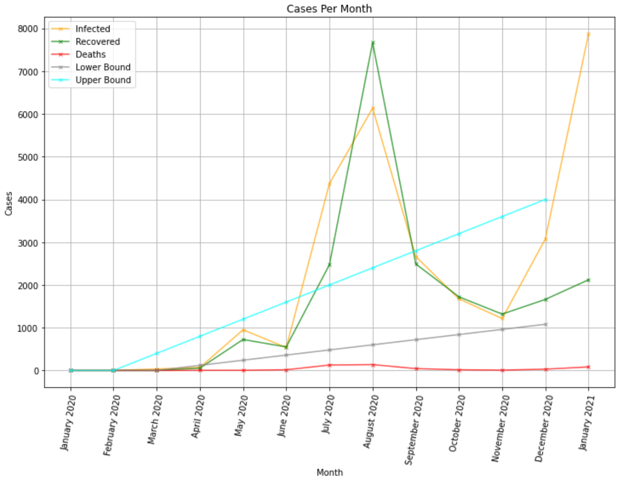


Fig 11 Infected and recovered cases

Fig 11 shows monthly infected and recovered totals for Zambia from January 2020 to January, 2021. The results show that the trends for infections and recoveries is that from around March 2020 to around November 2020, the numbers for recoveries seem to closely follow the infections. After November, however, the recoveries seem to start lagging behind infections. This information is important for decision making

4.3 Experiment 1 and 2

The results of variable-reversal technique for the month of September is shown in Fig 12. The technique was used to estimate the values of beta and gamma by moving the variables beta and gamma to the LHS of the equation. The value of beta is 6.54118, value of gamma is 0.0079 and value of R0 is 8.2667 for September.

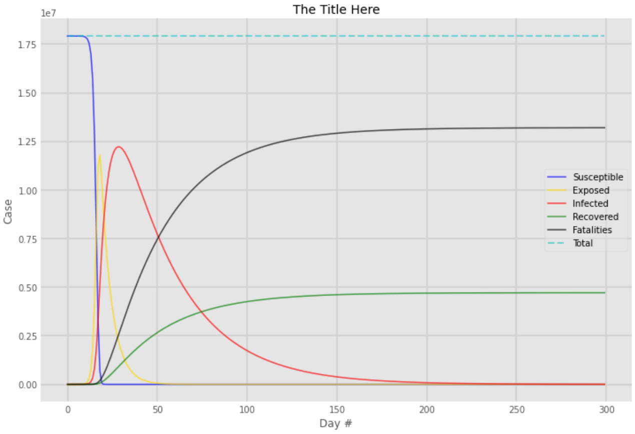


Fig 12 Variable-reversal technique for September

The results of the exponential-curve-fitting technique is shown in Fig 13. The technique attempts to estimate the values of beta and gamma by plotting a curve that best fits the actual data that was implemented. This technique used Point A at 60 and Point B at 360 with beta equal to 0.031, gamma equal to 1.0 and R0 equal to 0.032.

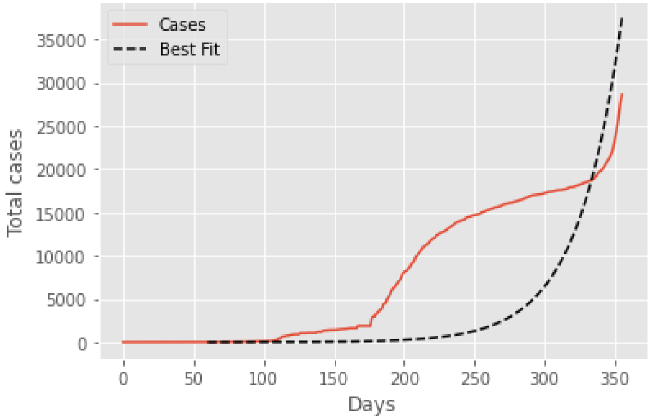


Fig 13 Exponential-curve-fitting technique

The results of the least-square technique are shown in Fig 14. The technique attempts to estimate the values of beta and gamma by finding the value of a slope between 2 arbitrary points on a plot of the actual data. The value of gamma is 0.00482878, value of beta is 0.0218498 and R0 is 4.52491.

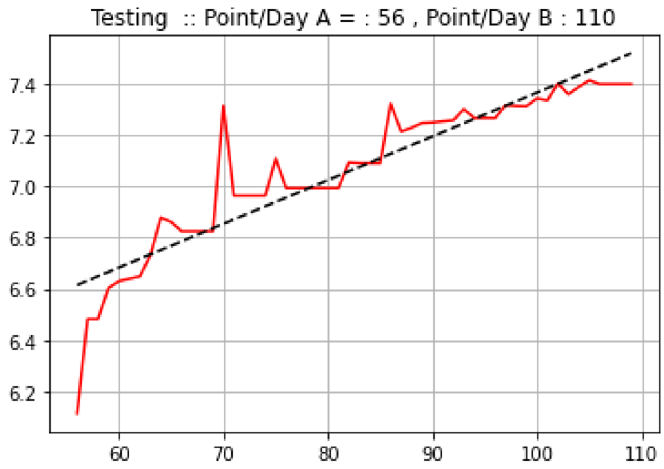


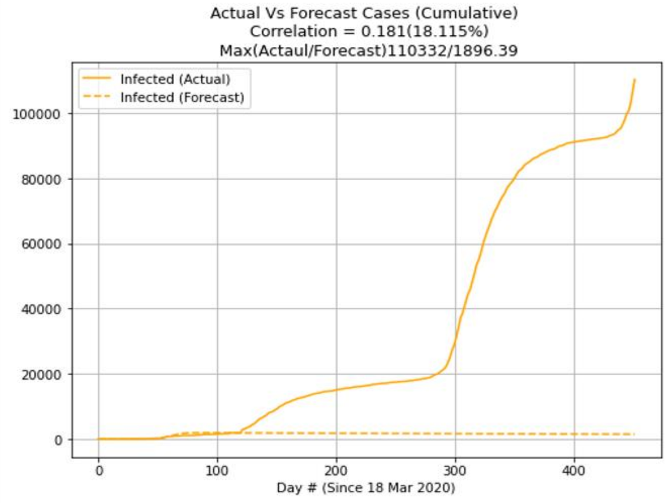
Fig 14 Least-square technique

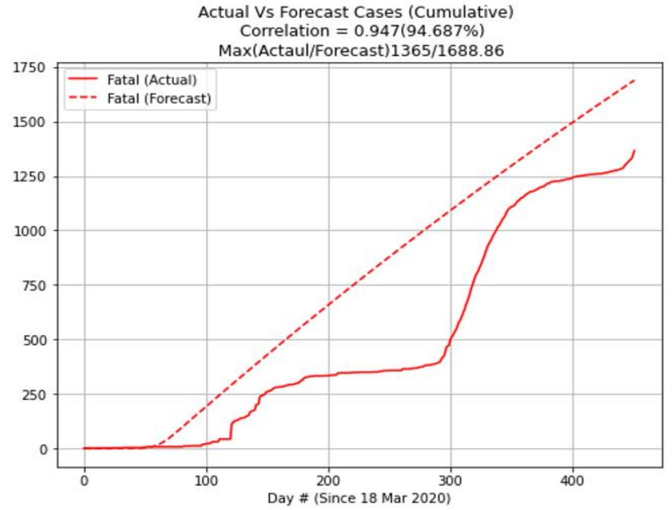
As can be seen from the results, the least-square and curve-fitting techniques produced parameters that resulted in SEIRD plots that were not close to fitting the actual HDX and ECDC datasets. The equation-reversal technique on the other hand, while being the most promising, was also producing parameters that resulted in SEIRD plots that were not fitting with the HDX and ECDC datasets.

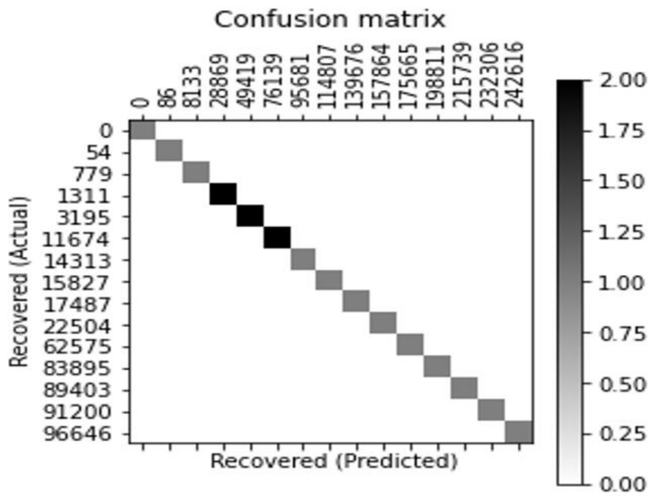
4.4 Forecasting model evaluation

The values generated by the forecasting model with infected, recovered and fatal cases were plotted on 3 line graphs as shown in Fig 15. The dotted lines represent the forecasted values, while the solid lines are for actual values. On top of each graph is a simple correlation coefficient between the 2 lines in each graph plot as well as values for the maximum of each plotted line.

From all 3 line-graphs, it is evident that the model's forecast values do not match up perfectly with the actual values. This fact is backed up by the difference between the maximum values for each graph. From the same graphs however, it can be seen that the correlation coefficient values for fatal and recovered cases are well above 50% while that for infections is not. The confusion matrices for the forecasted data show the most significant differences in the infected cases while the recovered cases also show a difference and the fatal values showing the most promising alignment. The model evaluation suggests that while the model is not able to forecast the actual cumulative values for infections, it is able to forecast the trend for fatalities and recoveries.







6.1 Conclusion

This study started out with the intention of analyzing and forecasting the COVID-19 pandemic in Zambia using data mining techniques and models. The analysis process began by creating a data mining dataset for COVID-19 for Zambia that focused on the provincial scope. From the analysis performed, various trends can be seen. These trends include the provinces that have the most infections, recoveries and fatalities. Other trends include the progression of cases at a country level from its onset. Another objective of the research was to create a forecasting model for COVID-19 for Zambia. To achieve this task, the research focused on experimentation to create a model based on the SEIRD epidemiology model.

While many arguments and points can be raised about the work in this study, the potential benefits and areas where data mining can be applied in the fight against COVID-19 in Zambia can also be seen. This point is very important when one considers the fact that the current COVID-19 pandemic is clearly going to be around for a while and no single country has so far managed to eradicate it and prevention still remains as the best solution. To implement preventive measures however requires constant analysis and forecasting of the pandemic in order for preventive measures to be effective.

Further research could concentrate on utilizing other pandemic analysis models such as time-based models or even utilizing the same ideas used in this study but supplementing them with other data mining models and techniques such as polynomial and hybrid regression. This is especially true for COVID-19, as it is an ongoing challenge and no analysis or forecasting model so far has been found that is suitable for all situations and especially for developing countries like Zambia.

REFERENCES

[1] W. H. O. WHO, “Coronavirus disease 2019 (COVID-19) Situation Report – 73,” Situaton Report, Apr. 2020. Accessed: Aug. 01, 2020. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200402-sitrep-73-covid-19.pdf

[2] W. Xu, J. Wu, and L. Cao, “COVID-19 pandemic in China: Context, experience and lessons,” *Health Policy Technol.*, vol. 9, no. 4, pp. 639–648, Dec. 2020, doi: 10.1016/j.hlpt.2020.08.006.

[3] E. Livingston and K. Bucher, “Coronavirus Disease 2019 (COVID-19) in Italy,” *JAMA*, vol. 323, no. 14, pp. 1335–1335, Apr. 2020, doi: 10.1001/jama.2020.4344.

[4] A. S. Fauci, H. C. Lane, and R. R. Redfield, “Covid-19 — Navigating the Uncharted,” *N. Engl. J. Med.*, vol. 382, no. 13, pp. 1268–1269, Mar. 2020, doi: 10.1056/NEJMe2002387.

[5] D. Baud, X. Qi, K. Nielsen-Saines, D. Musso, L. Pomar, and G. Favre, “Real estimates of mortality following COVID-19 infection,” *Lancet Infect. Dis.*, vol. 20, no. 7, p. 773, Jul. 2020, doi: 10.1016/S1473-3099(20)30195-X.

[6] T. P. Velavan and C. G. Meyer, “The COVID-19 epidemic,” *Trop. Med. Int. Health*, vol. 25, no. 3, pp. 278–280, Mar. 2020, doi: 10.1111/tmi.13383.

[7] D. Loconsole *et al.*, “Recurrence of COVID-19 after recovery: a case report from Italy,” *Infection*, vol. 48, no. 6, pp. 965–967, Dec. 2020, doi: 10.1007/s15010-020-01444-1.

[8] A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, “Artificial intelligence and machine learning to fight COVID-19,” *Physiol. Genomics*, vol. 52, no. 4, pp. 200–202, Mar. 2020, doi: 10.1152/physiolgenomics.00029.2020.

[9] Z. Ceylan, “Estimation of COVID-19 prevalence in Italy, Spain, and France,” *Sci. Total Environ.*, vol. 729, p. 138817, Aug. 2020, doi: 10.1016/j.scitotenv.2020.138817.

[10] M. Gatto *et al.*, “Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 19, pp. 10484–10491, May 2020, doi: 10.1073/pnas.2004978117.

[11] D. L. Katharine Houreld, “In Africa, lack of coronavirus data raises fears of ‘silent epidemic,’” *Reuters*, Jul. 08, 2020. Accessed: May 29, 2022. [Online]. Available: https://www.reuters.com/article/us-health-coronavirus-africa-data-insigh-idUSKBN24910L

[12] J. Muhaidat, A. Albatayneh, R. Abdallah, I. Papamichael, and G. Chatziparaskeva, “Predicting COVID-19 future trends for different European countries using Pearson correlation,” *Euro-Mediterr. J. Environ. Integr.*, May 2022, doi: 10.1007/s41207-022-00307-5.

[13] A. L. Fradkov, “Early History of Machine Learning,” *IFAC-Pap.*, vol. 53, no. 2, pp. 1385–1390, Jan. 2020, doi: 10.1016/j.ifacol.2020.12.1888.

[14] N. A. Kudryashov, M. A. Chmykhov, and M. Vigdorowitsch, “Analytical features of the SIR model and their applications to COVID-19,” *Appl. Math. Model.*, vol. 90, pp. 466–473, Feb. 2021, doi: 10.1016/j.apm.2020.08.057.

[15] E. F. Doungmo Goufo, R. Maritz, and J. Munganga, “Some properties of the Kermack-McKendrick epidemic model with fractional derivative and nonlinear incidence,” *Adv. Differ. Equ.*, vol. 2014, no. 1, p. 278, Oct. 2014, doi: 10.1186/1687-1847-2014-278.

[16] F. Brauer, “The Kermack–McKendrick epidemic model revisited,” *Math. Biosci.*, vol. 198, no. 2, pp. 119–131, Dec. 2005, doi: 10.1016/j.mbs.2005.07.006.

[17] A. E. Martianova, V. Y. Kuznetsova, and I. M. Azhmukhamedov, “Mathematical Model of the COVID-19 Epidemic,” Nov. 2020, pp. 63–67. doi: 10.2991/assehr.k.201105.012.

[18] J. Fernández-Villaverde and C. I. Jones, “Estimating and Simulating a SIRD Model of COVID-19 for Many Countries, States, and Cities,” National Bureau of Economic Research, w27128, May 2020. doi: 10.3386/w27128.

[19] C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos, “Data-based analysis, modelling and forecasting of the COVID-19 outbreak,” *PLOS ONE*, vol. 15, no. 3, p. e0230405, Mar. 2020, doi:

10.1371/journal.pone.0230405.

[20] G. Giordano *et al.*, “Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy,” *Nat. Med.*, vol. 26, no. 6, pp. 855–860, Jun. 2020, doi: 10.1038/s41591-020-0883-7.

[21] S. Moein *et al.*, “Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan,” *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, Feb. 2021, doi: 10.1038/s41598-021-84055-6.

[22] I. Cooper, A. Mondal, and C. G. Antonopoulos, “A SIR model assumption for the spread of COVID-19 in different communities,” *Chaos Solitons Fractals*, vol. 139, p. 110057, Oct. 2020, doi: 10.1016/j.chaos.2020.110057.

[23] S. Du *et al.*, “Predicting COVID-19 Using Hybrid AI Model,” Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3555202, Mar. 2020. doi: 10.2139/ssrn.3555202.

[24] T. Fischer and C. Krauss, “Deep learning with long short-term memory networks for financial market predictions,” *Eur. J. Oper. Res.*, vol. 270, no. 2, pp. 654–669, Oct. 2018, doi: 10.1016/j.ejor.2017.11.054.

[25] A. Jain, Department of Computer Engineering, SVKMs NMIMS MPSTME Shirpur, Maharashtra, India, G. Kulkarni, Department of Computer Engineering, SVKMs NMIMS MPSTME Shirpur, Maharashtra, India, V. Shah, and Department of Computer Engineering, SVKMs NMIMS MPSTME Shirpur, Maharashtra, India, “Natural Language Processing,” *Int. J. Comput. Sci. Eng.*, vol. 6, no. 1, pp. 161–167, Jan. 2018, doi: 10.26438/ijcse/v6i1.161167.

[26] J. Asadi and F. Tarokh, “How The Artificial Intelligence is Helping to Cure: The Analytic Means of Health in Current Pandemic World,” p. 6, 2020.

[27] G. F. Ficetola and D. Rubolini, “Containment measures limit environmental effects on COVID-19 early outbreak dynamics,” *Sci. Total Environ.*, vol. 761, p. 144432, Mar. 2021, doi: 10.1016/j.scitotenv.2020.144432.

[28] Gaurav Pandey, Poonam Chaudhary, Rajan Gupta, and Saibal Pal, “[2004.00958] SEIR and Regression Model based COVID-19 outbreak predictions in India,” 2020. https://arxiv.org/abs/2004.00958 (accessed Jun. 02, 2020).

[29] M. Miller, “2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository,” *Bull. - Assoc. Can. Map Libr. Arch. ACMLA*, no. 164, pp. 47–51, Mar. 2020, doi: 10.15353/acmla.n164.1730.

[30] Z. Zhao, X. Li, F. Liu, G. Zhu, C. Ma, and L. Wang, “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,” *Sci. Total Environ.*, vol. 729, p. 138959, Aug. 2020, doi: 10.1016/j.scitotenv.2020.138959.

[31] K. Jahanbin and V. Rahmanian, “Using twitter and web news mining to predict COVID-19 outbreak,” *Asian Pac. J. Trop. Med.*, p. 4, 2020.

[32] E. Chen, K. Lerman, and E. Ferrara, “COVID-19: The First Public Coronavirus Twitter Dataset,” *ArXiv200307372 Cs Q-Bio*, Mar. 2020, Accessed: Jun. 03, 2020. [Online]. Available: http://arxiv.org/abs/2003.07372

[33] Wim Naudé, “Artifcial intelligence vs COVID-19: limitations, constraints and pitfalls,” *AI & SOCIETY (2020) 35:761–765*, Apr. 2020, doi: https://doi.org/10.1007/s00146-020-00978-0.

[34] D. Phiri, S. Salekin, V. R. Nyirenda, M. Simwanda, M. Ranagalage, and Y. Murayama, “Spread of COVID-19 in Zambia: An assessment of environmental and socioeconomic factors using a classification tree approach,” *Sci. Afr.*, vol. 12, p. e00827, Jul. 2021, doi: 10.1016/j.sciaf.2021.e00827.

[35] M. Mwale, B. Kanjere, C. Kanchele, and S. Mukosa, “Simulation of the Third Wave of COVID 19 Infections in Zambia using the SIR Model,” In Review, preprint, Feb. 2022. doi: 10.21203/rs.3.rs-1337105/v1.

[36] Josephat Kalezhi, Mathews Chibuluma, Christopher Chembe, Victoria Chama, Francis Lungo, Douglas Kunda (2022) Modelling Covid-19 infections in Zambia using data mining techniques, Results in Engineering, Volume 13, https://doi.org/10.1016/j.rineng.2022.100363.

[37] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[38] Josephat Kalezhi, Mathews Chibuluma, Christopher Chembe, Victoria Chama, Francis Lungo and Douglas Kunda, “A Cross Platform Contact Tracing Mobile Application for COVID-19 Infections using Deep Learning” International Journal of Advanced Computer Science and Applications (IJACSA), 13(8), 2022. http://dx.doi.org/10.14569/IJACSA.2022.0130872

[39] European Centre for Disease Prevention and Control, “Homepage | European Centre for Disease Prevention and Control,” 2021. https://www.ecdc.europa.eu/en (accessed Jun. 30, 2021).

[40] HERA, “HERA - Humanitarian Emergency Response Africa - COVID-19 Project,” *HERA - Humanitarian Emergency Response Africa - COVID-19 Project*, 2021. https://hera-ngo.org/ | https://data.humdata.org/hxlproxy/data/download/time_series_covid19_recovered_global_narrow.csv | https://data.humdata.org/hxlproxy/data/download/time_series_covid19_confirmed_global_narrow.csv | https://data.humdata.org/hxlproxy/data/download/time_series_covid19_deaths_global_narrow.csv (accessed Jul. 01, 2021).

[41] Lighton Phiri, “lightonphiri - Overview,” *GitHub*, 2021. https://github.com/lightonphiri/data-zambia-shapefiles/archive/master.zip (accessed Jun. 30, 2021).

[42] CSSEGISandData, *CSSEGISandData/COVID-19*. 2021. Accessed: Jun. 30, 2021. [Online]. Available: https://github.com/CSSEGISandData/COVID-19

[43] JHU CSSE, “JHU CSSE,” 2021.

https://systems.jhu.edu/tracking-covid-19/ | https://github.com/CSSEGISandData/COVID-19 (accessed Jun. 30, 2021).

[44] Takaya, H. (n.d.). covsirphy: COVID-19 data analysis with phase-dependent SIR-derived ODE models (2.20.3) [Python]. Retrieved June 10, 2021, from https://github.com/lisphilar/covid19-sir

[45] Branco, P. (n.d.). covid-seird: A small package that implements the SEIRD Epidemiological Model on COVID-19 data. (0.0.8) [Python; OS Independent]. Retrieved June 10, 2021, from https://github.com/paulorobertobranco/covid_seird

[46] H. Froese, “Infectious Disease Modelling: Beyond the Basic SIR Model,” *Medium*, Apr. 22, 2020. https://towardsdatascience.com/infectious-disease-modelling-beyond-the-basic-sir-model-216369c584c4 (accessed Apr. 17, 2021).