RESEARCH THESIS

Surrogate Assisted by Genetic Algorithm for Simulation of Hospital-resource Allocation: Lockdown Policy Analysis for COVID-19 Pandemic Case

Author:
Timilehin OGUNDARE
(2373256)

Supervisor:
Prof. Terence VAN ZYL

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

submitted to
the Faculty of Science, in fulfilment of the requirements for the degree of
MSc in Computer Science And Applied Mathematics

in the

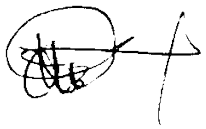
Wits Institute of Data Science (WIDS)
School of Computer Science and Applied Mathematics

October 27, 2021

Declaration of Authorship

I, Timilehin OGUNDARE (2373256), declare that this research thesis titled, "Surrogate Assisted by Genetic Algorithm for Simulation of Hospital-resource Allocation: Lockdown Policy Analysis for COVID-19 Pandemic Case" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this research thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this research thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the research thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 17-05-2022

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Abstract

Faculty of Science
School of Computer Science and Applied Mathematics

MSc in Computer Science And Applied Mathematics

Surrogate Assisted by Genetic Algorithm for Simulation of Hospital-resource Allocation: Lockdown Policy Analysis for COVID-19 Pandemic Case

by Timilehin OGUNDARE (2373256)

The government and healthcare sector plays a major role in managing the transmission upsurge and unpleasant effects of COVID-19. The dynamics of the pandemic is an emerging question for these parties, as this will guide their planning and strategy implementation of disease prevention and treatment policy. The emphasis of this study is to develop a simulation-based system that can estimate the future transmission dynamics of COVID-19. The simulation will predict the current and future trends of COVID-19 in South Africa, Italy, India, Mexico, and Fiji (SIIMCF). The prediction will assist hospital management with capacity planning and assist government in policy formulation around lockdown measures. Agent-Based Modeling and Simulation (ABMS) is an excellent computational tool for this simulation process. However, there are several constraints to using Agent-Based Modeling and Simulation (ABMS) to simulate a real-world event. Such constraints include the duration of the time required to perform simulation, the need for High performance computational resources, and the complexity of parameter tuning. Our study consider surrogate model as a substitute for Agent-Based Modeling and Simulation (ABMS). Our findings show the surrogate can overcome the constraints identified by ABMS for simulating the COVID-19 pandemic. Further, we optimize the surrogate with the Genetic Algorithm (GA) by running a local search and extracting the best hyperparameter values for the surrogate. The optimization of the surrogate will enhance the quality of its results when compared using real-world COVID-19 datasets. To evaluate the performance of the ADRIANA system, we used three evaluation metrics, which are, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2).

Publications

The work in this dissertation has been published in the following conference proceedings:

- Timilehin *et al.* 2021 **Ogundare Timilehin and Terence Van Zyl.** Surrogate Parameters Optimization for Data and Model Fusion of COVID-19 Time-series Data. In 24^{th} International Conference on Information Fusion, South Africa, 2021, (Arxiv link: http://arxiv.org/abs/2109.04207).

Acknowledgements

Special thanks to the Almighty **God** for granting me the privilege and resources needed to accomplish this research, as He has said:

*...that by two immutable things, in which it was impossible for God to lie, we might have a strong consolation, who have fled for refuge to lay hold upon the hope set before us (**Heb 6:18**)*

I would also want to dedicate this thesis to my late Father (OGUNDARE Jacob), and my entire family for their unwavering support throughout the course of this study.

My supervisor Prof. Terence Van Zyl, deserves special recognition. He has been a fantastic and supportive supervisor. He provided his full supervision, time, resources, and guidance. I really appreciate you Prof. I would not have imagine another supervisor for this study.

Contents

List of Figures

List of Tables

List of Abbreviations

DES	**D**iscrete **E**vent **S**imulation
COVID-19	**C**orona **V**irus **D**isease 2019
WHO	**W**orld **H**ealth **O**rganization
SARS	**S**evere **A**ccute **R**espiratory **S**yndrome
PPE	**P**ersonal **P**rotective **E**quipments
DSS	**D**ecision-making **S**upport **S**ystem
HRA	**H**ospital **R**esource **A**llocation
CI	**C**onfidence **I**nterval
SEIR	**S**usceptible **E**xposed **I**nfectious **R**ecovered
SIRS	**S**usceptible **I**nfectious **R**ecovered **S**usceptible
SIR	**S**usceptible **I**nfectious **R**ecovered
LSTM	**L**ong **S**hort **T**erm **M**emory
ICU	**I**ntensive **C**are **U**nit
CRD	**C**onfirmed **R**ecovered **D**eath
SIIMCF	**S**outh Africa **I**taly **I**ndia **M**exico **C**olombia **F**iji

Chapter 1

Introduction

1.1 Background

The COVID-19 pandemic has undoubtedly posed a significant economic and health-related concern to many parts of the world. COVID-19 emerged as an outbreak in Wuhan, China, and has spread to become a pandemic as declared by the World Health Organization on 11^{th} March, 2020 [1]. According to a report published by Johns Hopkins University [2], the overall number of confirmed cases, recovered cases, and death cases globally are 148 million, 85.7 million, and 3.12 million respectively as of 24^{th} April, 2021, with the first fatality case announced on 10^{th} January, 2020 [1]. The top five countries in the world in terms of confirmed cases as of 30^{th} March, 2021 are the United States of America, India, Brazil, France, and Russia [1], [2].

Below is a definition of epidemic and a pandemic:

> *An epidemic is referred to as the rapid spread of infectious diseases within the same demographic group. However, a pandemic represents the rapid spread of an infectious disease across different countries and continents with more population being affected (- WebMD [3]).*

As highlighted by the above definition, COVID-19 disease has infiltrated countries globally, regardless of their social development status as a developed, developing, or under-developed country [1], [4]. The affected countries have established various committees, structures, and initiatives to investigate different components of the pandemic. The global organization in charge of monitoring COVID-19 is the World Health Organization (WHO) [1]. Likewise, some of the organizations that monitor the dynamics of the COVID-19 in South Africa includes: National Institute for Communicable Diseases (NICD) [5], Africa Centres for Disease Control and Prevention [6], Department of Health in South Africa [7], and World Health Organization - South Africa division [8].

Due to the severe impact of the disease on the health sector and economic imbalance, research into the computational and medical studies of the disease has been

a key emphasis. One of the negative effects of the disease is its effect on the world economy. Even in the financial sector, the damage caused by COVID-19 has led to the worst trading day for the FTSE 100 (-9.10%) and S&P 500 (-9.5%) since 1987 [4].

One major strategy for limiting the virus spread is the awareness of governments and healthcare sector about the expected number of confirmed cases. Following that, policymakers and hospital administrator can devise and implement a workable preventive measures.

The treatment of COVID-19 patients could place an undue strain on hospital resources due to high transmission rate. Therefore, hospitals should have effective means of administering treatment resources, because of dynamics of hospital operational structure, and novelty in diseases dynamics (for example, COVID-19 is different from SARS and Ebola in features).

As a result, providing recommendations based purely on prior experiences is challenging because there are no general case-specific findings applicable to all hospitals. In addition, the dynamic interrelationships between various factors in hospital pandemic planning must be considered. For example, changing the number of ventilators or beds affects staffing, pharmaceutical demands, PPE, etc [9].

The majority of COVID-19-affected countries adopted severe lockdown[1] measures at the initial stage of the disease outbreak [4], [10]. Likewise, patients in the hospital who satisfy clinical and epidemiological feature of COVID-19 are quickly isolated. For instance, the first lockdown in South Africa was imposed from 26^{th} March, 2020 to 16^{th} April, 2020 [5] in an increasing degree of strictness ranging from level-1 to level-5.

Likewise, in India, the first lockdown was enforced from 25^{th} March, 2020 to 14^{th} April, 2020 [2], [4]. The daily number of confirmed cases in South Africa reduced from 11.8% percent to 6.3% percent as a result of the lockdown measures [4], [10]. A government would be unethical to enforce lockdown on the country indefinitely, as this can cause a nationwide economic slump [5], [6], [11]. As a result, the government must understand when to raise, lessen, or completely discontinue lockdown procedures.

Furthermore, there are implicit uncertainties in patient inflow, patient duration in the hospital, and hospital operations. Therefore, the relationship between these elements is monitored using a random process rather than a deterministic approach [9]. These factors strongly contribute to the degree of accuracy of prediction and evaluation of model performance. Moreover, several existing frameworks for forecasting COVID-19 and hospital demand capacity do not consider the operational constraints of their model's flexibility and feedback mechanisms [9], [12].

[1] Lockdown is an emergency protocol that requires population groups to remain indoors for essential or social reasons.

As a result, we aim to create a simulation-based decision-making system that can integrate the aspects highlighted above and answer our research questions, which will be discussed in a later section of this dissertation.

Several research projects have been conducted to provide scientific answers to forecast the future of COVID-19. Researchers have used statistical models, conventional machine learning models, mathematical models, epidemiology models, and deep learning techniques. For instance, the work of Ahmar and Val [13] use ARIMA and Sutte-ARIMA to forecast COVID-19 prevalence and stock market variables in Spain, arriving at a mean absolute percentage error (MAPE) evaluation of 3.6%. In addition, Ceylan *et al.* [12] used ARIMA to predict the confirmed cases of COVID-19 in France, Spain, and Italy with a MAPE range between 4% to 6%. In Chapter 2, we offer an in-depth overview of the current findings in this research area.

In this study, we examined implementing a simulation-based system using an Agent-Based Modeling and Simulation (ABMS) technique. Agent-Based Modeling and Simulation (ABMS) are computer models that attempt to capture the behavior responses of several agents in a simulation environment. With ABMS, the action and role of the agents within the system must be known and defined, as well as the interconnections that exist between the agents [14].

Over the last few years, ABMS has found relevance in simulation experiments. Exploring an ABMS is a good option for our study. However, after carefully examining the computational cost of implementing an ABMS, the inherent implications do not appear suitable for proposing a solution to COVID-19 because of several considerations. Some issues identified are, implementing ABMS is computationally expensive since it needs long simulation times, parameter tuning is complex, challenges in model validation, and the inability to perform effectively with homogeneous data, which is a key feature of the datasets analyzed in this study [14], [15].

To address the computational cost that is involved in using ABMS, we developed a simulation model that can model the function (of an ABMS) and perform the operation of an ABMS in a short period. As a result, our study explores the development of a simulation model named ADRIANA, which is made up of three different simulation models namely the SEIR compartmental model, discrete event simulator, and an optimized surrogate model. We aim to fuse these models and their individual results into the final ADRIANA. We also verify that the predictions of the ADRIANA system are similar to real-world data. Chapter 3 of this dissertation contains details about the structural components, formulation, and mechanism of each model in the ADRIANA system.

In developing the ADRIANA system, we explored the Genetic Algorithm (GA) as the optimization framework that searches the hyperparameter space for the optimal values for the surrogate. GA is a universal optimizer that researchers have

proven to be a robust optimizer for model development [16].

1.2 Problem Statement

The COVID-19 pandemic should be carefully controlled by authorities of affected regions. We considered South Africa (for Africa), Italy (for Europe), India (for Asia), Mexico (for North America), Colombia (for South America), and Fiji (for Oceania) which we abbreviate as SIIMCF in this dissertation[2].

Managing the spread of the COVID-19 disease requires a deliberate effort on the part of government to recommend preventative and precautionary measures for residents. As a result, they should thoroughly investigate the dynamics surrounding the pandemic to establish a realistic approach to reduce disease transmission.

It is critical to run a simulation that reflects the activities of the COVID-19 pandemic agents, while predicting future transmission pattern. As mentioned, we consider a substitute to ABMS to avoid its drawbacks. Hence, we develop and optimize a surrogate, also termed a machine learning model that can perform similarly to the ABMS.

The surrogate is a less computationally intensive data-driven framework that can be executed in a short time with minimal human interaction. The optimization phase is important in surrogate development. Optimization can assist machine learning algorithms to optimally detect trends in historical datasets and support algorithm's predictions by learning patterns from datasets. A data-driven surrogate can reflect time-variant trends in datasets without an explicit understanding of the elements that contribute to the pattern. We use a Genetic Algorithm (GA) as an optimizer to extract optimal hyperparameters for a surrogate model.

1.2.1 Why is Optimization Important in Surrogate Development?

The primary objective of a surrogate model is to predict an occurrence with a high level of precision. We can only accomplish this through an optimization process. Therefore, surrogate model optimization is accomplished by adjusting and tuning the hyperparameter values to minimize inconsistency between the observed and predicted values.

During the training stage, the model's parameters are determined. Examples are biases and weights for neural networks. They are the internal component of a model, and their values vary depending on the input values. Contrary, hyperparameters are assigned with initial values before the model is trained. Examples of hyperparameters are learning rates, number of clusters, number of estimators, tree depth, etc. In

[2] Abbreviation for the six countries explored in this study (South Africa, Italy, India, Mexico, Colombia, and Fiji, with South Africa as the benchmark country).

general, hyperparameters determine the structure and output of a model. The best hyperparameters are obtained through hyperparameter optimization, while the best model parameters are obtained through model training.

Therefore, hyperparameter optimization is important for tuning the surrogate to optimal standard. Obtaining the ideal hyperparameter values can contribute to a decrease in the model error and an improvement in model accuracy. Some of the optimization techniques include, Genetic Algorithms, gradient descent, and exhaustive search. Figure 1.1 depicts the processes required in determining the best design methods for an optimization process.

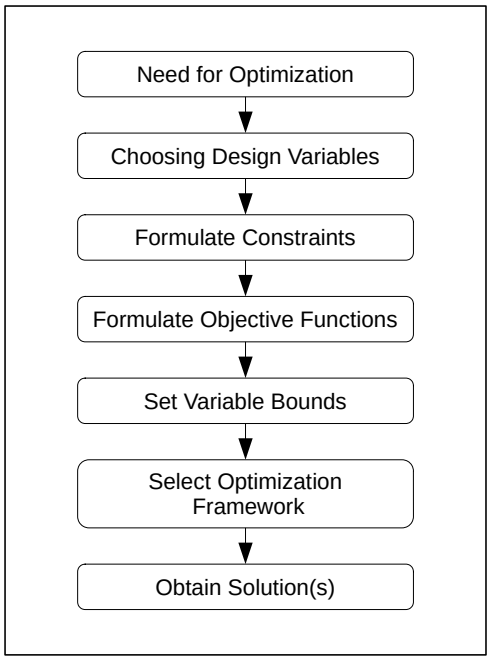


FIGURE 1.1: Flowchart Showing Optimal Design Procedures [17]

1.3 Purpose of this Study

The intention of our study was to:

- estimate the length of a surge in hospital capacity and resource demand using an SEIR epidemiology model, a discrete event simulation, and an optimized surrogate model. In addition, we predicted future transmission pattern of COVID-19 in South Africa and the other SIIMCF countries, to capture the third wave of the disease in South Africa and present a global perspective of the pandemic for the next 13 months (starting from April 24, 2021);
- use the prediction to inform and enhance the decision of hospital management and policymakers concerning lockdown measures; and

- assist various agencies such as hospital management, economic advisors, social advisers, governments, in developing an ideal course of action that can absorb the effect of COVID-19 diseases in their respective domains of operation.

Current research in this line is largely focused on models centered on statistical, epidemiological, and mathematical techniques. We extend this by forecasting using an optimal neural networks architecture. Additionally, we investigate some models that have been developed to respond computationally to the COVID-19 pandemic.

1.4 Research Questions

To provide an optimal computational solution that will address the research problem and accomplish the purpose of our study. The following research questions are being investigated:

1. Compared to using the ABMS to simulate a pandemic case, how effective are surrogate models in estimating the future transmission dynamics of COVID-19 with a global perspective?
2. Optimizing the surrogate hyperparameters is an essential procedure. How far can we justify this optimization of the surrogate to drive forecasts?
3. To what degree can the ADRIANA system be calibrated and its predictions used to handle future pandemic prediction problems?

1.5 Research Methodology Overview

Our analysis forecasted the transmission curve of COVID-19 in South Africa and the SIIMCF countries using both simulated and an economic open-source dataset explained further in Chapter 4 of this dissertation. At the time of writing, South Africa is anticipating the third wave of the virus, which was appropriately captured in this study.

1.5.1 Data Collection

The quantitative datasets that were used for our research were extracted from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [18]. The dataset is presented in a time-series format which contains the daily cumulative statistics of COVID-19 confirmed cases, recovered cases, and death cases (CRD) for 191 countries starting from 22^{nd} *January*, 2020

till day X_{t-1}, where X_t is the current day. We then extracted the subtractive daily values for the CRD cases for day n by differencing the cumulative value of day$(n-1)$ from the cumulative value of day(n). That is, daily value for day(n) = (cumulative value for day n) - (cumulative value for day $n-1$). Given that the dataset is a time-series, it is important that the datasets be made stationary [19].

We proceeded to scale the datasets to stationarity. We explored this to improve the training process of the surrogate model. After scaling the datasets, we trained and predicted using three optimal neural network architectures, five conventional regression machine learning models, and two statistical regression models. The performance hierarchy of the models is shown in Figure 4.33. We provide the techniques that make up the ADRIANA system in Chapter 3, while we report the results obtained in Chapter 4 of this dissertation.

1.6 Case Study

Our research primarily focuses on predicting the transmission dynamics of COVID-19 in South Africa. In addition, we explore the same technique to predict the future trends of COVID-19 in five(5) other countries which we referred to as SIIMCF, with each country representing a continent.

1.7 Significance and Motivation of this Study

It is important to provide strategies that can help the prevention of infectious disease in populations. By extension, these policies are based on insights provided by applicable frameworks that are concerned with infectious diseases.

We are driven to do this study as part of the scientific community because of the historical contributions of computational modelers in solving real-world problems. Secondly, there is a need for an accurate COVID-19 disease forecast mechanism that can advise action plans to address the continuous spread of the disease globally, which was captured in our study.

1.8 Research Contributions

This study contributes to the growing domain of knowledge on applying machine learning optimization techniques in forecasting pandemic cases. This research aims to build an accurate simulation of the COVID-19 pandemic case and observe the rate of infections and recoveries, which will inform the future trend of the disease and be useful to policymakers when planning preventive measures. Our findings will make the following contributions:

(i) Use optimized machine learning techniques, time-series modeling approach, and event simulation tool to forecast the transmission curve of the COVID-19 pandemic at a global scale by considering one country form each continent of the world. This study evaluates the fusion of conventional machine learning, neural networks, compartmental models and statistical regression tools to forecast the transmission dynamics of the COVID-19 pandemic.

(ii) Develop a discrete simulation of COVID-19 patient procedures in the hospital based on their acute level. This projection informs hospital management about peak durations of confirmed cases, and the quantity of resources needed to optimally treat COVID-19 patients.

(iii) Inform policymakers about the best time to reduce or increase lockdown measures, and other response actions during a pandemic outbreak. Additionally, we observe the implication of being pragmatic in enforcing lockdown measures on the population as suggested by the ADRIANA system.

1.9 Dissertation Structure

This chapter is an introductory overview of the study. It discusses the problem statement, purpose of the study, objectives of the research and the research questions. The remaining section of this dissertation is organized as follows. Chapter 2 reviewed and identified the limitations of existing studies that address COVID-19 pandemic case. In Chapter 3, we detailed the methodologies, datasets, implementation processes, and how we arrived at our results. In Chapter 4, we discuss the simulation results by the appropriate representations, and how our research questions have been answered by the results. Finally, Chapter 5 presents a summary of the whole research process, findings, recommendations, and suggested improvement to this study.

Chapter 2

Research Background and Related Works

2.1 Introduction

This Chapter attempts to demonstrate, analyze, and discuss the outcomes of existing computational modeling-based studies used to estimate the transmission dynamics of COVID-19. This will let us explore some key analyses that will help our study.

We conducted a bibliographic review for two main purposes. First, we explore publications on COVID-19 to extract computational insights, and secondly, we identify other study areas around COVID-19. General observation presents medicine, computational science, and operations research as the three most significant study areas regarding COVID-19. This has resulted in many research publications from diverse domains of knowledge and specialization. We presented these publications categorization in Figure 2.1.

Several studies have presented scientific solutions and simulation experiments that can support the decisions of COVID-19 stakeholders. For example, hospitals treating COVID-19 patients must have a structured insight to inform their plan for sufficient resources planning to treat and manage infected individuals.

Furthermore, government policymakers must have a clear understanding of the disease's transmission trend in their specific geographic location. The majority of COVID-19 research has sought to determine if there is a general element that governs COVID-19 transmission in all concerned countries around the world. These studies suggest that there are various factors associated to the transmission dynamics of COVID-19 in their respective geographies [20], [21].

Hence, developing a simulation for COVID-19 is strongly recommended to capture solutions and forecast disease transmission patterns. Majority of simulations

in this research area use a hybrid of discrete event simulation and optimization approaches to arrive at a near-optimal hyperparameters value representation of COVID-19's stakeholders, such as the healthcare system. Simulating and predicting the future development of the COVID-19 pandemic scenario necessitates a thorough comparison and evaluation of the result obtained from statistical tools, deep learning frameworks, and conventional machine learning techniques [22].

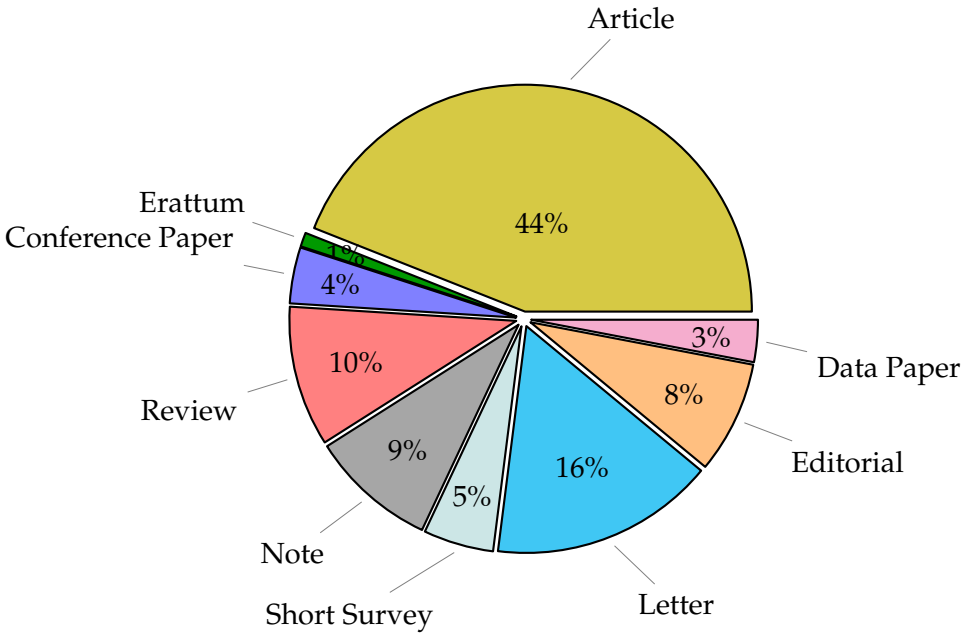


FIGURE 2.1: Categorizing COVID-19 Publications Based on Focus Area According to Scopus Database [23]

2.2 Notable Findings in Existing Studies

Current research that focuses solely on statistical regression models (such as the Auto-Regressive (AR), Sutte-ARIMA, Moving Average (MA), and Integrated-ARMA (ARIMA)) for predicting COVID-19 transmission rate and hospital resource demand, has been proven to be relatively dependent on assumptions [24], which are:

(i) the variance of error terms in model architecture which accumulates with time [19].

(ii) the inclusion of incorrect coefficients for estimating unknown model parameters which can produce wrong estimates of R^2 values during the model assessment.

To decode the change in variance of error terms, a non-linear set of models, such as neural network models is required.

Alkhamis *et al.* [25] developed a discrete-event simulator and optimization framework that formulates into a decision-making support system (DSS). The DSS recommends strategies to minimize the waiting time of COVID-19 patients at Kuwaiti hospitals. Based on their simulation prediction, their findings reported a 40% reduction in COVID-19 patient waiting time and a 28% improvement in hospital service delivery. Similarly, Yeh Lin *et al.* [26] proposed a simulation model and a GA-based optimization technique that predicts an optimal schedule for Taiwanese hospital nurses during COVID-19 capacity treatment planning.

A study conducted by Andrew Crane *et al.* [9] demonstrates a system that estimates COVID-19 transmission trends and simulates hospital resource capacity demands. They created a discrete event simulator and used the SIR framework as the epidemiology model. Their investigation sampled key analyses for model parameters across a 6-day doubling period, i.e. $(T_d = 6)$ [9]. They obtained a 95% confidence interval (CI) with 97.5% and 2.5% percentile estimates using 1000 simulation cycle. Their major output is a prediction that shows the total number of ward beds, number of ventilators, number of ICU beds, number of daily admissions, and peak demand time of hospital resources as a function of the transmission forecast.

Furthermore, research undertaken by Kunar *et al.* [22] explores the COVID-19 patterns in Canada. They created a regression model that predicts the transmission rates and used the SIRS epidemiology framework to group the population into epidemiological compartments. They also used a wavelength transformation approach [27] on the datasets before data fitting and simulations to minimize random noise and conserve time-frequency features. Their findings showed a linear trend in future transmissions of COVID-19. Their simulation model was primarily built on an LSTM neural network which was validated with testing datasets with an accuracy of 92.67% and a RMSE of 45.70 for long-term predictions. They were able to estimate 70-day future predictions of COVID-19 infections and recoveries in Canada. Their results were used to guide the hospital administrators of expected demand for hospital resources.

2.3 Assumptions and Deductions from Existing Studies

Investigating existing studies relevant to model development concerning COVID-19, we observe the assumptions made by the researchers and also unpack their results as of the time the research was conducted.

2.3.1 Assumptions of Existing Studies

(i) Population groupings are defined explicitly, and the number of susceptible populations does not decline evenly towards zero.

(ii) While compartmental models have a short prediction time scale, birth and death cases (other than those caused by COVID-19 disease) can be ignored, and the fatality count caused by COVID-19 is a minimal percentage of the total population.

(iii) Asymmetric features in the historical datasets used for model development causes computation errors.

(iv) The total population N is not constant, but susceptible populations are considered an adjustable parameter for generating newly infected individuals.

(v) The temporal progression of populations and key parameters influencing disease spread is identified.

2.3.2 Deductions of Current Studies

(i) Using mainly statistical models to forecast the future transmission of COVID-19 cases raises the risk of biased predictions.

(ii) Most simulation models in the current study included graphical interface to enable easy-to-use hyperparameter tuning. The graphical interface was incorporated because respective geolocation where simulation will be applied has distinct and unique features which are considered in the model development.

(iii) Due to the dynamic response rate of the different population groups in responding to policies guiding disease spread, the results of computational models present approximated results. Population dynamics motivates researchers to use SIR, SIRS, SEIR, SEIRD, etc., for their epidemiological simulations.

Choosing a key performance indicator (KPI)[1] that defines the dynamics of disease transmission in the COVID-19 simulation process is a challenge for researchers, because of the notion that there is no set rule for determining KPI [20]. However, researchers in this research area were able to develop features that constitute an ideal KPI for the simulation of COVID-19. Such KPIs include the number of discharged patients, the number of unattended-to patients who returned home, the average waiting time of patients in the hospital, patient risk predictors based on characteristics (such as age, previous disease infection, etc.), and the level of restricted mobility.

It is worth noting that 88% of the countries have had their researchers work on COVID-19 transmission and prevalence dynamics [24]. Researchers established that statistical methods did not perform adequately in estimating the transmission trend

[1]Process of measuring the performance of our model activity, by combining different factors into a production process to provide correlated information.

of COVID-19 [22]. In addition, the results of non-optimized models underperform the optimized models [21], [24]. However, assessing the performance of COVID-19 simulation models has resulted in some observations (such as a low R^2 value) in several West African countries [21]. Specifically caused by inconsistencies in reporting daily cases, which resulted in uneven presentation of the confirmed, recovered, and death cases of COVID-19.

2.4 Identified Concerns with Conventional Forecast Models

This section discusses the general challenges that have been encountered with conventional unoptimized models for pandemic forecasting.

- Due to the non-linear feature of the analyzed data, a significant amount of model tuning is needed to get a satisfactory prediction probability. The model tuning process itself is complex because the researchers are expected to have a reasonable insight into the structure of the dataset.
- When there are random processes or other societal variables that contribute to data capture, the models are vulnerable to undetected errors. By implication, when compared against recently observed values, this might result in outlier data.
- Datasets are typically expected to fulfill a certain data standard, such as non-null values, outliers, stationarity, and white noise, among other features for proper modeling, and particularly regression analysis, which may not hold in practice.

Although optimization-simulation approaches have been explored in many COVID-19 simulation studies, a typical issue is the computation cost, particularly when considering the complexity of implementing ABMS. Therefore, meta-models, also known as surrogates, have been embraced and deployed by researchers to obtain satisfactory approximation results without undertaking time-consuming and costly simulations [22], [28].

In Table 2.1, we provide the primary literature evaluated by our study in assessing the existing studies that predict the demand and allocation trend of hospital resources for treating COVID-19, and estimating the future transmission pattern. In addition, the study focus, simulation modules, aims, and important discoveries of existing research are presented in Table 2.1.

TABLE 2.1: Literature Showing Existing COVID-19 Pandemic Simulation Models

Reference	**Research Focus**	**Objectives**	**Simulation Modules**
Andrew Crane, Gary M.D *et al.* - (CHIME) [9]	Disease Transmission Forecast	Obtaining an Optimum HRA Measurement	SEIR, and DES
Epidemic Forecasting in Canada - Kunar *et al.* [22]	Disease Transmission Forecast	Using AI to Predict COVID-19 Future Occurrence	Regression, SIRS Model
Bernard and Douglas - (Epidemic Rationizer) [20]	Hospital Resource Management	Rationing Hospital Resources Amongst Infected Patients	DES
Mihaela Van Der Schaar - (Adjutorium) [29]	Disease Transmission Forecast	Using Machine Learning to Manage Limited Hospital Resources	DES, ABMS, and Surrogate
Cleo, Lucia *et al.* [21]	Disease Transmission Forecast	Obtaining an Optimal Future Prediction of COVID-19 Based on Previous Pandemic Database Data	Regression, ARIMA, SIRD Model
Andrea, Franco *et al.* [28]	Investigating the Challenges in Predicting the Spread of COVID-19 Pandemic	Comparing the Results of Exponential Model and Compartmental Models to Predict COVID-19 with focus on R_0	SIR Model

Legend for Table 2.1:

- HRA - Hospital Resource Allocation
- SEIR - Susceptible Exposed Infectious Recovered
- DES - Discrete Event Simulation
- AI - Artificial Intelligence
- SIRS - Susceptible Infectious Recovered Susceptible
- ARIMA - Auto-Regression Integrated Moving Average
- SIRD - Susceptible Infectious Recovered Death

Note: The researchers presented the performance measures of these models as an average performance due to some factors, such as the best result of the models are extracted based on testing performance across different countries.

We identified three noteworthy challenges in the examined literature:

- Feature estimation - It was established by analyzing the environmental factors that influence the reproduction number R_0 for prediction models, which were mostly simulated with variations of the epidemiological models, such as SIRD, SIQED, SIS, SIRS, etc. Government policy plays a notable role around the mobility of the population and their willingness to adhere to the policy.
- Variation of model simulation: The nature and after-effects of COVID-19 are dependent on many parameters (such as individual response, hygiene, individual immune resistance levels, control measures, etc.) of the affected location; therefore, using a single model to simulate COVID-19 transmission globally may cause an inaccurate subsystem. Researchers should investigate the characteristics that affect their location during the development of their model.
- Uncertainties - Data distribution for model training and development is continuous. Data is presented daily from history till $day(n) - 1$, where $day(n)$ is the present day. Therefore, there is possibility of data sentiment. The effect can be bias model predictions, which reduces the performance and accuracy of the model.

2.5 Discussion and Conclusion

We examine the existing study which explores conventional machine learning models, statistical techniques, epidemiological models, and deep learning neural networks. These models have been applied to simulate hospital resource demand and allocation for COVID-19 infected individuals. Additionally, we examined the prediction models that estimate future transmission of COVID-19 in various countries, and their shortcomings noted.

Using machine learning models in analyzing various tasks, particularly in the simulation of COVID-19 occurrences, has increased among researchers. Furthermore, we discovered that one of the key benefits of applying machine learning is its capacity to capture and simulate the interaction between complex input and output representations of historical data. Empirically, we concluded that machine learning can be applied to simulate the dynamics of the COVID-19 pandemic irrespective of the location.

The data description, quantitative models, research techniques, and components involved in the development of the ADRIANA simulation system are all thoroughly addressed in the next chapter of this dissertation.

Chapter 3

Overview of Research Framework

3.1 Introduction

In this chapter, we examined the methodologies that were followed in our study. We also detailed the approach and the results that align with our research questions. For the simulation experiments, we expanded the procedures, datasets, and model assessment techniques used to achieve our study goal.

3.2 Study Motivation

At the time of writing this dissertation, COVID-19 has exposed the global community to different problems [7], [8]. Furthermore, there are demand dynamics for hospital resources in different countries, based on the ratio of confirmed cases to available resources in hospitals, and increases in disease transmission within the population [1], [7]. Hence, it is vital to have a computational framework that can help address uncertainties in the determining the future trend of COVID-19 globally, and measures the expected demand level of hospital resources required to treat COVID-19 patients.

Using machine learning to simulate various pandemic occurrences is a contemporary method that has produced acceptable results globally [9], [29]. Previous studies have applied neural network architecture, kernel methods, regression analysis, decision trees, and time series models to achieve similar objectives as evidenced in Chapter 2 of this dissertation. Hence, the current work aims to integrate epidemiological compartmental models, discrete event simulation, and machine learning into a single system to make recommendations around lockdown policy and hospital resources management. The forecast accuracy of a model is essential to achieve effective policy and resources management during COVID-19 pandemic. As a result, different intervention structures have been established by the governing authorities of the affected regions to reduce COVID-19 spread [8].

3.3 Research Strategy

COVID-19 virus is an unprecedented event with a global impact [7], [8]. Since the first reported South African case in March 2020, South Africa has experienced fluctuating transmission pattern of the disease. Interestingly, the documented instances of the virus are not on the decreasing trend as at time of this report [5]–[7]. There was a first wave and a second wave (Table 4.5 and Figure 4.14), and there is a high probability projection for the occurrence of third wave [1], [5], [8]. Consequently, COVID-19's transmission rate within the population largely determines the lockdown measures and patient's population in hospitals [5].

This study follows an exploratory process guided by previous works in which we investigated existing models to investigate the events around COVID-19. We formulated the model fusion of the ADRIANA system as the interrelation between three sub-models outlined in Section 1.1 in Chapter 1 of this dissertation. The operation and structural components of ADRIANA sub-models are explained below.

- The SEIR compartmental model is used to simulate the number of susceptible (S), exposed (E), and infectious (I) persons in the population. In essence, considering a population sample N, the SEIR simulation estimates the number of people who can be infected, die, or recover from the disease. Hospitals can apply the estimates of the compartmental model for capacity planning during a pandemic surge. Different parameter configurations are also applied to SEIR model to evaluate the model's projection to different configurations.

  The number of infected individuals who recovered or died from the disease was simulated by SEIR model depending on criteria such as an individual's risk score, previous incidence of disease, age, and response to therapy. The compartmental model's internal operational parameters are initial disease prevalence, incubation rate, recovery rate, population size, and reproduction number. For this study, the SEIR model made a significant assumption that all fatalities occurred in hospitals, and none occurred before patients were brought to hospitals for treatment.

- The ADRIANA system's discrete event simulation (DES) module modeled the sequence of operations and activities that an infected individual undertakes in the hospital. We based our event simulation experiment on several factors that contribute to the operation of the DES. For example, when an infected person visited the hospital, we established factors (such as age group and health history) to identify whether the patient's health status is moderate, severe, or critical. Therefore, we can determine the patient's admission to either ward

or ICU. In our simulation, individuals in moderate infection stage may be discharged without being admitted to the hospital. Likewise, those in the severe stage were admitted to the ward, and those in the critical stage were admitted to the ICU.

The disease incubation duration for COVID-19 was estimated to be between 2 and 5 days, and an average duration of stay in the hospital for a COVID-19 patient was between 1 to 27 days [1]. Based on these parameter configuration, the DES was able to simulate the entire sequence of patient's admission processes in the hospital. Also, we estimated the initial population of the DES as the number of infectious population (I) simulated by the SEIR model.

- The surrogate model was used to learn transmission pattern from dataset to predict the future transmission trend of COVID-19. The benchmark surrogate for this study is LSTM neural network. Our research explored the optimization of surrogates to ensure optimal configuration of hyperparameters.

The compartmental model, DES, and surrogate are all interrelated in this study. The compartmental model grouped the population into epidemiological compartments and obtained additional details of the population to estimate recoveries or fatalities, which determines healthcare burden. In addition, we used the compartmental model to analyze the forecast of the surrogate model since the compartmental model estimated the total confirmed, recovered, and death cases from a population distribution N. Further, we identified the SEIR parameter configuration that generated values similar to the surrogate result, which enabled us to reproduce the compartmental results for different population samples.

In essence, the surrogate analyzes past real-world data to predict future disease transmission pattern. We use the GA to improve the surrogate result by searching and selecting the best hyperparameter suggested by the GA. The goal of the GA's cost function is to minimize the root mean squared error (RMSE) as the loss function during the hyperparameter search process.

3.3.1 Agent-Based Modeling and Simulation (ABMS)

Agent-Based Modeling and Simulation (ABMS) is a computation modeling method that has attracted recognition in modeling experiment in the last few decades [30]. An Agent-Based Model and Simulation (ABMS) is a framework in which agents continually interact together in the same environment [31]. For example, agents improve their mutual action through information exchange in swarm or ant colony optimization [32]. Their interaction is influenced by achieving an optimal end-state within the environment in which several agents are connected [33].

As detailed by Gary *et al.* [15], a typical ABMS-formulated system comprises three main parts. They are (1) the agents, their characteristics, and behaviours in the system, (2) a protocol establishing the interaction mechanism between the agents, and (3) the environment in which the agents operate.

3.3.1.1 When is ABMS Useful in Research Work?

The application of ABMS to research is based on, but not limited to, when:

(1) It is essential to interpret the natural concept of a research idea to understand the purpose and goals of the study. (2) It is important to analyze how well a system interpret and simulate event processes to ensure reliability and comprehensiveness of the system. (3) A research methodology must make predictions based on the dynamic interactions of agents within the system. (4) There is a need to test for dynamic "what-if?" scenarios when agents' patterns of behaviour in the system are altered.

3.3.1.2 Understanding the Agent in ABMS

To recognize an agent in modeling, it must have some inherent characteristics, which include, but are not limited to: the ability to reside within its environment, ability to observe, learn and represent its behavior based on the sensorial knowledge of the environment, ability to be autonomous, and possess high interaction with other agents in the environment. In general, whether the agents are human, autonomous, or organizational, they are the decision-actors in an ABMS-based system. The agent representation in an ABMS system is shown in Figure 3.1.

ABMS is used in various disciplines, including system science, computer science, traditional modeling, simulation studies, and other research fields [14], [15], [34]. Table 3.1 provides different application areas of ABMS and their summaries.

3.3.1.3 Benefits and Critique of an ABMS - In relation to our study

Considering the benefit that ABMS provides in real-world modeling problems such as COVID-19, our study carefully analyzed the associated computational implication. Our analysis highlighted some of the strengths and drawbacks of using ABMS.

3.3.1.4 Benefits in Implementing ABMS

ABMS is excellent in simulating experiments for hypothesis testing, and also functions well to decode the relationship between agents in a system [33]. We can deduce that ABMS can be exploited as a type of contextual magnifying glass to better grasp the reality of our study event, the COVID-19 pandemic, as it offers the natural representation of the system [30], [39]. Additionally, ABMS is adaptable, could simulate

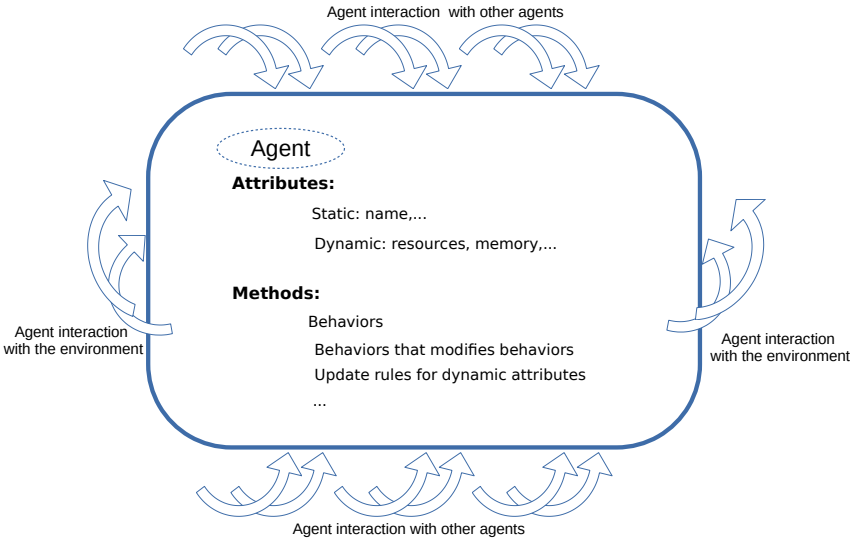


FIGURE 3.1: A Typical Agent Representation in an ABMS System

TABLE 3.1: Some Fields of Application of the ABMS

Application Field	**Description**
Epidemic Modeling	Bio-attacks - Evaluation of a scalable citywide multi-agent model (Carley *et al.* [35])
Energy Analysis	Simulation based analysis of offshore wind energy (Mast *et al.* [36])
Biomedical Study	ABMS Immune-simulator showing interaction between innate and adaptive immunity (Folcik *et al.* [15])
Organizational Decision Making	ABMS approach to allow negotiations for achieving global objectives in location planning of intermodal freight hubs (Van Dam *et al.* [37])
Industrial Chemistry	molecular self-organization modeling via ABMS (Troisi *et al.* [34])
Market Analysis	Using ABMS to predict future tourism market place in sub-orbital space tourism region (Charania *et al.* [38])
Air Traffic Control	An Agent-Based Model focusing on analyzing control policies and performance of an air traffic management resources (Conway *et al.* [14])
Ecology	Simulation of interaction between predator-prey and whales (Testa *et al.* [39])
Anthropology	ABMS describing the structure and dynamics of a prehistoric settlement patterns in lake Titicaca basin of Peru (Griffin *et al.* [33])
Crime Analysis	ABMS simulation approach to predict burglary rate in Leeds and Vancouver (Malleson *et al.* [40])

complex systems [40], and allows for the development of system-specific features which is vital to our study.

Considering these ABMS' strengths to our research has positioned us to provide scientific results to COVID-19 event. ABMS has many advantages, some of which have been highlighted. Though we did not explore the implementation of ABMS, we can ascertain that the beneficial features of an ABMS in the context of our research were duly represented in the ADRIANA system.

3.3.1.5 Critiques and Limitations of ABMS

COVID-19 event as the main focus of this study, we considered the relevance and contributions of an ABMS in modeling phase. Although, we highlighted the benefits of adopting an ABMS as a research instrument in our study, we equally considered the drawbacks of implementing an ABMS, which are noted below.

The modeling and integration of various agents and their interactions in the system incurs high computational cost. These computational costs are due to long simulation time of the system, high computing capacity required to develop and execute the simulation, and complex technique used to evaluate the simulation's accuracy. Despite recent technological advancements, running the ABMS still requires a high-capacity power specification.

Additionally, relating the ABMS simulation to our study includes human actors that could lead to irregular activities, biased decisions, and complex identities. In essence, the conduct of human agents in an ABMS system is difficult to ascertain and evaluate [39]. Another limitation of the ABMS regarding our study is that more practical processes are required to have a full system. ABMS analyses a system at unit level rather than an integrated level. This puts further emphasis on the computational constraints of implementing the ABMS in our study.

Therefore, relating the application of ABMS to our study, we identified the following agents: population, susceptible individuals, exposed individuals, infected individuals, recovered individuals, policymakers, hospitals, hospital equipment, wards, and intensive care units (ICU). Furthermore, factors that contribute to the interaction of the agents include infection rate, recovery rate, reproduction number, algorithmic hyperparameters, algorithmic operations, numbers of confirmed, recovered, death cases, the intensity level of the lockdown, ICU occupancy rate, ward occupancy rate, individual risk predictor, and individual features (such as immune strength).

3.3.2 Discrete Event Simulation

Discrete event simulation (DES) is a tool for simulating the process and performance of a real-world event [41]. It has been applied in different fields including health,

networks, and capital investment [42]. In addition, researchers have used DES to simulate complex real-world problems [42], [43].

The functionality of a DES involves representing a system as a sequence of asynchronous processes [43]. The events can be hospital admission, patient stay in hospital ward or intensive care unit (ICU), patient discharge from hospital, and patient transfer to another hospital [42]. Throughout simulation events, DES assumes a stable system environment [44]. The entities (objects that move within the system), resources (items required to trigger an event), and events (processes of the object) are the typical components of a DES system [43].

The architecture of the discrete event simulation presupposes that the simulated system changes at distinct points in simulation time. The simulation system is asynchronous, implying that it does not operate on a perpetual clock but on simulation time intervals [42].

Our analysis adopted a DES paradigm, with the patients modeled as a separate unit with related features which might change based on the simulation constraints. The entire DES approach involves: the arrival of infected patients at the hospital, admitting a patient to a ward or ICU (depending on disease severity), estimating the time of arrival, estimating treatment time, survival probability (based on factors such as individual risk predictors), estimating the number of treatment resources, and categorizing the patient as either dead or recovered.

The DES has an event queue and allows complex decision logic to be introduced in the simulation process, which is not always possible with other types of simulation approaches [41]. In addition, DES enables us to test different "what-if?" conditions during result analysis.

3.3.3 Trade-Off: Discrete Event Simulation, and Agent-Based Modeling and Simulation in Epidemiology

The ABMS and DES are two commonly used techniques for simulating complex real-world events [34], [35]. Our study shows that DES functionality is constructed in a top-down strategy. The top-down strategy by the DES allows for a less complex model construction and evaluation with a lower computational budget. Contrary, the functionality of ABMS is constructed on a bottom-up design, which enhances the communication between agents and the simulation environment [18].

A comparative study of ABMS and DES reveals that for pandemic simulation, ABMS requires larger computing resources than DES [44]. Therefore, we considered the use of DES and a surrogate as a substitute for ABMS to reduce computational costs while maximizing results.

3.4 Surrogate Optimization Through Genetic Algorithm

3.4.1 Background

Nature has been the source of inspiration for many human innovations. Artificial Neural Networks and Genetic Algorithms are two examples of these innovations [26], [45]. Genetic Algorithm (GA) is a type of evolutionary algorithms (EA) [26]. The GA decodes the actual or suggested solution to a problem and then applies a recombination operator to the data to preserve critical features of the solution.

Genetic Algorithms (GA) are search-based and meta-heuristic algorithms that function on natural evolution and genetic principles [26]. Holland J. at Michigan University, United States of America, was the first to introduce Genetic Algorithm (GA) to the research field in the 1960s and 1970s [41]. The search ability of GAs is based on evolution process by initiating the search with a random set of solutions and producing successive generations as results. Genetic Algorithms (GA) are usually regarded as function optimizers and are applied to a different form of applications.

In GA's, the survival of the fittest or a natural selection mechanism represents the core principle guiding its search-selection operation [41]. Since the operation of GA is similar to the concept of solution evolution by forming generations of solution populations, it requires a fitness function as a decision component [45]. The GA is well-suited for searching solutions with high complexity, which frequently involve discrete, large, and non-linear features. However, while GA does not guarantee an absolute optimal solution, the findings of GA are near the global optimum in the search space [43].

An individual represents a member of the solution to the problem to be evaluated and resolved [16]. Predefined rules governs the optimal selection of the fittest individuals from a population for each generation [26]. The selection process begins with the starting population and continues until the selection function reaches a termination criteria. The pseudo-code illustrating the step-wise implementation of GA is presented in Figure 1.

Basic features included in the structural formation of the GA contribute to classifying GA as heuristic optimization, and generate-and-test optimization algorithm [16], [26]. The key structural features of GA are individual representation, evaluation of fitness function, population, survivor selection technique, gene, chromosome, parents selection mechanism, parents variation operators, mutation, and recombination [16], [43]. In the current research, the state-of-the-art (SOTA) operation of the GA is described as $(1 + (\varphi, \varphi))$. Here, the crossover φ is applied to increase the momentum of searching the hyperparameter space and filtering the run-time analysis to optimize the upper bound of the search result [16], [43].

Algorithm 1 Pseudo Code for Genetic Algorithm

Parameter Definition
$1 < N < \infty,\ n > 0,\ X_{best}$;
SELECT(α), RECOMBINE(β), MUTATE(γ);

BEGIN
INITIALIZE population X_n with random solution
EVALUATE each candidate C_n of the population
REPEAT UNTIL (TERMINATION CONDITION is satisfied)
for $i = 1$ **to** N **do**
(1) SELECT parent as P_{nBest} (*i.e*, $\alpha(C_n)$)
(2) RECOMBINE pairs of PARENT as S_n (*i.e*, $\beta(P_{nBest})$)
(3) MUTATE the resulting offspring as R_0 (*i.e*, $\gamma(S_n)$)
(4) EVALUATE new candidate as $\hat{C}_n$ (*i.e*, $R_0(offspring)$)
(5) SELECT individuals for next generation
if $\hat{C}_n$ is better than X_n **then**
UPDATE X_{best} with $\hat{C}_n$
end if
end for
RETURN the best solution X_{best}
END

3.4.1.1 Notion Guiding Natural Selection by the Genetic Algorithm

The natural selection process of the GA begins by identifying the fittest individuals from a population sample. The selection is based on defined condition(s) within the simulation system. The selected fittest individuals produce offspring with natural characteristics inherent in their parents, which are then transferred to the succeeding generation to produce another population. If parents of an offspring selection are fitter, the generated offspring will have a better chance of being picked as the fittest individual. This process is repeated until the fittest individuals are identified from a generation, or the predefined termination conditions are satisfied. The flowchart in Figure 3.2 illustrates the functional phases of the GA before terminating its operation.

3.4.2 Phases of Genetic Algorithm Operation

The application of GA is an important phase in optimizing the surrogate. The five (5) functional phases involved in implementing the Genetic Algorithm are outlined below. The phases explored by GA are initial population, fitness function estimation, selection, crossover, and mutation. Depending on the type of surrogate's problem, each phase may include a sub-unit.

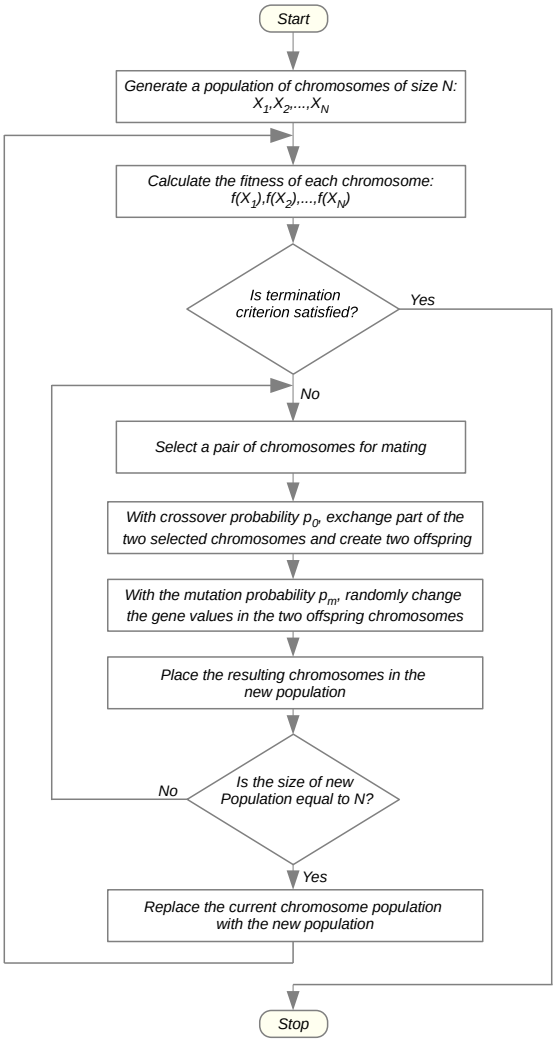


FIGURE 3.2: Flowchart Showing Operation of Genetic Algorithm [17]

3.4.2.1 Initial Population

The GA's process begins with a population, which is a collection of randomly generated individuals. Each individual represents a solution to the optimization problem we want to solve. Genes are a set of characteristics or factors that are unique to each individual. To generate a distinct chromosome, genes are arranged into a single string (or solution). The genes are encoded into chromosomes with binary representation, which subsequently produces the population in which GA acts upon as shown in Figure 3.3.

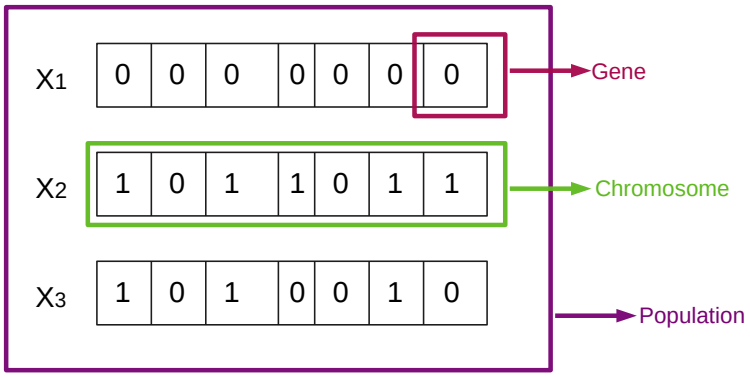


FIGURE 3.3: Population Distribution of Genetic Algorithm

3.4.2.2 Fitness Function

The fitness function determines the inferential quality of a model [26]. It determines whether a model will converge to an optimal result or not. The GA performs best at stronger fitness levels. There are two primary types of fitness functions in GA optimization. The first type is a static fitness function, which optimizes using a defined set of test cases. The second category is a mutable fitness function, which is the coexistence of a sequence of test cases in the search space [46].

The computational field and problem dynamics in which GA will be explored determines the fitness function to be considered. For example, the Pareto fitness function developed to address a multi-objective optimal design concept, as shown in Equation 3.1 uses the maximin function as the fitness function [46].

$$F(x_i) = \max_{j \neq i; j \in Q} \left[\min_{l \leq g \leq h} \left\{ f_g(X_i) - f_g(X_j) \right\} \right] \tag{3.1}$$

where $f(x_i)$ represent fitness of the i^{th} design, Q represent independent point in the current population, and j, l, g, h represents specific data points in the search space.

3.4.2.3 Selection Technique

In GA's optimization process, selection is the method of identifying the fittest genomes (or parents) in a population sample to transmit their genetic features to the next generation through a crossover operator. The fittest genomes are chosen based on their estimated fitness score in relation to a defined reference. The following are the key methods for implementing the selection process by the GA:

(I) After evaluating the fitness function for each individual in the population, a fitness value is assigned to the individual. Then, the fitness value is standardized to ensure that a uniform aggregate value is obtained throughout the population. To normalize a fitness value in the search space, an individual fitness value is divided by the total fitness values.

(II) We compute the accumulated normalized fitness values for each individual in the population by summing the individual fitness value and the fitness values of all previous individuals. A hypothesis to ensure that the normalization process is successfully executed recommends that the cumulative fitness value for the last individual in the population must be equal to one (1), else the normalization procedure is inaccurate and will be repeated [41].

(III) The GA selects a real number p at random between 0 and 1.

(IV) The selected individual must have an accumulated normalized value n greater than or equal to p, i.e. $n \geq p$.

The GA uses different techniques in its selection process, such as fitness proportionate, roulette wheel, Boltzmann selection, stochastic sampling, steady-state selection, elitism selection, and tournament selection [16], [41]. The stochastic sampling technique was considered in this study, and is implemented by random selection of individuals in a generation. Next, the fitness score for the selected individuals is computed, and the individual with highest fitness score is finally selected [26], [47].

3.4.2.4 Crossover Operation

Crossover, also known as recombination in Evolutionary Algorithms (EA), is a genetic operator used to blend the genetic features of two selected parents in a population to generate a new offspring [41], [48]. It is the most important stage of the GA's operation. The new offspring generated through crossover are mutated before integration into the population. The mutation procedure will be covered in later section of this dissertation.

Various algorithms in evolutionary computation captures genetic information using different data structures, allowing genetic representation to be merged with various crossover operators [17]. Data structures that can be merged with crossover operators are often vectors of real values and bit arrays [41]. The GA stores the genetic features in a chromosome encoded by a bit array.

The crossover technique for bit array is a good representation of the genetic recombination process. A random crossover point p_0 from the gene formation is chosen to match each pair of parents in a bit array representation of the chromosomes. Figure 3.4 is a typical demonstration of the crossover procedure. In this example,

we assumed that the crossover point was on gene four (4). Next, new offspring are produced from the crossover operation by exchanging their parents' genes until a crossover termination level is reached.

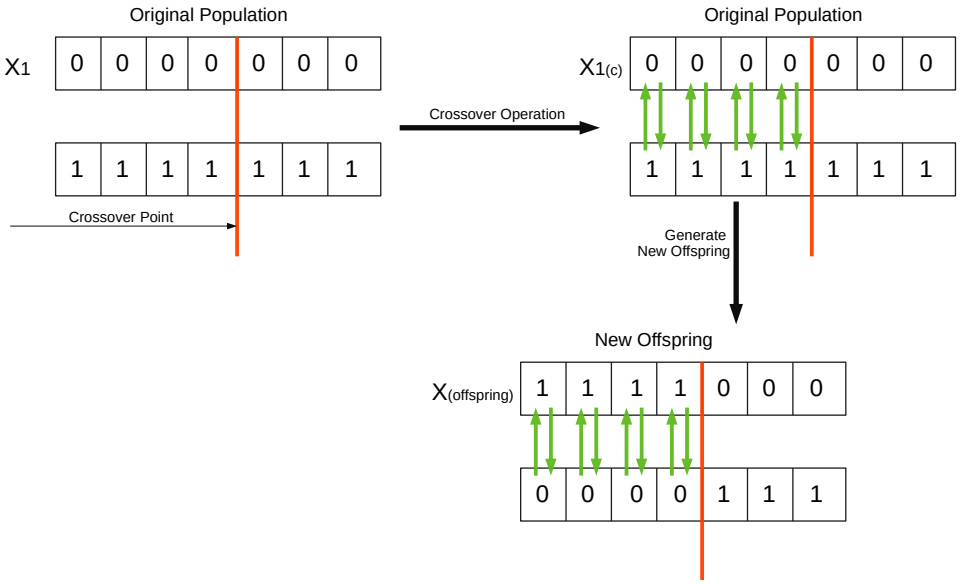


FIGURE 3.4: Representation of Crossover by Genetic Algorithm [17]

3.4.2.5 Mutation

Mutation is one of the genetic operators that functions by adding new characteristic features to the genetic search space in an evolutionary sense [47]. Mutation helps other genetic operators avoid feature loss[41]. For example, it prevents chromosomes from exhibiting many similarities in their features. The mutation operator exhibits different characteristics in a population sample during the GA's operation, especially when population features are static because of crossover reiteration and reproduction operators [16], [41].

One of the major impacts of mutation is that it handles variations in individual chromosomes when compared to their parents. Mutation operates on a bit-array level by copying values from one string to another with transferable mutated bit features [41]. However, the mutation probability (p_m) is typically a small value.

Empirically, the type of mutation operator used is determined by the environment in which the Genetic Algorithm functions. For example, every population contains several mutation operators for different genomes. The following mutation operators are applicable to different fields of Genetic Algorithm's operation: swap, interchanging, bit flip, creep, reversing, scramble, random resetting, uniform, inversion, and insert mutation [17], [41]. The typical mutation processes of the Genetic Algorithm is depicted in Figure 3.5.

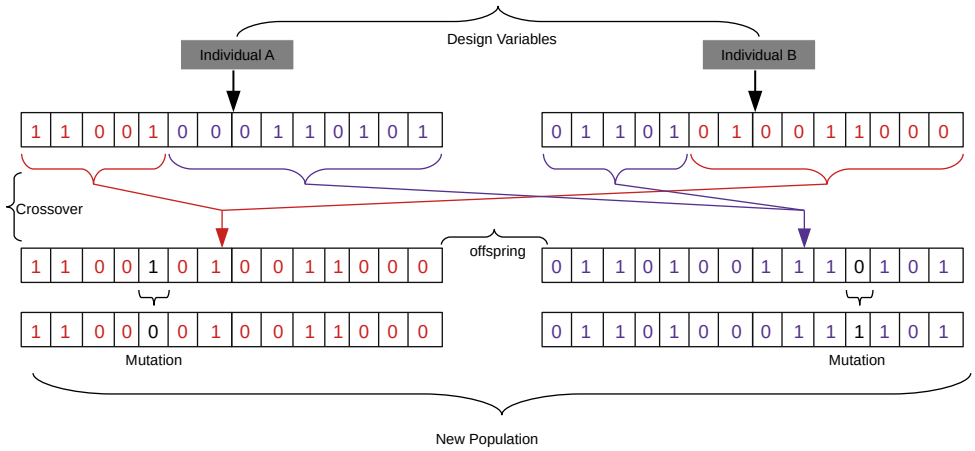


FIGURE 3.5: Typical Mutation Process of the Genetic Algorithm [17]

3.4.2.6 Termination Operation

The Genetic Algorithm terminates when the population converges (that is, it does not produce different offspring from the preceding generation) [16], [26]. The final implementation phase of a GA is the termination phase, which is determined by different factors.

In general, GA stops its operation when newly produced populations converge to a point in the search space [26]. The convergence indicates that the characteristics of GA's new offspring are not significantly different from those of previous generations. Typically, the termination operation of GA occurs when, but is not limited to:

- A predefined termination criteria has been satisfied.
- An individual solution in a population reaches a certain degree of fitness.
- A defined number of generations have been completed.
- There is consistency in the features of new offspring and subsequent populations, and will not provide better result, that is, population variation features fade to a diminishing point.

3.4.3 Convergence Metric of Genetic Algorithm

The Genetic Algorithm (GA) is a stochastic search algorithm. More particularly, depending on the complexity of the solution space, there is no certainty that the GA will uncover the absolute solution to the problem description [26]. Thus, it is realistic to assume that the GA converges to a mean value after exploring the search

space [16]. The mutation operator is one of the core components that contribute to the convergence dynamics of the GA [17], [47]. In essence, the convergence of the GA to a global optimum can be expected if it has been running for a long time and acted upon by a mutation operator.

Computationally, there is an average upper limit on the maximum iterations required for the GA to reach convergence [16]. In general, the upper bound value decreases as the population size in the solution space increases. Researchers demonstrated this by proposing and exploring convergence assumptions that apply to their optimization problem [41], [47].

3.4.4 Operational Constraints of Genetic Algorithm

The Genetic Algorithm considered as a universal optimizer for solving optimization problems has certain constraints [16], [41]. Some of which are listed below:

- Efficient modeling of the objective fitness function is quite challenging. As a result, using GA to represent complex problems may demand prolonged time and high computational requirements, both of which may be unsuitable for analysis.
- When modeling and optimizing complex problems with the GA, identifying a point when a global optimum has been attained may be challenging. One major reason for the challenge is that, while successive generations may attain high fitness values, they can be misinterpreted as a global optimum, resulting in a skewed optimization result [41].
- Developing GAs to solve an optimization problems with binary values only (e.g., success or failure, as in the decision tree) may be challenging, because there is no convergence condition for such optimization design [17].
- In several optimization problems, the GA exhibits biased solutions towards converging to local optima rather than the global optimum of the problem specification. Therefore, GA cannot substitute short-term fitness for long-term fitness solutions.

3.4.5 Applications of Genetic Algorithms

The Genetic Algorithm has been widely applied to different areas of research and in different domains of knowledge [26], [41]. It has become one of the sought-after tools used by researchers for optimizing models [26]. Some of its application areas involves machine learning, multi-modal optimization, engineering, finance, economics, process scheduling, wireless sensor network, and natural science [16], [41], [47]. The key vocabularies used in Genetic Algorithm is presented in Table 3.2

TABLE 3.2: Vocabularies Used in Genetic Algorithms

Vocabulary	**Description**
Genes (bits)	Unit part of a solution
Alleles	Values of genes
Locus	Location of a gene
Genotype	Encrypted solution
Phenotype	Decrypted solution
Chromosome(individual, string)	Solution (coding)
Population	Collection of solutions

An overview of the relationship between the operation of Genetic Algorithm and simulation experiment is presented in Figure 3.6.

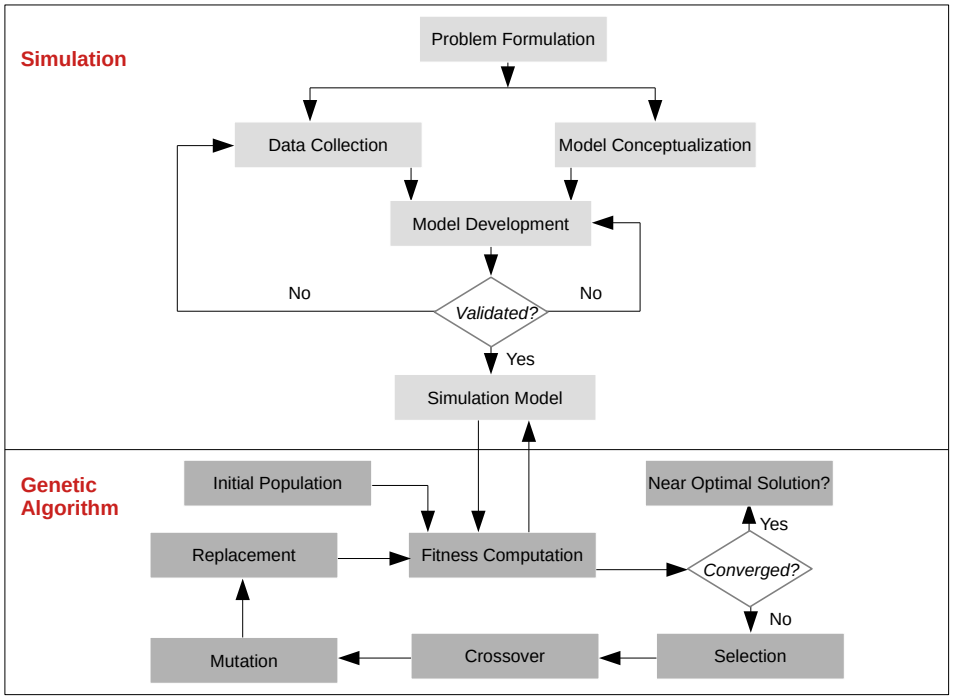


FIGURE 3.6: Overview of the Relationship Between GA and Simulation Development

3.4.6 Underpinnings of Genetic Algorithm

Evolutionary Computation (EC) is an advanced modeling discipline with several techniques and processes, where a well formed modeling framework is subjected to complexity [16], [41]. Therefore, different high-level EC libraries have been developed to provide answers to complex computational problems.

This study explored the deep learning framework, heuristic machine learning approaches, and statistical models. One of the key goals of this study was to confirm that the surrogate is calibrated to give a good solution. The optimal hyperparameter calibration of the surrogate model influences surrogate quality. We adopted the Genetic Algorithm (GA) as a framework for optimizing the deep learning architecture. We used the GridSearch technique with a 10-fold cross-validation parameter search to enhance the performance of the heuristic model. To arrive at the optimal hyperparameter values for other forecast models explored in this study, we analyzed the internal features of the models.

Our study explored Distributed Evolutionary Algorithm in Python (DEAP) framework, which is one of the derivatives of the Genetic Algorithms and relates to Evolutionary Computation (EC).

The Distributed Evolutionary Algorithms in Python (DEAP) framework is a tool created and built on the Python programming language which researchers embrace to create advanced Evolutionary Computation systems [49]. DEAP[1] framework focuses on providing functional tools for accelerated evolutionary algorithms, with a clear implementation process (i.e. in pseudocode).

The DEAP architecture is also notable for its emphasis on code simplicity and compactness, which attracted our attention to its implementation as the evolutionary optimization engine. One of the primary strengths of DEAP architecture is that it provides transparent parallel processing in optimization process [49], [50]. Researchers use different novel techniques in DEAP to solve real-world problems. They considered DEAP to be a useful tool in the optimization process [50]. The general architecture and core modules of DEAP algorithm is illustrated in Figure 3.7.

3.4.7 Implementation Analysis of DEAP Architecture in Our Study

The core internal architecture of the DEAP framework in our study comprises three major parts: the base, creator, and toolbox module.

The ***base*** module contains data structures and objects extensively used in Genetic Algorithm computation; however, it has not been implemented in the Python official library [50]. The ***creator*** module enables the formulation of classes in real-time (or run-time) using composition or inheritance features. As a result, user-specific GA functionality may be activated by dynamically adding functions and data as arguments to pre-existing classes. In practice, the creator module assists in generating populations and genotypes from different data structures encountered throughout the optimization experiment [50].

[1]Freely available at: https://deap.readthedocs.io/ and http://github.com/DEAP/deap

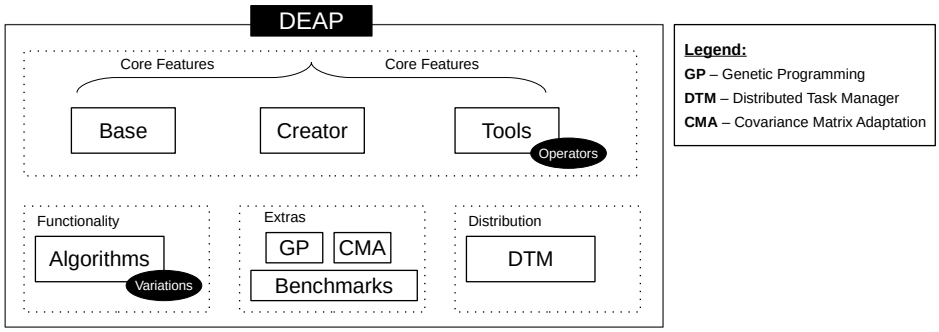


FIGURE 3.7: Core Architectural Modules of DEAP [49]

The ***toolbox*** module is a container for the operators (or tools) used to simulate the genetic operation of the DEAP framework [49]. During implementation, we populated the toolbox with useful computational DEAP tools. Furthermore, the DEAP framework uses Distributed Task Manager (DTM) in the implementation process to create and allocate sub-tasks among different internal modules [50].

During DEAP implementation, we represented individuals in a population as a Bernoulli randomly generated 10-bits string, with the first 6-bits representing the window size optimization feature of the neural network and the last 4-bits representing neuron optimization of the neural network. The goal was to maximize the prediction accuracy of the network model while minimizing the network's computed Root Mean Squared Error (RMSE). The toolbox contains parameters such as individual binary bits, mutate operator, population, select operator, evaluate operator (fitness function), and crossover operator.

The implementation procedures of DEAP architecture in this study involve random generation of individuals through a Bernoulli binary function generator. The binary value generated has gene length of 10, which produces the individuals that forms the population. Next, a crossover probability (*cxpb*) was applied to the population in the present generation. A random mutation probability (*mutpb*) was applied to shuffle the index of the individual population, and finally, the fittest parameter is chosen based on the pre-defined selection criteria. The selection criteria is a minimum RMSE with an optimal accuracy evaluated during the GA's iteration process. These feature combinations are allowed to function on training data to compute the optimal window size and neuron size to feed the surrogate neural network.

As shown in Table 4.9 and Appendix B, the DEAP implementation in our study

has number of generations of 45, population size of 10, gene length of 10, mutation probability of 0.1, and crossover probability of 0.3. In addition, Figure 3.8 represents the transitional process of individuals between different generations in GA as explored by DEAP architecture.

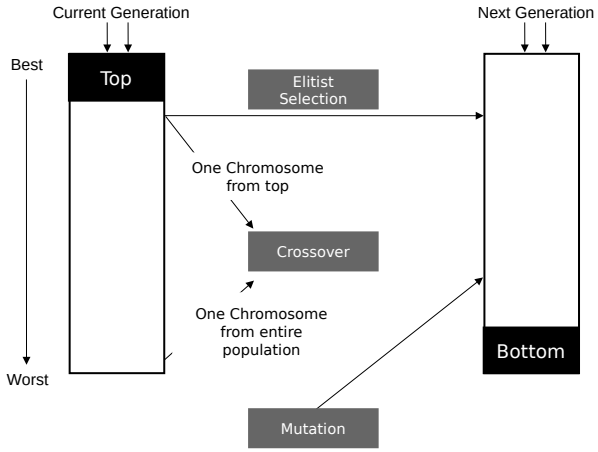


FIGURE 3.8: Genetic Algorithm's Transitional Process Between Generations [49]

3.5 Epidemiology Compartmental Model

Applying compartmental model to simulate a pandemic event is a recent development by researchers in the last decade [24], [51]. The compartmental simulation is used to examine the transmission dynamics of infectious disease in a population sample N. The complexity level of the disease determines the compartmental model settings required to extract insight of the disease spread dynamics and to evaluate required control strategies.

The SEIR model is an early development of the epidemiology mathematical compartmental model, which was developed by public health physicians (R.A. Ross, A.G McKendrick, W.H Hamer, and W.O, Kermack) and supported by the statistical contribution of J.Brownlee between 1900 and 1935 [52]. Compartmental models have been extensively applied as an efficient tool to simulate disease infection rate, and associated dynamics [51], [53].

Our study explores the epidemiological compartmental model by grouping homogenous populations into four different compartments, namely, susceptible (S), Exposed (E), Infectious (I), and Recovered (R). Individuals in a population with a function of time t can be categorized as susceptible $S(t)$, exposed but not infectious $E(t)$, eventually infectious $I(t)$, or recovered (otherwise immune) $R(t)$. The SEIR

model has been applied to simulate population groupings of different infectious diseases such as Ebola, Dengue, Flu, and Measles [54], [55]. Categorizing individuals into various compartments considers the relationship between the compartments and the transition of individuals to next compartment.

Our research covered the compartmental simulation of the COVID-19 pandemic. During SEIR simulation process, we ensure that there is no flow of new susceptible individuals. Therefore, the number of infectious population diminishes to zero over time. During a pandemic, one of the primary difficulties of interest to policymakers and public health practitioners is identifying the severity level of the disease in the population, which guides the formulation of disease prevention policy.

These concerns can be further subdivided to ask how long the pandemic would last. How many individuals in the population will be affected by the disease and need treatment? What is the average maximum number of individuals requiring urgent treatment at a given time t? We provided applicable solutions to resolve these concerns by developing a compartmental model to group population based on transmission dynamics.

3.5.1 SEIR Compartmental Model Simulation Process

We assumed that the dynamics of COVID-19 disease transmission are deterministic. Therefore, the population's behavior depends entirely on the historical response to the pandemic and the rules governing the simulation function. We also assume that the number of individuals in a compartment is a differential function of time t while defining the internal components of each compartments of the model. In addition, at the inception of a pandemic, there are few infectious individuals in the population, and the transmission rate is mostly regulated by random interactions between the non-infectious and infectious populations. We considered these interaction dynamics during SEIR model development.

The time t is an independent variable in the SEIR model development. We mathematically expressed the rate of population movement within compartments as a time derivative. As a result, estimating the duration of stay of individuals in each compartment is based on using ordinary differential equations [56]. To develop COVID-19 compartmental transmission simulation, we ensured that our assumptions were valid (through research) regarding the determining factor that drives the rate of flow of individuals between the compartments. Simply stated, we built a mathematical expression in the simulation that relates the infection rate with the recovery rate in the compartmental simulation as expressed in next paragraph.

Next, we implemented the assumptions concerning infection dynamics, referred to as the mass action frequency [57]. According to the mass action frequency hypothesis, the rate at which a disease expands in a population is jointly proportional

to the product of susceptible persons S and infected individuals I. Therefore, taking r as a numerical constant and the degree of new infection as C, we can express the law of mass action frequency as:

$$C = rIS \tag{3.2}$$

To demonstrate the law of mass action frequency, suppose a population of size N, with an individual average contact rate of δN which is sufficiently assumed to transmit infection in time t. Then, computing the probability of a spontaneous contact of an infectious individual with a susceptible individual is given as S/N. Furthermore, computing the rate of new infections in unit time per infectious individual is estimated as $(\delta N)(S/N)$. Therefore, we are able to generate new infection rate in the population as $\delta SI = (\delta N)(S/N)I$.

Conversely, the probability of a susceptible individual interacting with an infected individual is I/N. As a result, the degree of new infections in the susceptible population is given as $(\delta N)(I/N)$, and the rate of new infection in the population is $\delta SI = (\delta N)(I/N)S$. It is worth noting that both mathematical methods described above are for estimating the magnitude of new infections in a population during a pandemic. Essentially, different compartment structures have unique complexities. These complexities determine the different accuracy levels of estimating new infection rate. Figure 3.9 illustrates the operational transition diagram of the SEIR architectural structure.

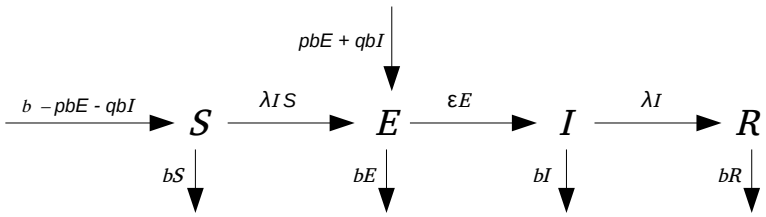


FIGURE 3.9: SEIR Compartmental Population Transfer Diagram [57]

In Figure 3.9, the total population is grouped into the susceptible, exposed, infectious and recovered partitions, represented by $S(t)$, $E(t)$, $I(t)$, and $R(t)$ respectively. The birth rate and mortality rates are assumed to be equal and are symbolized by b. Furthermore, the disease is not expected to lead to the death of the starting individual in order to sustain the overall population density, that is, $S(t) + E(t) +$

$I(t) + R(t) = 1$. Furthermore, the expression expresses the bi-linear mass action of the compartment's operation as shown in Equation 3.2. The variables p and q respectively measure the proportion of new offspring from the infected and exposed compartment.

The birth flow of individuals into the susceptible compartment is denoted as $b - pbE - qbI$, while the birth flow of individuals into the exposed compartments is represented as $pbE + qbI$. Generally, for proper compartmental modeling operation, the conditions $0 \leq p \leq 1$ and $0 \leq q \leq 1$ must be satisfied. The compartmental transfer diagram illustrated in Figure 3.9 is the Euler differential equations for simulating the SEIR model represented by Equation 3.3.

$$\begin{cases} S^{'}(t) = b - rIS - pbE - qbI - bS & \text{Susceptible compartment} \\ E^{'}(t) = rIS + pbE + qbI - (\epsilon + b)E & \text{Exposed compartment} \\ I^{'}(t) = E - (\gamma + b)I & \text{Infectious compartment} \\ R^{'}(t) = \gamma I - bR & \text{Recovered compartment} \end{cases} \tag{3.3}$$

Further, we investigate Equation 3.3 in the following numerical field:

$$\Sigma = \{(S, E, I, R) \in \mathbb{R}^4_+ : S + E + I + R = 1\}$$

We observed that Σ is positively even in Equation 3.3. The variable $\epsilon > 0$ represents the rate at which exposed populations becomes infectious. Likewise, $\gamma \geq 0$ represents the recovery rate of infectious individuals in the population. Therefore, we express $1/\epsilon$ to represent the mean incubation period or the latent period, and $1/\gamma$ to represent the mean infections. If $\gamma = 0$, it indicates that there is no recovery from the disease and Equation 3.3 reduces to a Susceptible, Exposed, Infectious (SEI) model.

Likewise, if ϵ tends towards infinity, (i.e. $\epsilon \longrightarrow \infty$), then latent period is insignificant, which reduces Equation 3.3 to a Susceptible, Infectious, Recovered (SIR) model. Also, if $p = q = b = 0$, then Equation 3.3 translates to a classical SEIR model. These claims have been proven in a studies of by Greenhalgh *et al.* [55] and Hethcote *et al.* [56].

The SEIR simulation derived a cardinal reproduction number $R_0(p, q)$ which determines the overall dynamics of Equation 3.3. Reproduction number[2] is the number of secondary occurrences produced by an infectious individual in a susceptible population during an pandemic. R_0 is dimensionless with a unit of $time^{-1}$[51], which is mathematically expressed as:

[2]Evaluating the rate at which an individual at the infectious stage of disease produces another infectious individual in a closed population.

$$R_0 \propto \left(\frac{contact}{time}\right).\left(\frac{time}{infection}\right).\left(\frac{infection}{contact}\right)$$

More clearly:

$$R_0 = \rho.d.\tau$$

where ρ represents the average contact rate between susceptible and infected individuals in a closed population, τ shows the disease transmission probability[3], and d is the duration which an infectious individual manages an infection.

3.6 Design Study for Machine Learning Forecasting

Machine learning is the engineering approach where computer systems observe patterns and learn from past occurrences [58]. A system's learning function is regarded as its ability to improve on a task T based on experience E, with respect to a performance metric P, to generate intelligent predictions about the observed pattern [58], [59]. A task can be a COVID-19 forecast, a crime rate forecast, or anything else measured using prediction accuracy like the root mean square error and the experience represented by the historical dataset.

More specifically, the machine learning field is concerned with solving two related questions, which are. (1) How can computer systems be engineered to enhance their operation through experience? and (2) What are the fundamental statistical and computational information rules governing the learning system? Analytical study of the machine learning approach is required to answer the posed questions.

The sub-sections below describes the machine learning frameworks that were explored in this study. They consist of Neural networks (subsection 3.6.1), regression (subsection 3.6.4), decision tree (subsection 3.6.4.2), and kernel (subsection 3.6.3) machine learning methods.

3.6.1 Neural Networks

Neural networks (NN) are mathematical and computational models that process information using a learning paradigm similar to brain structural formation [59]. Neural networks (NN) are showing great strides in time series forecasting [60]. In addition, NN does not require accurate data-assumption conditions before being applied, and also can efficiently simulate imperfect or incomplete datasets [60].

[3]The probability of a susceptible individual getting infected when contact is made with an infectious individual.

3.6.1.1 Recurrent Neural Network

Recurrent neural networks (RNNs) are very effective and persistent information models for addressing complex tasks [59]. They are a type of neural networks that adequately represent data sequences using an internal memory structure [59]. The fundamental architecture is similar to feed-forward neural networks (FFNNs), with feedback loops connecting current and previous results [45], [59]. RNNs structural formulation includes an input gate, a hidden layer, and an output gate.

Kumar *et al.* [22] and Ella *et al.* [45] demonstrated that RNN could accurately forecast COVID-19 transmission trends by using the long short-term memory (LSTM) network, a variant of RNN. Their analysis also indicates that the neural model outperforms the statistical ARMA and ARIMA models.

Figure 3.10 illustrates the structural component of the RNN, showing the hidden layer transferring information from one input unit to the other in the network sequentially. W_x represents the weight for the hidden layer, W_y is the weight vector for the output layer, and W_h represents weight vectors at different time intervals. The input and output vector at time t are denoted by x_n and y_n respectively, and the activation function represented by h_n.

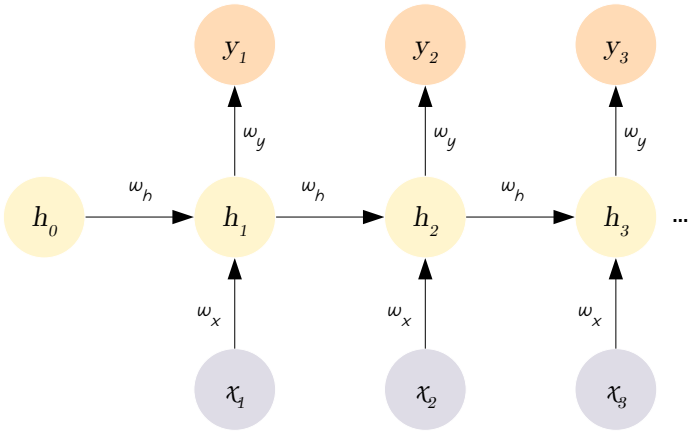


FIGURE 3.10: Schematic Diagram of the Recurrent Neural Network Architecture [61].

The operation of the hidden layer at timestamp t for RNNs is computed with $h(t) = f(h_{t-1}, x_t)$, where the present state, previous state, and input state are represented by h_t, h_{t-1}, and x_t respectively. Likewise, the operation of the output layer at timestamp t is computed as $y(t) = W_{hy}.h_t$, where y_t is the output and W_{hy} is the weight at output layer.

Our study explores multilayer Perceptron (MLP), and two RNNs, the Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). They are described below.

3.6.1.2 Long Short-Term Memory Neural Network

Long Short-Term Memory (LSTM) is a variant of RNN that is used to adequately represent the temporal sequence of data to determine long-term dependencies [22]. LSTM has been applied to various labeling and sequence prediction tasks. The three major gates in the architecture of LSTM are input, forget and output gate. LSTM operates by mapping an input sequence $x_i = (x_1, x_2, ..., x_n)$ to an output sequence $y_i = (y_1, y_2, ..., y_n)$ by evaluating the network activation function.

A study conducted by Kumar *et al.* [22] revealed that LSTM networks could accurately forecast the COVID-19 transmission rate. Their study demonstrated that LSTM provides a notable improvement in forecast outcomes in comparison to traditional forecasting approaches. In our study, we stacked four (4) LSTM layers in our LSTM implementation. Furthermore, to avoid model overfitting, we introduced a dropout layer between the stacked layers. Finally, we added early-stopping constraint to the LSTM architecture to terminate the training process whenever the validation metrics converges.

3.6.1.3 Multilayer Perceptron Neural Network

The Multilayer Perceptron (MLP) neural network, commonly known as the feed forward neural network (FFNN), is the core of contemporary neural network architecture. They are neural network types in which information flow is transmitted in a forward direction from the input node to the output node [61]. The main goal of MLP is to estimate a predictor function f, that relates an input vector $\tilde{x}$ to a output variable y, (i.e. $y = f(\tilde{x})$) [59]. MLP allows nonlinear correlation between data features to be accurately modeled, which is dominant in the COVID-19 datasets.

Rikz *et al.* [45] conducted a study that established the strength of MLP to forecast COVID-19 transmission patterns. They integrated a multilayer feed forward neural network with an enhanced interior search algorithm based on probabilistic learning to improve the model performance [45].

3.6.1.4 Gated Recurrent Unit

For sequence learning tasks, the Gated Recurrent Unit (GRU) uses a network architecture similar to LSTM. Unlike LSTM, GRU has only two gates, which are the reset gate and the update gate. GRU can solve the concerns of gradient expansion and vanishing inherent in traditional RNNs while still learning long-term dependencies

[62]. Additionally, compared to LSTM, the operation of GRU is less complex and faster for training and task execution [61].

3.6.2 Neural Network Activation Functions

Given a set of inputs to the neural network, then the activation function, also termed the transfer function, is used to regulate a network node's output. The activation functions applied for model development are determined by the type of problems to be analyzed. Some of neural network's activation functions include hyperbolic tangent (tanh), exponential linear unit (ELU), leaky rectified linear unit (Leaky ReLU), softplus, and sigmoid shrinkage [62]. Our study considered the ReLU activation function for the neural network models.

3.6.3 Kernel Methods - Support Vector Machine (SVM)

Support Vector Machines (SVMs) are supervised machine learning algorithms that have demonstrated success in classification (using pattern recognition) and regression-related tasks [63]. In the past, using SVMs for computational tasks was supported by factors such as (1) Whilst other machine learning techniques, such as ANN, converge to local minima, SVM solutions converge to a global optimum. (2) Rather than minimizing the training error, SVMs will minimize the upper limit of its model's general performance, making it more efficient with good generalization results. (3) SVM has Convex optimization feature to achieve optimal results [64], and (4) Outliers in datasets have less influence on SVM output. Additionally, SVM is efficient at dealing high dimensional data [63], [64].

SVMs operates by dividing data into two distinct classes using a hyperplane as demonstrated in Figure 3.11. The data points in the training set can be classified as belonging to either of the two split classes. Essentially, SVMs split data points into classes to construct an optimum hyperplane that maximizes the distance between the data points and the hyperplane.

3.6.4 Regression Model for COVID-19 Forecasting

A regression model is a type of modeling technique that analyses the relationship that exists between a predictor variable and a target variable [65]. It has been applied in different computational disciplines, such as time series and forecasting studies [65], [66]. Typically, a regression model has an intercept a, slope of the regression line $\vec{b}$, independent variable x, and an error term ϵ. Mathematically, we present the multi-linear regression model (explored in this study) by:

$$y(x_i) = a + b_1 x_{i-1} + b_2 x_{i-2} + ... + b_n x_{i-n} + \epsilon \tag{3.4}$$

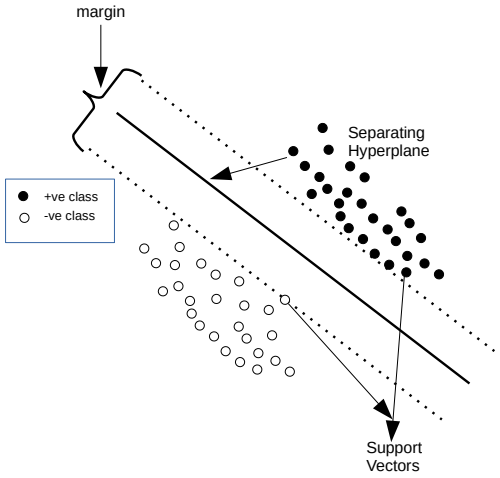


FIGURE 3.11: Schematic diagram of Support Vector Machine [63]

A study by Zhenyu *et al.* [65] has demonstrated the use of regression models in COVID-19 forecasting. They considered linear regression, Lasso, multi-polynomial, and multi-linear regression model for their prediction to guide in hospital capacity planning and policy implementation to mitigate the disease transmission in the United States of America. Our study also implemented the linear regression model with implementation details discussed in Chapter 4.

3.6.4.1 Support Vector Regression

We applied support vector regression (SVR), a sub-model of SVM for prediction task. Further, as demonstrated in Equation 3.5, the implementation of an estimator function ($\epsilon-$ sensitivity) is essential, and is considered for modeling SVM's regression task.

$$L_\epsilon(x_i) = \begin{cases} 0, & if\ |y_i - f(x_i)| \leq \epsilon,\ \epsilon \geq 0 \\ |y_i - f(x_i)| - \epsilon, & else \end{cases} \tag{3.5}$$

where $f(x_i)$ is the predicted values, y_i is the observed values, and ϵ is the error estimate.

The error estimator function for SVR as proposed by Nesreen *et al.* [67] is illustrated by $f(x) = \sum_{n=1}^{M}(\alpha_m^* - \alpha_m)x^T x + b$, where the Lagrange multipliers are represented by α_m and α_m^*, the bias term is denoted by b, and the input values are represented by x.

Lagrange multipliers are significant for error estimation because SVR is applied

to solve constraint optimization problem [63]. Lagrange multipliers works by introducing a random variable into the model architecture [63], [67]. In essence, minimizing the loss function $J = C\sum_{n=1}^{M} |L_\epsilon(x_i)| + \frac{1}{2}||w||^2$ resulted in optimization of the error estimation function [67].

The work of Suvarna *et al.[66]* demonstrated how SVR is effectively used to predict the transmission dynamics of COVID-19. They compared the result obtained from SVR with the neural networks, and they achieved a similar result.

3.6.4.2 Random Forest

Random forest (RF) is one of the modern models that have been used for classification and regression analysis [68]. The random forest regressor (RFR) is a variant of the RF. It addresses regression problems by generating multiple decision tree branches during model training [64], [68].

Implementing RFR in our research experiment as one of the surrogates also resulted in high performance of ADRIANA system, especially when optimized. Due to the similar model design of the RFR and DTR, we also considered the decision tree regressor (DTR) in our study. We configured the parameters required to fit the datasets for both models. The tree (in RFR and DTR) represents their structural environment, mathematically illustrated in Equation 3.6.

$$y(x_i) = \frac{1}{H}\sum_{n=1}^{H} T_n x_i \tag{3.6}$$

where H specifies the number of trees in the forest, T_n represents individual trees in the forest.

3.7 Time Series Models for COVID-19 Forecasting

Different studies on COVID-19 research have also examined the use of statistical techniques to forecast the transmission patterns of the disease, as highlighted in Chapter 2. However, the emergence of the pandemic disease shows a specific pattern in datasets that models should decode to give an effective forecast. For the COVID-19 prediction, time series models such as the Auto-Regressive Integrated Moving Average (ARIMA), Auto-Regressive Moving Average (ARMA), Auto-Regressive (AR), and Moving Average (MA) have been explored by researchers. However, these time series models were shown to be based on presumptions, causing their inability to accurately forecast transmission pattern of COVID-19 [13], [24].

We considered ARIMA and AR time series models to forecast the COVID-19 transmission pattern. We observed the degree of performance of time series model

in our research. Despite formulating these time series models with the best parameter calibration, they did not give optimal prediction performance and could not correctly fit the data compared to other machine learning models that we explored.

Kumar *et al.*[22] conducted an investigation that predicted the transmission wave of COVID-19 in Canada using LSTM neural networks and ARMA time series model. They concluded that time series models did not meet their expectations as statistical results were analyzed against real-world data. The limitation of time series models were also evidenced in the works of Cleo *et al.*[18] and Marta *et al.*[24].

The parameter values are critical part of executing the ARIMA model. The (AR)-(I)-(MA) is made up of three parameters: p for the AR, d for the I, and q for the MA, interpreted as follows: p is the number of previous time steps to be considered for current prediction, d the number of times the series should be differenced to achieve a stationary series, and q is the autoregression error estimation variable. Appropriate statistical tools have been considered in our study to estimate the optimal p, d, q model values, which are presented in Table 4.9 of Chapter 4.

3.8 Model Performance Evaluation Indices

A suitable model assessment approach is required to build an accurate machine learning model. Performance evaluation is a major part of machine learning development process. The surrogate model's development life cycle includes unifying the model's components into a single unit, testing the model with new observations, and assessing the model using set of evaluation metrics. The feedback from the analysis will influence the model improvement until the model reaches the desired accuracy.

In addition, machine learning performance measures how efficient the model can correlate an input to an output. Also, it measures model's generalization to new observations that were absent during the model training phase [69]. Datasets were splitted into training and testing sets to assess the accuracy of a model. Our study used 70% of the dataset for training and 30% of the dataset for testing the surrogates, which is consistent with the SIIMCF's CRD cases.

To assess model's quality, both the training and test datasets must be evaluated to extract the error dynamics. We considered Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) as the surrogate evaluation metrics.

Mean Absolute Error (MAE) is an estimate of the average magnitude of the model error without considering its direction in the hyperplane [70]. MAE is computed by summing the absolute difference between the observed and predicted values, and is mathematically represented by:

$$\left(\frac{1}{n}\right)\sum_{i=1}^{n}|y_i - \hat{n}|, \tag{3.7}$$

where n is the number of observations, y_i is the observed values, and $\hat{n}$ is the predicted values.

Root Mean Square Error (RMSE) measures the standard deviation error of model's prediction by estimating the distance between a regression line and data points [69]. Additionally, we used RMSE for error estimation to minimize loss function in the surrogates [62]. RMSE is mathematically represented by $\sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - \hat{n})^2}$, where y_i represents the set of observed values, $\hat{n}$ represents the set of predicted values, and n is the number of observations.

R-Squared (R^2) measures the co-linearity between the predicted and observed values [71]. It estimates the model accuracy by examining the estimated values between a range of 0 and 1. An R^2 value closer to 1 indicate a better model fit with less error variance between the two sets of observations (the observed and predicted values) [71]. Generally, R^2 is proportional to the model accuracy. R^2 is mathematically represented as $\frac{1-SS_{res}}{1-SS_{tot}}$, where SS_{res} is the sum of squares of the residual errors of the model, and SS_{tot} is the total sum of errors.

3.8.1 Model Evaluation Using Confidence Interval

Confidence intervals (CIs) are techniques of quantitative validation in modeling and simulation experiment [72]. They are range of values that include a numerical concentration of interest [72]. The confidence level is a percentage-based evaluation of an unknown value based on statistical similarity to standard confidence interval values.

3.8.1.1 Confidence Interval Computation

To compute confidence interval, if we denote the lower limit and the upper limit of a population distribution as $\tilde{L_b}$ and $\tilde{U_b}$ respectively, population mean to be ℓ, error margin to be τ, and critical value to be ξ_l. Then, we can express the CI to be $[(\ell - \tau), (\ell + \tau)]$. The error margin τ depends on data feature from population distribution.

Assuming a dataset has a normal distribution and a standard deviation of σ, then the confidence interval is computed mathematically as illustrated in Equation 3.8:

$$\left[(\ell - \xi_l)\frac{\sigma}{\sqrt{\eta}}, \quad (\ell + \xi_l)\frac{\sigma}{\sqrt{\eta}}\right] \tag{3.8}$$

Where ξ_l represent the critical value (obtained from statistical tables) for the population distribution with confidence level l, and η is the frequency of the population distribution.

3.8.1.2 Confidence Interval Validation

In this study, we ensured high estimation accuracy of model performance. We explored eight (8) steps to compute the confidence intervals of our model results. The steps are outlined below:

(1) We selected the output of a model variable χ_i for validation.

(2) Based on model execution, we computed values of model execution samples η.

(3) We ran the model η-times, observed the output variable χ_i for every $i^{th}-$iteration to generate the samples $\chi_1, \chi_2, \chi_3, ..., \chi_n$.

(4) We computed the standard deviation σ, and the mean ℓ of the model from the distribution $\chi_1, \chi_2, \chi_3, ..., \chi_n$ obtained from (3).

(5) We obtain the critical value (from statistical table) ξ_l by examining the distribution of the model output in (3), the distribution size η, and the standard deviation σ.

(6) We obtained the confidence level c.

(7) Based on the parameters selected, the confidence interval CI was estimated by applying Equation 3.8.

(8) We verified if the model output variable Y was within the computed confidence interval. If $\tilde{L_b} \leq \mathrm{Y} \leq \tilde{U_b}$, then the model was valid, else, the model was invalid for population distribution $\chi_1, \chi_2, \chi_3, ..., \chi_n$.

Illustration: Given error margin (∇), and a variable χ_i (i.e, point estimate), we computed lower bound L and upper bound U of the series $[\chi - \nabla, \chi + \nabla]$ as shown in Table 3.3. From Table 4.10, we extracted the LSTM's forecast (as the series χ_i) for confirmed cases in South Africa for day $70^{\pm 5}$ with values from day-65 to day-75. We used the following parameter configuration, sample size (n) is 11 with values shown in Table 3.3, the mean (ℓ) of the distribution (χ_i) is estimated as 1363.5, the standard deviation (σ) is estimated as 421.8, error margin (∇) computed as 249, the degree of freedom is computed as $n - 1$ is 10, and the critical value ξ_l (used for the t-distribution to estimate the confidence level) of 0.95.

We estimated the confidence interval as values between 1364 ± 249, resulting in data points ranging from [1120 to 1610], with 1120 and 1610 being the lower bound and upper bound respectively.

TABLE 3.3: Estimation of Confidence Interval

Day	**Lower Bound** (L)	**Actual Value** (χ_i)	**Upper Bound** (U)
65	855	1104	1353
66	685	934	1183
67	714	963	1212
68	1053	1302	1551
69	1166	1415	1664
70	1279	1528	1777
71	1308	1557	1806
72	826	1075	1324
73	1308	1557	1806
74	826	1075	1324
75	2160	2490	2658

3.8.2 Mean Absolute Percentage Error

The Mean Absolute Percent Error (MAPE) is a good evaluation metric to estimate the prediction accuracy of two data series [73]. MAPE is the average or mean of predicted absolute error measured in percentage. MAPE is estimated by summing the absolute percentage error of the two series. In essence, smaller MAPE indicates better prediction accuracy [73].

The MAPE formula is divided into two parts, which are the M part and the APE part. The M is the mean or average of the estimated APE values. M is calculated by dividing the APE by the number of data points in the series. Therefore, if abs is the absolute value, then APE is computed by:

$$\frac{abs(actuals - predictions)}{actuals}$$

where $actuals$ and $predictions$ are different data series. For example, considering data series (actual and prediction) with four (4) data points, as shown in Table 3.4, then the MAPE is estimated as follows:

TABLE 3.4: MAPE Computation Example

	Point1	**Point2**	**Point3**	**Point4**	**Total**
Actuals	150	90	100	80	**420**
Predictions	130	150	90	100	**470**
APE	13%	66%	10%	25%	**12**%
MAPE					**29**%

How we applied MAPE to estimate simulation (SEIR and DES) outputs in the ADRIANA system is highlighted in Section 3.9.1 of Chapter 4.

3.9 Summary

The strategies used in this study were aimed at answering the research questions. The following is a quick synopsis of research process to answer our research questions.

(I) Compared to using the ABMS to simulate a pandemic case, how effective are surrogate models in estimating the future transmission dynamics of COVID-19 with a global perspective?

Our research has shown that we can use compartmental model, DES, and surrogate model, as substitute to ABMS to forecast COVID-19 transmission pattern. The application and interrelation of a compartmental model, discrete event simulation and optimized surrogate significantly influenced the strength of the ADRIANA system. Although, LSTM neural network is the benchmark model for this study, we also considered nine (9) separate models.

(II) Optimizing the surrogate hyperparameters is an essential procedure. How far can we justify this optimization of the surrogate to drive forecasts?

The surrogate optimization gave us the best hyperparameter configuration for optimal result of the ADRIANA system. Our study also evidenced that optimized model outperforms an unoptimized model for prediction. In addition, we calibrated the SEIR parameters to fit the surrogate training and test predictions before running the same configuration for a given number of days (n).

(III) To what degree can the ADRIANA system be calibrated and its predictions used to handle future pandemic prediction problems?

Surrogate models can be calibrated to attain a minimal loss function with improved accuracy. The result obtained by hyperparameters configuration with minimal error identifies the accuracy level that can be attributed to the model for prediction task.

3.9.1 Summary of the Implementation Procedures

This section provides an overview of the implementation specifics explored in our study. It highlights the essential approaches that we applied to obtain our results. The source code for engineering the ADRIANA system can be accessed through this link for reproducible result.

1) We retrieved the daily cumulative COVID-19 CRD datasets for the SIIMCF countries from the data repository of Johns Hopkins University's Center for

Systems Science and Engineering (CSSE) [18]. The datasets were uniform with the COVID-19 cases reported by monitoring platforms such as NICD [5], WHO [5], and International Journal of Infectious Disease [74].

2) We transformed each SIIMCF's confirmed, recovered, and fatality case from (1) into a time series form.

3) As stated in Section 4.6.2, the SEIR model was initialized (before tuning) with random parameters and a starting population N. The parameter configuration of the SEIR model was essential to avoid out-of-range values and ensure quality model simulation. The parameter configuration was consistent to produce results that closely match the real-world datasets extracted in (1).

4) The DES fetched the infectious population from the SEIR result in (3) and simulated the patient hospitalization events from admission through discharge or death as illustrated in Figure 4.25. We used the infectious population in the SEIR MODEL to be COVID-19 confirmed cases. The infectious population modeled by DES produced the number of recoveries and fatalities. We developed a lambda function in python to record the number of recoveries and fatalities simulated by the DES.

5) We used the MAPE function described in Section 3.8.2, and illustrated by Table 4.8 to estimate accuracy of the real-world datasets in (1) with the simulated datasets in steps (3) and (4). Additionally, we used confidence interval technique described in Section 3.8.1 to estimate the acceptable range of values by comparing datasets in (1), and datasets simulated in steps (3) and (4).

6) Next, we performed exploratory data analysis and preprocessed the datasets in steps (1), (3) and (4) to a standard form to allow proper fit by the surrogate for accurate prediction.

7) The surrogate is fitted with three types of datasets which are: (1) The real-world dataset extracted in step 1. (2) The simulated datasets analyzed in step 6, and lastly (3) The combination of real-world dataset and the simulated dataset to form a multivariate input with a univariate output. The surrogate was trained with 70% of the data, and tested with 30% of the data. The data splitting (into training and test) was consistent to all the three categories of datasets used in this study. Next, the surrogate is calibrated to obtain and allocate the best hyperparameter configurations for optimal surrogate result.

8) Next, we applied evaluation metrics on the surrogate, assessed and compared its performance with each of the datasets described in (7). The surrogate evaluation results for the datasets described in (7) are presented in Chapter 4.

9) Finally, we plotted the surrogate and epidemiological results graphically for visual representation.

10) The procedures outlined in (2)-(7) were applied to the simulation experiment of the SIIMCF countries.

The ADRIANA system is made up of the models mentioned in stages (3), (4), and (7) as illustrated in Figure 4.30. In addition, we expanded on the ADRIANA system's architecture by simulating the surrogate with three dataset types as discussed in Section 4.1. Additionally, we compared the results obtained by fitting the surrogate with the three categories of datasets described in Section 4.1.

Note: Relating to the implementation procedure highlighted in Section 3.9.1, and ADRIANA system architecture illustrated by Figure 4.30, the following are noted:
(1) The DES estimates number of recovered and death based on the infectious population (I) of the SEIR model. We considered the infectious population as the confirmed cases.
(2) Considering that our study is concerned with recommendation for hospitals and government regarding COVID-19 transmission, the SEIR and DES capabilities are significantly useful to hospitals in guiding capacity requirements. For example, hospitals can estimate the number of patients expected for ward or ICU admission in a region by configuring DES parameters applicable to them. Likewise, the transmission trend forecast of the surrogate is significantly useful to the government to guide policy development around lockdown measure. For example, governments can estimate when they should increase or decrease lockdown levels based on high or low transmission pattern of COVID-19.
(3) We can trust the ADRIANA system's projection to estimate COVID-19's transmission dynamics, both for patient hospitalization and disease transmission trend, because we carefully followed steps (1)-(7) highlighted in Section 3.9.1.

3.10 Ethical Consideration

We do not require ethical clearance for this study, because our datasets were open-sourced and freely extracted from John Hopkins COVID-19 data repository [2]. Throughout the course of this research, we had no interaction with COVID-19 patients or hospitals.

3.11 Conclusion

This chapter provided the methodology that was explored in our study, and how the research questions have been answered. In chapter 4, we discuss model implementation processes, and results of our study.

Chapter 4

Results and Discussion

4.1 Introduction

The focus of this research was to establish an accurate estimate of the future transmission dynamics of COVID-19 in South Africa (the catchment area of this study) and one country from each continent. We considered South Africa (for Africa), Italy (for Europe), India (for Asia), Mexico (for North America), Colombia (for South America), and Fiji (for Oceania) which we represented as SIIMCF countries.

The transmission dynamics determines the hospital resource demand needed to treat patients as shown in Figure 4.25. There are underlying objectives that must be satisfied to achieve the goal (highlighted in Section 1.3 of Chapter 1) of this study. The objectives of ADRIANA's development were achieved by: (1) We used an SEIR compartmental model to simulate and categorize population into compartment by applying parameter estimation. (2) We used a discrete event simulation to capture COVID-19's patients journey in the hospital, and (3) We developed a surrogate that can present the future transmission trend of COVID-19.

The interaction of the sub-models (described in Section 1.1 of Chapter 1) within the ADRIANA system is illustrated by Figure 4.30. Furthermore, the implementation procedures followed in this study is summarized in Section 3.9.1 of Chapter 3.

We performed the experimental study using ten prediction models, two were statistical models and eight were machine learning models. These models were trained, tested, and evaluated to achieve the best predictions, with the procedures illustrated in Figure 4.1. More specifically, the long short-term memory (LSTM) neural network model serves as the benchmark model. The next subsection discusses the research findings as well as the procedures followed to obtain our research result.

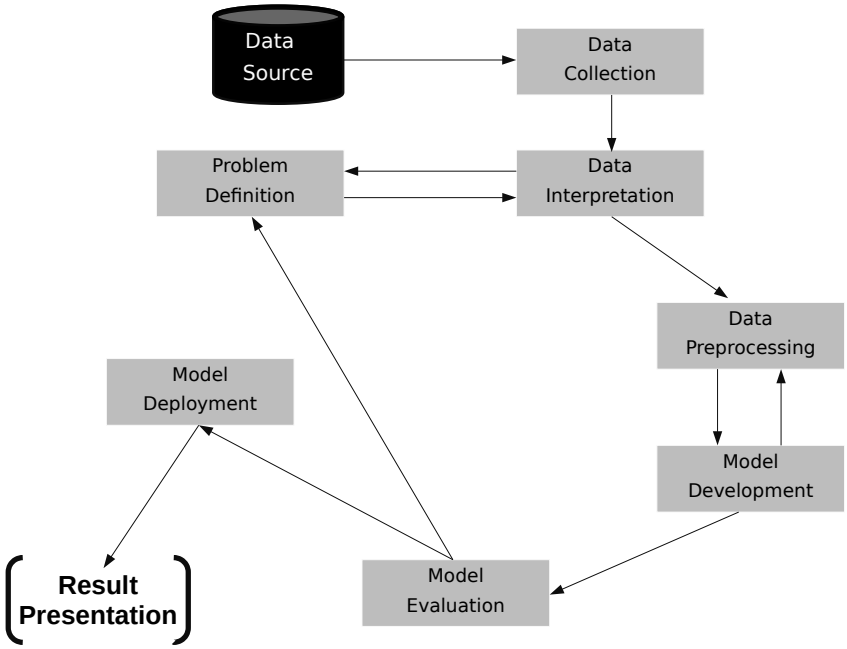


FIGURE 4.1: Synopsis of Model Development Phase

NOTE: The following acronyms describes the operation of the surrogate with the three categories of datasets:
(1) S-REAL: Represents a surrogate fitted with real-world dataset.
(2) S-SIMUL: Signifies a surrogate fitted with datasets from the SEIR and DES simulation experimens.
(3) S-SIREAL: Signifies a surrogate fitted with real-world and simulation datasets to form a multivariate solution as input and gives a univariate series as output. It is noteworthy that the dependent variable (y_i) for the S-SIREAL is computed as the average of the real-world and simulation datasets at any point i in the series.

4.2 Data Collection and Preprocessing

We explored two sources of data in this study, which are the real-world datasets and the simulated datasets. The real-world datasets was extracted from the COVID-19 data repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [18]. The datasets contains the daily cumulative statistics of COVID-19 confirmed cases, recovered cases, and death cases (CRD) for 191 countries across the globe starting from January 22^{nd}, 2020 up till April 23^{rd}, 2021.

The simulated dataset were obtained from the simulation experiment of the SEIR model and discrete event simulator as described in Section 3.9.1. We performed data preprocessing operation for the real-world and simulated datasets. Both the real-world and simulated datasets were presented in time-series format for simulation experiment.

4.2.1 Data Cleaning

The datasets contained the cumulative values of COVID-19 confirmed, recovered, and death cases (CRD) in separate excel files. The applicable libraries in the python framework[1] were used to filter data for the SIIMCF countries from the excel document. We extracted the daily CRD cases by subtracting the cumulative value of the previous day from the current day, (i.e. daily CRD value = cumulative value of today - cumulative value of yesterday) because we fit the surrogate with the daily values.

From the excel files containing CRD cases, the top horizontal axis showed the recorded date for the cases, while the successive rows contained the respective values for different countries. We transformed the datasets to a dataframe (after the focus-country filter is applied) to ensure that both the date column and the confirmed, recovered, and death cases were presented on a horizontal axis. For the most part, the dataframe functions as temporary internal memory storage for the machine learning application framework, in which further analysis can be performed [58].

It should be noted that the CRD cases are applicable to the surrogate model as highlighted in Section 3.9.1. Therefore, we were able to forecast the confirmed, recovered, and death cases for the SIIMCF countries. We initialized the SEIR model parameters before the population was grouped into compartments. Furthermore, the starting dataset used by the DES model (to demonstrate patients' hospitalization) is the population in the infectious compartment of the SEIR model, as explained in Section 3.3 of Chapter 3.

4.2.2 Missing Values

The analysis of the CRD datasets for the SIIMCF countries contains non-missing or null values. Hence, the CRD data presentation was consistent. Furthermore, we examined confirmed, recovered, and death cases as a distinct scenario for the regression analysis.

[1] https://www.python.org/

4.2.3 Data Transformation

Data normalization is an essential aspect of data preparation for modeling operations such as scale-up, scale-down, and numerical feature extraction [58]. Many machine learning models require scaling to achieve their maximum performance. However, the CRD dataset contains data points with different scales. Therefore, for an efficient surrogate training process to learn data features, normalization is considered through scaling.

Data can be normalized using a variety of approaches, including Z-score, decimal scaling, min-max scaling, and integer scaling [75]. We considered the min-max scaling technique. The min-max scaling approach is flexible to outliers and fits the dataset within a data boundary. In our instance, the lower and upper bounds are respectively 0.00001 and 1. The mathematical representation of the min-max featured scalar $\hat{A}$ adopted in our study is:

$$\hat{A} = \left(\frac{A - min(A)}{max(A) - min(A)} \right) \times (d - c) + c$$

where A is a set containing range of values for the dataset, c and d are the preassigned lower bound and upper bound values.

4.3 Implementation Procedures and Experimental Result

In this section, we focus on the implementation specifics and strategies that were explored. We present the observations, assertions, data features, models, and statistical analyses that were explored. First, we investigated the COVID-19 dynamics on a global scale before emphasizing South Africa as the benchmark country of this study.

4.3.1 Exploratory Data Analysis

This section offers a summary of the data analysis techniques used in our work. Our research examined the characteristics of the datasets that are required to fit a surrogate for optimal results. Table 4.1, Table 4.2, Table 4.3, and Table 4.4 shows the analysis and features of the datasets used in our study.

In addition, we assessed the CRD datasets for outliers to estimate the degree of data variation within the datasets. We provide the assessment result in subsection 4.3.4 of this chapter.

Similarly, we applied the Pearson correlation coefficient framework to obtain the relationship between CRD variables. This informed our understanding of the relationship between the confirmed, recovered, and death cases of COVID-19 for each SIIMCF country. Figure 4.2 presents the correlation analysis for South Africa's

TABLE 4.1: Summary of COVID-19 Real-World Dataset for SIIMCF Countries [2]

Country	Cases	Min	Max	STD	Q-0.25	Q-0.50	Q-0.75	Skew	Kurtosis
South Africa	Confirmed	0	21980	4453.9	596	1676.5	4046.3	1.8	2.9
(Africa)	Recovered	0	45858	4822.5	389.7	1441.5	4041.5	3.2	15.5
	Fatalities	0	844	148.7	13	73	151	2.3	5.9
Italy	Confirmed	148	40902	9837.2	454	3492	15336	1.1	0.3
(Europe)	Recovered	883	39266	8987.7	366	1943.5	14998	1.1	0.1
	Fatalities	31	993	250	15	219	456	0.6	0.8
India	Confirmed	1858	346786	48199.6	3934.5	19731	52394.3	3.2	13.8
(Asia)	Recovered	2258	219838	33247.5	1718.8	17991.5	50302.8	1.7	4.4
	Fatalities	39	2624	419.7	90	291	650	1.4	2.7
Mexico	Confirmed	0	28115	4458.9	1526.7	4797	6599	1.5	3.9
(North	Recovered	893	27160	5786.6	1155.5	4011	6258.8	-8.5	141.7
America)	Fatalities	0	3050	415.8	146.5	436.5	669.5	1.6	5.6
Colombia	Confirmed	0	21078	5048.1	646.8	6052	9146	0.5	-0.4
(South	Recovered	6151	89557	7376.9	103.3	4902	9463.8	1.5	54.8
America)	Fatalities	0	430	123.2	18	163	236	0.3	-0.9
Fiji	Confirmed	0	8	0.8	0	0	0	5.6	38.7
(Oceania)	Recovered	0	6	0.6	0	0	0	5.4	33.2
	Fatalities	0	1	0.1	0	0	0	15.1	226.5

TABLE 4.2: Summary of COVID-19 Simulated Dataset for the SIIMCF Countries

Country	Cases	Min	Max	STD	Q-0.25	Q-0.50	Q-0.75	Skew	Kurtosis
South Africa	Confirmed	0	20870	4362.4	534	1772.1	4078.6	1.6	2.3
(Africa)	Recovered	0	45762	4471.4	382.4	1255.1	4341.2	3.1	14.7
	Fatalities	0	682	135.5	11	63	142	2.1	5.3
Italy	Confirmed	136	40550	9347.8	442	3481	15268	1.1	0.5
(Europe)	Recovered	836	38280	8988.4	356	1842.5	14594	1.3	0.2
	Fatalities	30	954	241	15	210	465	0.7	0.9
India	Confirmed	1813	340765	48201.3	3924.2	19633	52388.3	3.4	11.2
(Asia)	Recovered	2290	216398	33310.2	1713.4	17580.8	50350.4	1.6	4.9
	Fatalities	33	2622	420.2	79	267	640	1.3	2.3
Mexico	Confirmed	0	28230	4480.4	1496.3	4396	6560	1.3	-3.2
(North	Recovered	894	27169	5285.4	1130.3	4022	6272.6	7.5	128.3
America)	Fatalities	0	3085	410.2	133.7	422.6	680.3	1.2	5.4
Colombia	Confirmed	0	21046	5034.7	654.2	6044	9154	0.3	0.2
(South	Recovered	6124	89625	7338.4	112.6	4305	9679.6	1.3	-2.1
America)	Fatalities	0	415	132.8	14	173	232	0.4	0.3
Fiji	Confirmed	0	7	0.5	0	0	0	5.3	22.9
(Oceania)	Recovered	0	4	0.3	0	0	0	3.5	28.6
	Fatalities	0	1	0.1	0	0	0	13.3	189.3

CRD dataset, which indicates that the pairwise distribution of the CRD datasets was strongly correlated since all the correlation values are 0.68 or greater. The correlation coefficient (both real-world and simulated datasets) obtained for other SIIMCF countries are presented in Appendix I and Appendix J of this dissertation.

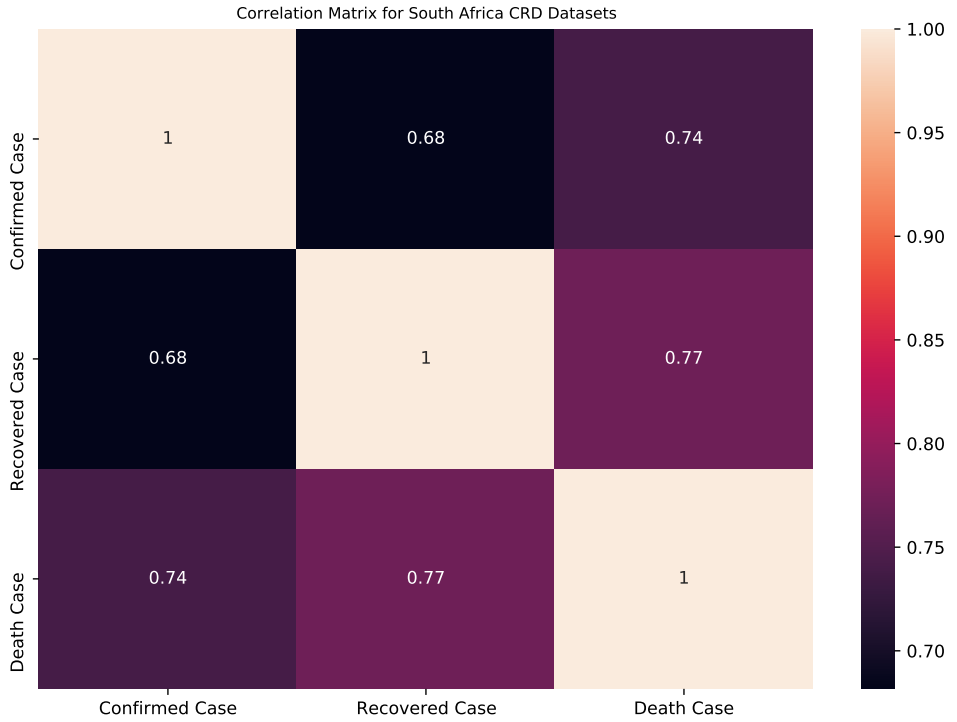


FIGURE 4.2: Correlation Matrix for South Africa CRD Datasets

4.3.2 Correlation Analysis and Stationarity Test

In regression analyses, correlation and covariance are techniques used to identify relationships between two or more variables [65].

Before fitting the surrogate model, we analyzed the CRD datasets for stationarity. Stationarity transforms the series to a constant mean, variance, and probability distribution [76]. Stationarity provided us with better accuracy in the model, as unwanted trends in the dataset are eliminated. We examined stationarity for the CRD dataset using the autocorrelation function (ACF), Partial Autocorrelation Function (PACF) and statistical tests, as shown in Tables 4.3 and Table 4.4.

4.3.2.1 Autocorrelation Function

The Autocorrelation Function (ACF) defines the correlation of a regression parameter with itself at different time-lags [77]. The ACF technique is used in regression

analysis to extract key features in time series analysis and to determine the moving average (MA) component of the ARIMA model [77], [78].

The auto-covariance function of a regression is represented by Equation 4.1, then we estimated the auto-correlation function as:

$$\widetilde{Cv_k} = \frac{1}{\eta}\sum_{p=1}^{\eta-k}(\chi_p - \bar{\chi})(\chi_{p+k} - \bar{x}), \quad \text{for some lag } k, \tag{4.1}$$

We noted that in Equation 4.1, a variance (χ_p) evaluated with the denominator of η equals autocovariance (χ_p) evaluated at $lag = 0$ (i.e $\widetilde{c_0}$). Therefore, we computed the autocorrelation function (**ACF**) as represented in Equation 4.2.

$$\breve{Aut}_k = \frac{\widetilde{Cv_k}}{\widetilde{Cv_0}} = Cor(\chi_p, \chi_{p+k}) \tag{4.2}$$

4.3.2.2 Partial Autocorrelation Function

Partial autocorrelation function (PACF) estimates the linear relation of a sequence (χ_t) with a lagged variance of itself χ_{t+p}, in which linear dependence $(\chi_{t-1}, \chi_{t-2}, ..., \chi_{t-(p-1)})$ has been eliminated [78]. PACF helps to estimate the AR component of the ARIMA model configuration. Mathematically, the PACF can be represented as shown in Equation 4.3.

$$f_j = \begin{cases} Cor(\chi_1, \chi_0) = r_1 & at\ j = 1 \\ Cor(\chi_j - \chi_j^{j-1}, \chi_0 - \chi_0^{j-1}) & at\ j \geq 2 \end{cases} \tag{4.3}$$

We used the visual representation of ACF-plot to determine if the CRD datasets are stationary. The PACF helped us to identify the optimal order of p, d, and q in the ARIMA model as described in subsection 3.6.4 of Chapter 3. The ACF and PACF visual result of confirmed case for South Africa is shown in Figure 4.3, while those of SIIMCF countries are presented in Appendix L and Appendix M of this dissertation.

Analyzing the ACF-PACF plot illustrated by Figure 4.3 for south Africa's confirmed case, the first column on the first row is the visual plot of the original series, the first column on both the second row and third row are the visual plot of the differenced series. The second column on the first row is the ACF statistical test to check if the series is stationary or not. The plot on the second column in the second row confirms that differencing d of order 1 (i.e. $d = 1$) is sufficient to transform the dataset to a stationary series. Finally, the second column on the third row is a PACF plot that suggests the number of previous time-steps the model ($ARIMA$) should consider to predict the current time-step. The same interpretation is consistent for the ACF-PACF representation for the SIIMCF countries as presented in Appendix L of this dissertation.

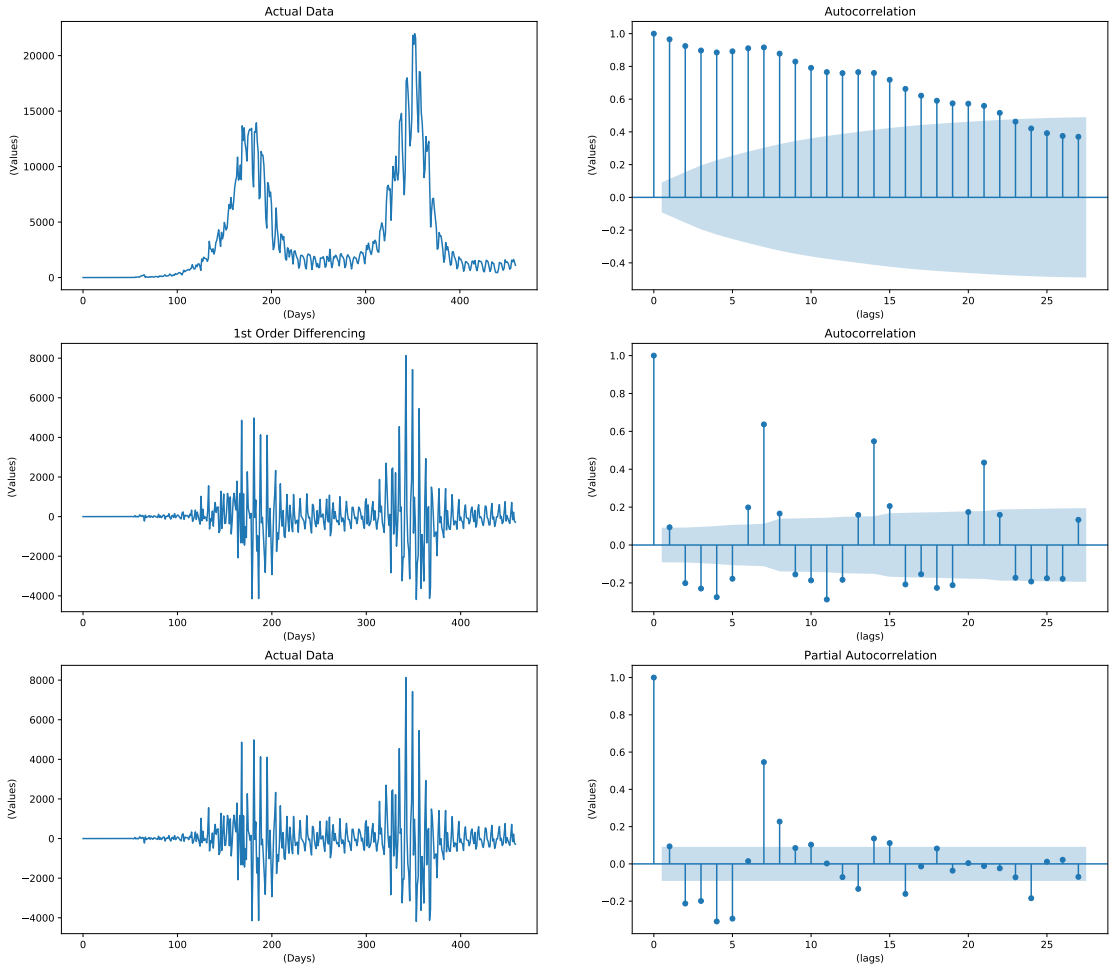


FIGURE 4.3: ACF-PACF Plot of Confirmed Case in South Africa

In addition, to estimate the optimal ARIMA parameters, we computed a mathematical iteration by fitting the ARIMA model on a combination of p, d, q values between the range of $[0, 8]$. Next, we applied a lambda function to extract p, d, q configuration that resulted in lowest Akaike Information Criterion (AIC) value as presented in Tables 4.3 and 4.4. In addition to the ACF and PACF hypotheses, we used the Adfuller statistical test (ADF-test) and the Kwiatkowski–Phillips–Schmidt–Shin test (KPSS-test) to also verify stationarity in the SIIMCF CRD datasets.

The ADF-test proposes that the regression series has a unit root feature, which implies the series is not stationary. We applied the ADF hypothesis by scanning the CRD datasets and estimated the probabilistic value (p-value) against a 0.05 significance level. In essence, if the p-value is greater than 0.05, then we accept the null hypothesis (H_0) that the series is not stationary, otherwise, we reject the null hypothesis (H_1) [75]. The null hypothesis indicate that there exist a unit root in the time series data [79]. The unit root is a time series property which determines non-stationarity of a time series dataset.

Additionally, we applied KPSS-test to check for stationarity. To evaluate KPSS

test, if the $p < 0.05$ significance level, we accept the hypothesis H_0 that the series is non-stationary, and reject the hypothesis H_1 if $p > 0.05$ significance level.

4.3.3 Transforming Data to Stationarity

We can use several statistical methods to transform CRD datasets to stationary series, including differencing, seasonal adjustment, transformation, and de-trending. However, we only considered differencing in our study as it was sufficient for our experiment. We estimated the order of differencing d by subtracting a value at a previous d^{th} data point from a value at a current data point t. Differencing technique is mathematically expressed by $\hat{X}_t = X_t - X_{t-d}$, where X_t is the current data point on the series, X_{t-d} is the previous d^{th} data point, and d is the order of differencing.

After analyzing the CRD datasets with ACF, ADF-test, and KPSS-test, we identified that CRD datasets for the SIIMCF countries are not stationary. Therefore, the datasets must be transformed to a stationary series before we train the models. The ACF and PACF stationarity tests are visually presented in Section 4.3.2 and Appendix L. Additionally, the p-value and minimum AIC parameter estimated for the ADF-test and KPSS-test are presented in Tables 4.3 and 4.4.

4.3.4 Outlier Data

Detecting Outliers is an essential procedure in time series analysis [78]. Outliers are data points that deviates significantly from the rest of the datasets [77], [78]. Outliers can be caused by reasons such as, theoretical estimation error and data presentation [80]. Outliers must be statistically assessed and eliminated to minimize skewed model performance. Outliers are evident in the CRD datasets examined in this study, as seen in Figure 4.4 (South Africa CRD datasets). We identified that scaling datasets (real-world and simulated data) to stationarity is a major technique to eliminate outliers from the CRD datasets. Data outlier analysis for other SIIMCF countries is presented in Appendix K of this dissertation.

4.4 Schematic Analysis of COVID-19 Global Scale

Our study considered predicting the future transmission trend of COVID-19 globally. We examined one country from each of the world's seven continents, excluding Antarctica because of its shared boundaries with other surrounding countries and inability to retrieve its datasets [4]. Additionally, we evaluated the future dynamics of the CRD cases between the SIIMCF countries to identify the future trend of the disease globally. In Table 4.5, we presented the major economic and social milestones of COVID-19 globally, from November 2019 to February 2021.

TABLE 4.3: Highlight of CRD Statistical Test for the SIIMCF Countries

Country	**Cases**	**ADF Test**	**KPSS Test**	**AR-Lag**
South Africa (Africa)	Confirmed	p-value = 0.12 Min AIC = 5,087.7	p-value = 0.04	1
	Recovered	p-value = 0.98 Min AIC = 4,714.1	p-value = 0.03	1
	Death	p-value = 0.96 Min AIC = 2,708.2	p-value = 0.02	1
Italy (Europe)	Confirmed	p-value = 0.47 Min AIC = 8,387.1	p-value = 0.01	2
	Recovered	p-value = 0.63 Min AIC = 5,874.6	p-value = 0.01	1
	Death	p-value = 0.07 Min AIC = 3,555.10	p-value = 0.03	1
India (Asia)	Confirmed	p-value = 0.82 Min AIC = 6,022.7	p-value = 0.01	1
	Recovered	p-value = 0.96 Min AIC = 3,495.1	p-value = 0.01	2
	Death	p-value = 0.72 Min AIC = 4,064.1	p-value = 0.02	1
Mexico (North America)	Confirmed	p-value = 0.67 Min AIC = 8,590.1	p-value = 0.03	1
	Recovered	p-value = 0.39 Min AIC = 6,572.2	p-value = 0.01	1
	Death	p-value = 0.32 Min AIC = 4,475.3	p-value = 0.01	1
Colombia (South America)	Confirmed	p-value = 0.64 Min AIC = 7,792.7	p-value = 0.01	1
	Recovered	p-value = 0.30 Min AIC = 6,709.4	p-value = 0.01	1
	Death	p-value = 0.28 Min AIC = 3,322.6	p-value = 0.01	1
Fiji (Oceania)	Confirmed	p-value = 0.56 Min AIC = 1,188.2	p-value = 0.02	1
	Recovered	p-value = 0.23 Min AIC = 535.2	p-value = 0.03	1
	Death	p-value = 0.21 Min AIC = 582.7	p-value = 0.02	1

TABLE 4.4: Highlight of Simulated CRD Statistical Test for SIIMCF Countries

Country	**Cases**	**ADF Test**	**KPSS Test**	**AR-Lag**
South Africa (Africa)	Confirmed	p-value = 0.11 Min AIC = 4,8067.4	p-value = 0.03	1
	Recovered	p-value = 0.65 Min AIC = 4,674.6	p-value = 0.04	1
	Death	p-value = 0.85 Min AIC = 2,452.3	p-value = 0.01	1
Italy (Europe)	Confirmed	p-value = 0.37 Min AIC = 8,201.3	p-value = 0.01	1
	Recovered	p-value = 0.34 Min AIC = 5,454.3	p-value = 0.03	1
	Death	p-value = 0.06 Min AIC = 3,576.30	p-value = 0.02	1
India (Asia)	Confirmed	p-value = 0.32 Min AIC = 5,687.2	p-value = 0.01	1
	Recovered	p-value = 0.66 Min AIC = 3,211.4	p-value = 0.01	1
	Death	p-value = 0.35 Min AIC = 4,039.1	p-value = 0.02	1
Mexico (North America)	Confirmed	p-value = 0.76 Min AIC = 7,549.6	p-value = 0.02	2
	Recovered	p-value = 0.42 Min AIC = 6,980.3	p-value = 0.01	1
	Death	p-value = 0.22 Min AIC = 4,385.3	p-value = 0.01	1
Colombia (South America)	Confirmed	p-value = 0.44 Min AIC = 7,721.7	p-value = 0.01	1
	Recovered	p-value = 0.20 Min AIC = 6,307.4	p-value = 0.01	1
	Death	p-value = 0.23 Min AIC = 3,212.4	p-value = 0.01	1
Fiji (Oceania)	Confirmed	p-value = 0.42 Min AIC = 1,167.2	p-value = 0.01	1
	Recovered	p-value = 0.23 Min AIC = 504.2	p-value = 0.02	2
	Death	p-value = 0.33 Min AIC = -584.7	p-value = 0.01	2

TABLE 4.5: COVID-19 Major Milestones on a Global Scale (Nov, 2019 to Feb, 2021)

Date	**Location**	**Milestone**
17th November, 2019	China	First confirmed COVID-19 case globally [8].
28th January, 2020	South Africa	NICD commenced COVID-19 testing on citizens [5], [7].
12th February, 2020	Barcelona	World's largest mobile phone exhibition, Mobile world congress (WWC) cancelled over COVID-19 concerns [2].
13th February, 2020	Spain	First death of COVID-19 recorded globally [2], [8]
5th March, 2020	South Africa	First confirmed case of COVID-19 recorded [5], [7].
10th March, 2020	Italy	Italian prime minister announced partial lockdown [2], [8].
11th March, 2020	World	WHO declared COVID-19 as a pandemic [8].
17th March, 2020	South Africa	Authority for regulating health product announced its intention to expedite review of vaccines, treatments and clinical trials [5], [8].
23rd March, 2020	China	China stops quarantine in Wuhan after two months of total lockdown [2], [5], [8]
23rd March, 2020	United Kingdom	Prime minister of the United Kingdom declared partial lockdown [2].
26th March, 2020	South Africa	First total lockdown (level 5) enforced by government [5].
27th March, 2020	South Africa	First death of COVID-19 in South Africa [5], [7].
March 2020 - Nov 2020	South Africa	First wave of COVID-19 in South Africa [5], [7].
28th March, 2020	Spain	Government enforced stricter lockdown measure [2], [8].
1st February, 2021	South Africa	Arrival of first set of vaccine [5].
December 2020 - April 2021	South Africa	Second wave of COVID-19 in South Africa [5], [7].
17th February, 2021	South Africa	Administering of first set of vaccine [5].

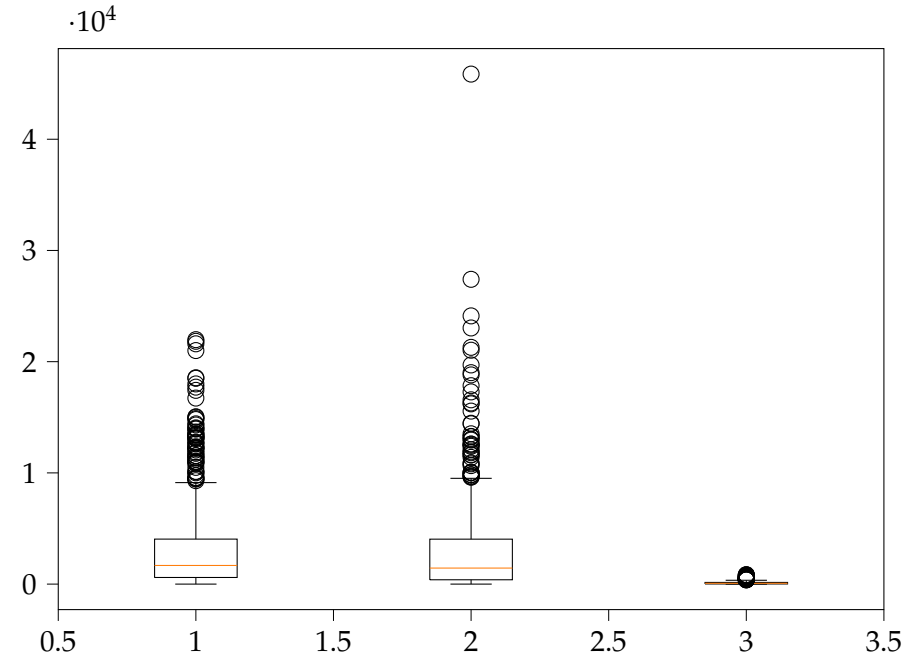


FIGURE 4.4: Outlier Detection of South Africa CRD Dataset

The monthly figures of COVID-19 confirmed, recovered, and death cases for the SIIMCF countries are presented visually in Figure 4.5, Figure 4.6, and Figure 4.7 respectively.

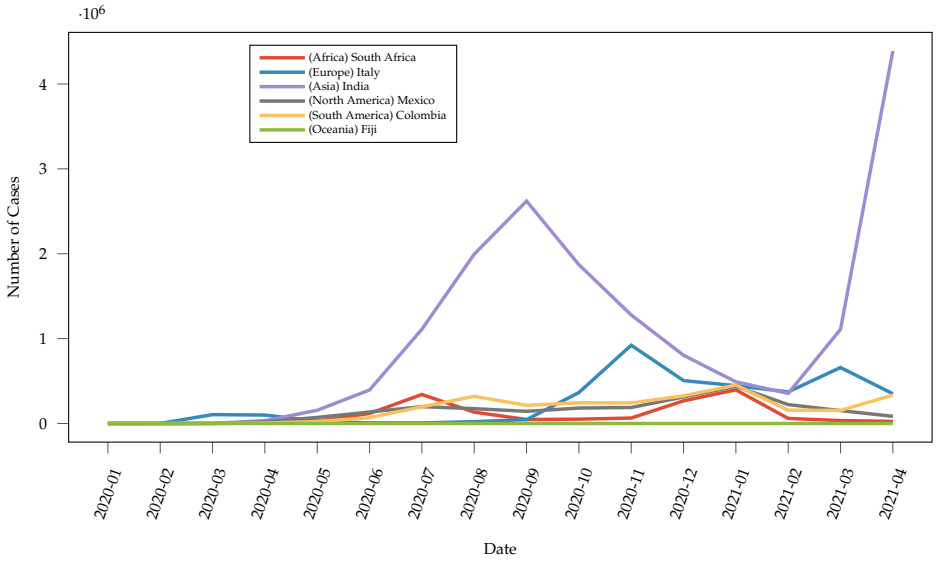


FIGURE 4.5: Intercontinental Monthly Dynamics of COVID-19 Confirmed Case (January 2020 - April 2021) [2]

4.4.1 Schematic Analysis of COVID-19 in Africa

The COVID-19 pandemic has expanded throughout the continents. We extended our research analysis to capture the existing key findings of the transmission pattern of COVID-19 in Africa. We considered Africa's COVID-19 analysis because South

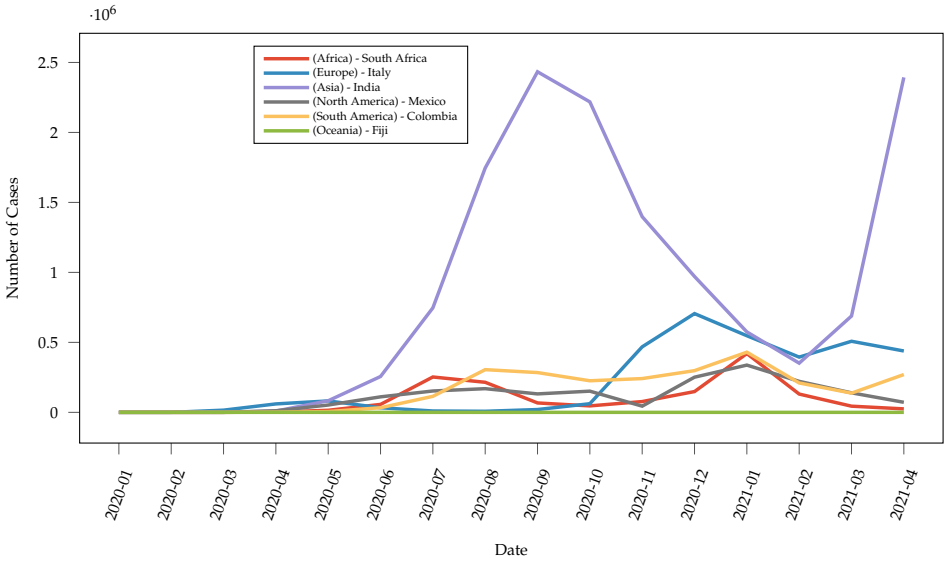


FIGURE 4.6: Intercontinental Monthly Dynamics of COVID-19 Recovered Case (January 2020 - April 2021) [2]

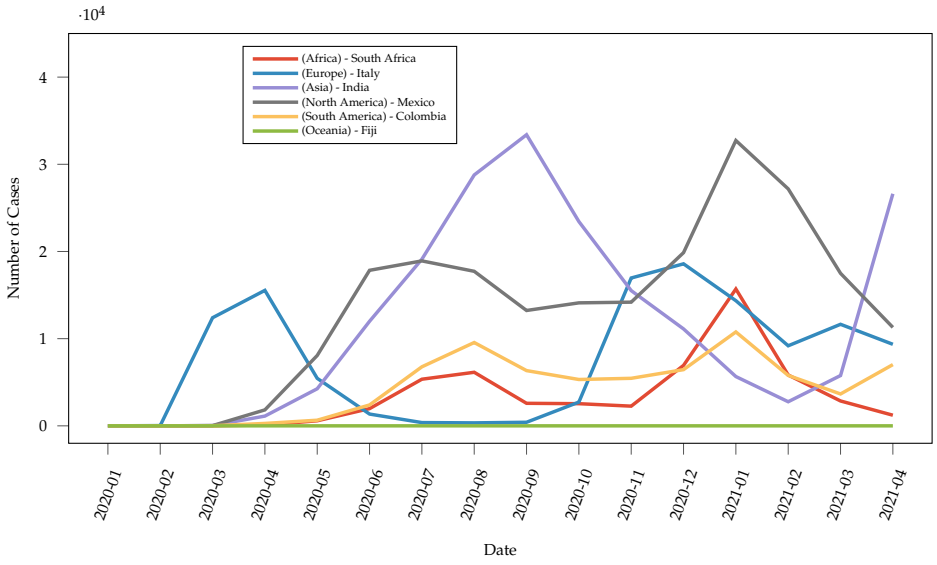


FIGURE 4.7: Intercontinental Monthly Dynamics of COVID-19 Death Case (January 2020 - April 2021) [2]

Africa as our study's catchment area, is located within Africa. According to a study conducted by a team of African-based researchers [81], acknowledged by the World Health Organization (WHO [8]), and reported by the British Broadcasting Corporation (BBC [4]), the top five (5) countries affected by COVID-19 in Africa is presented in Table 4.6.

Figure 4.8 is the schematic diagram showing the monthly cumulative confirmed cases, recovered cases, and death cases for top five (5) countries affected by COVID-19 in Africa.

TABLE 4.6: Top Five (5) African Countries Affected by COVID-19 [81]

Number	Confirmed Cases	Active Cases	Recovered Cases	Fatality Cases
1	South Africa	South Africa	South Africa	South Africa
2	Morocco	Algeria	Morocco	Egypt
3	Tunisia	Ethiopia	Tunisia	Morocco
4	Egypt	Libya	Egypt	Tunisia
5	Ethiopia	Morocco	Ethiopia	Algeria

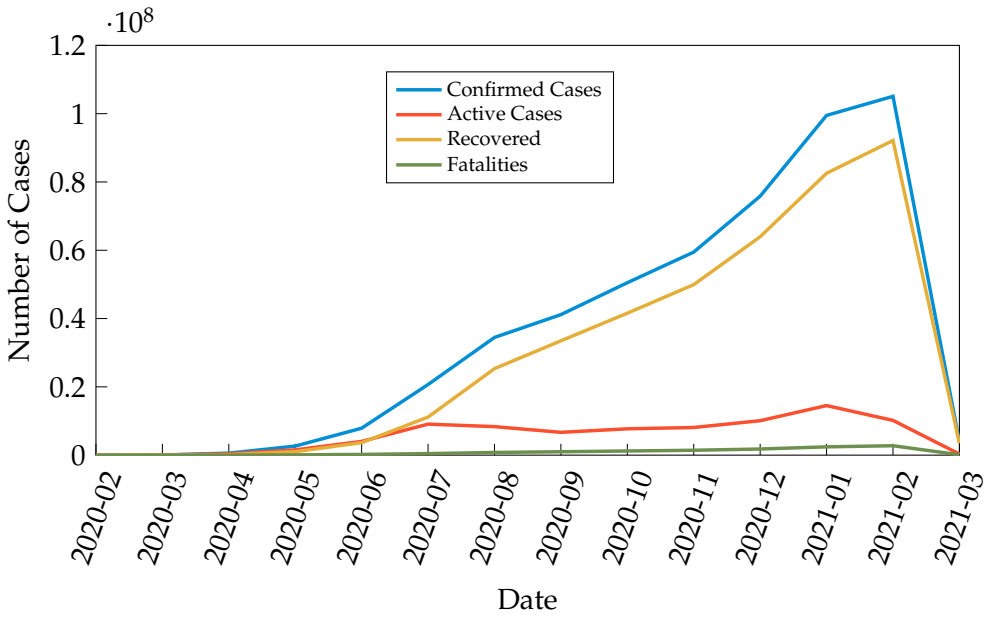


FIGURE 4.8: Cumulative Confirmed, Active, Recovery and Fatal Cases in Africa (February 2020 - March 2021) [82]

4.4.2 Schematic Analysis of COVID-19 in South Africa

Policymakers in South Africa considered lockdown as one of the major intervention strategies to reduce the transmission of COVID-19. In Table 4.7, we presented the average percentage daily increase of COVID-19's confirmed cases in South Africa for different lockdown levels [81], [82].

Additionally, in Figure 4.14, we illustrated the different periods of lockdown against the CRD cumulative case in South Africa. The information was acquired from the portals of the National Institute of Communicable Diseases (NICD) in South Africa [5], and World Health Organization (WHO) in South Africa [8]. From March 2020 to April 2021, the lockdown stages range from alert level-5 to alert level-1, representing the decreasing degree of public mobility strictness [7], [8].

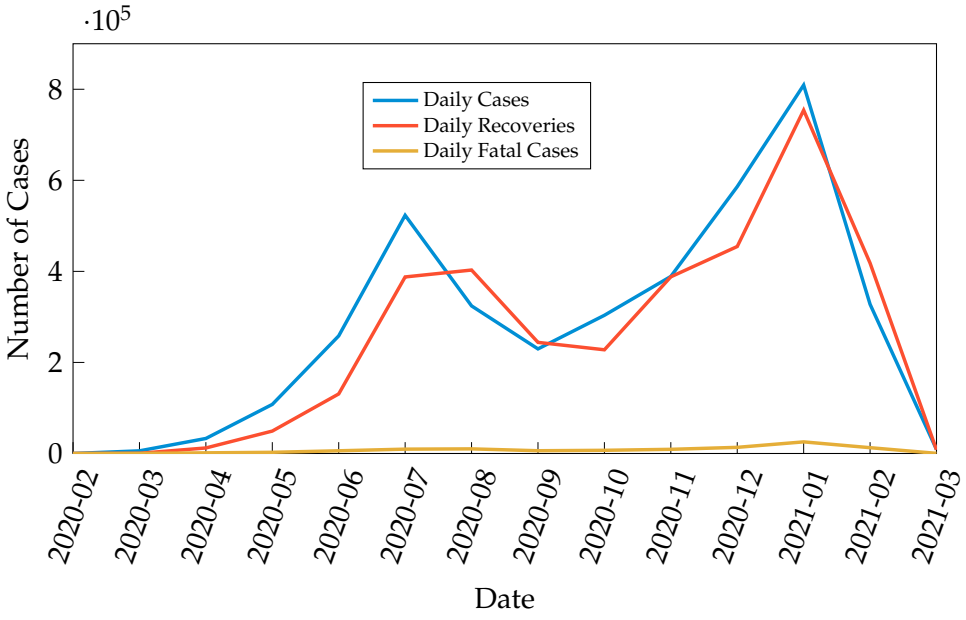


FIGURE 4.9: Daily Changes in Confirmed, Recovery, and Fatal Cases in Africa (February 2020 - March 2021) [82]

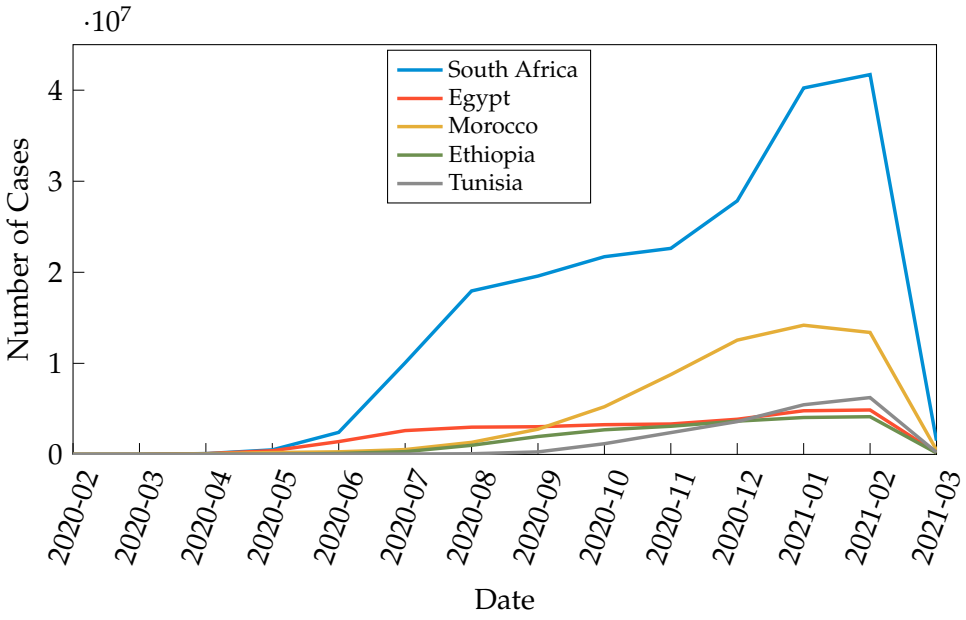


FIGURE 4.10: African Countries with Top Five Highest Confirmed Cases (February 2020 - March 2021) [82]

4.4.3 COVID-19 Risk Index Analysis in South Africa

The risk index primarily describes the overall result of the risk assessment of an event [83]. The risk index examined in our study is focused on fundamental risk factors of COVID-19 disease. They include variables that exist before the pandemic which has influenced the transmission dynamics of COVID-19 in South Africa [84]. In addition, risk index is applied to identify locations where the secondary impact of COVID-19 will be most severe [83], [84].

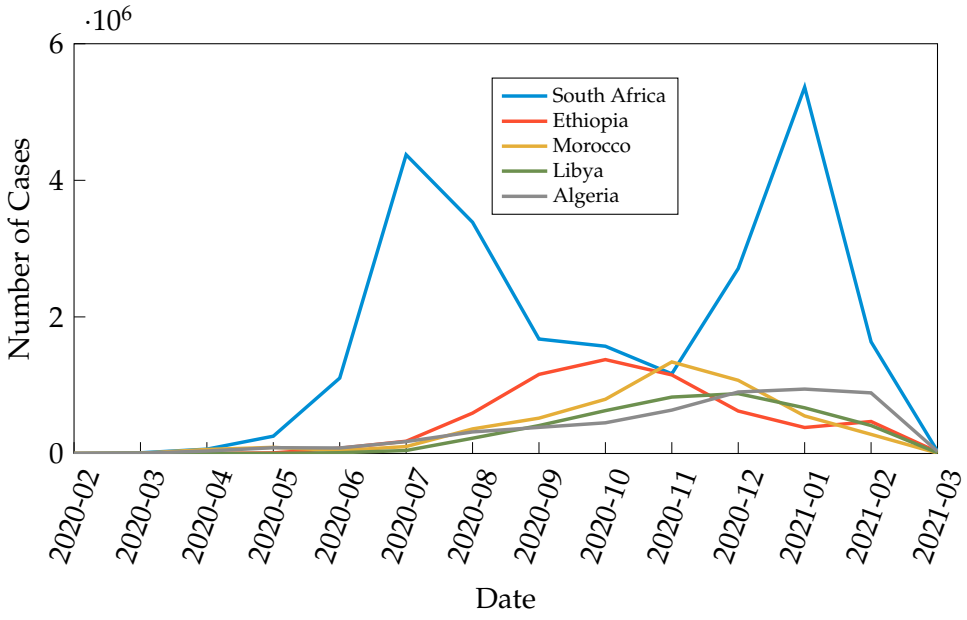


FIGURE 4.11: African Countries with Top Five Highest Active Cases (March 2020 - March 2021) [82]

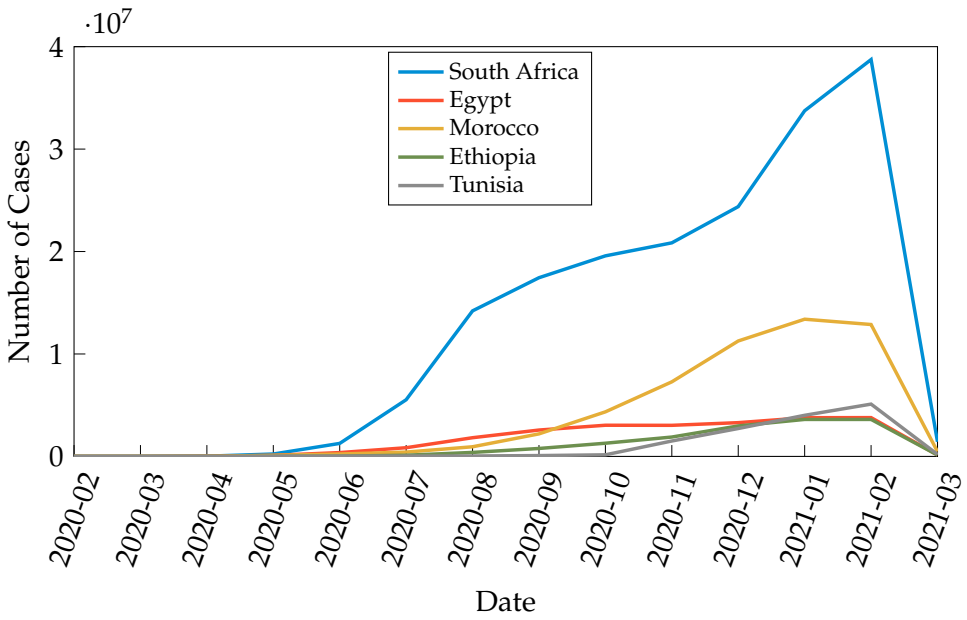


FIGURE 4.12: African Countries with Top Five Highest Recovery Cases (February 2020 - March 2021) [82]

To assess the risk index of COVID-19's third wave in South Africa, computing a risk index greater than 10 indicates that the predicted third-wave is statistically accepted. Generally, risk indexes with values of 10 and 7.5 reflect high and medium risk thresholds respectively [84].

The mathematical expression to estimate COVID-19's risk index in the SIIMCF countries are represented by Equation 4.4.

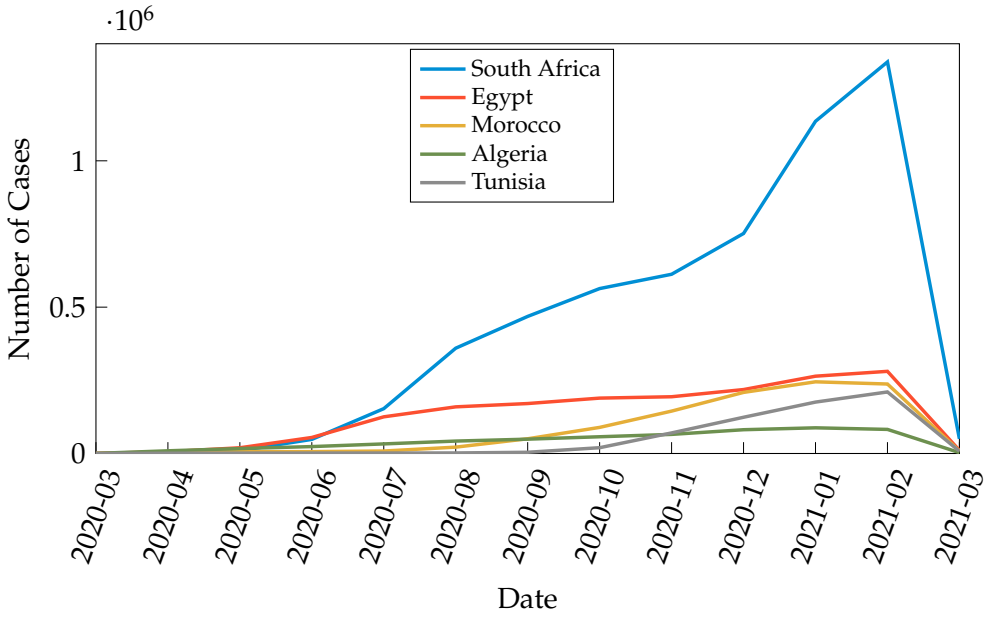


FIGURE 4.13: African Countries with Top Five Highest Fatal Cases (March 2020 - March 2021) [82]

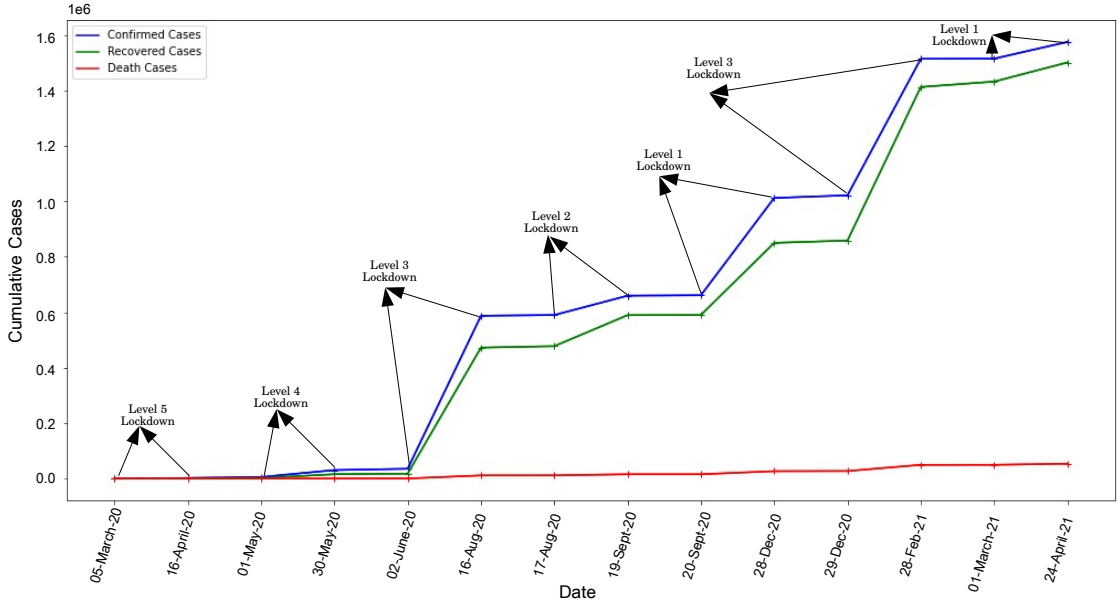


FIGURE 4.14: Lockdown Measures Relating to COVID-19 Confirmed Cases in South Africa - (March, 2020 to April 2021) [82]

$$\begin{aligned} \tilde{x}_i &= 100\left(\frac{\hat{a}_i - \hat{p}_i}{\hat{p}_i}\right) \\ \tilde{y}_i &= 10\left(\frac{1}{1+\frac{\hat{a}_i}{\hat{t}_j}}\right) \end{aligned} \tag{4.4}$$

$$\textbf{if } \tilde{y}_i > 8,\ risk\,index = \tilde{x}_i + \tilde{y}_i;\ \textbf{else}\ \ risk\,index = \tilde{x}_i$$

Where $\hat{a}_i$ represents the actual cases, $\hat{p}_i$ represents the present day prediction, $\hat{t}_j$

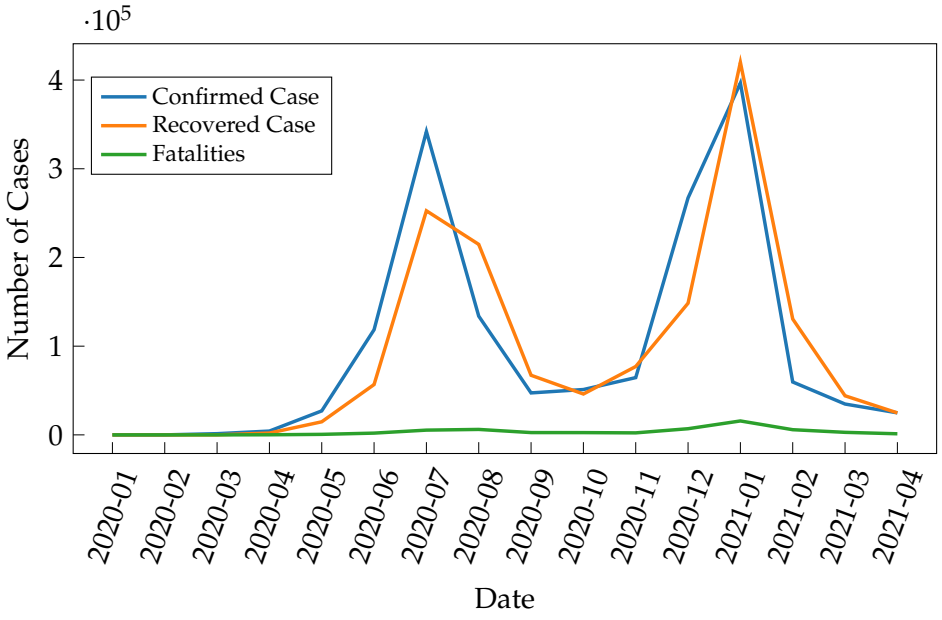


FIGURE 4.15: Monthly Confirmed, Recovered, and Death Cases in South Africa (January, 2020 - April 2021) [82]

TABLE 4.7: Average Daily Increase of COVID-19 Confirmed Case Under Different Lockdown Levels in South Africa [81]

Lockdown Level	**Value** (%)
Before Lockdown	0.01 - 41.43
Level 1	0.96 - 5.37
Level 2	0.93 - 5.38
Level 3	0.65 - 3.80
Level 4	0.34 - 0.36
Level 5	0.10 - 0.43

represents total prediction. The total prediction $\hat{t}_j$ and present-day prediction $\hat{p}_i$ are evaluated by using the susceptible infectious (SI) model technique. The SI model describes an epidemiological model that uses a first-order differential equation to evaluate the patterns of infectious disease over a defined time duration [53].

According to the SI model, each member in the susceptible population has an equal and high risk of becoming infected if they contact an infected individual. The fundamental parameters in the SI model to measure disease transmission are S_o and I_o at a given time $n = 0$, which represents the number of susceptible and infected populations, respectively.

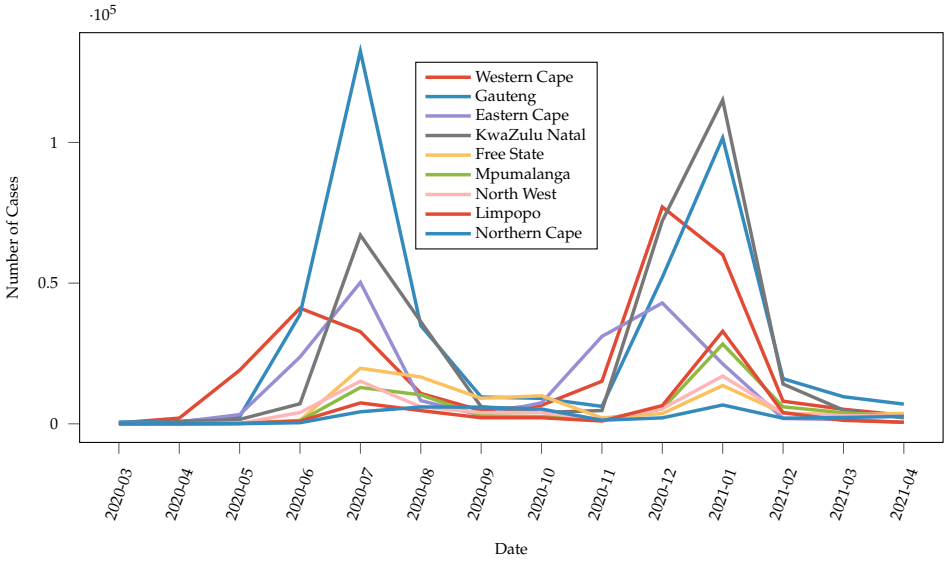


FIGURE 4.16: Monthly Observation of COVID-19 Confirmed Case Across the Provinces in South Africa (March, 2020 - April 2021) [82]

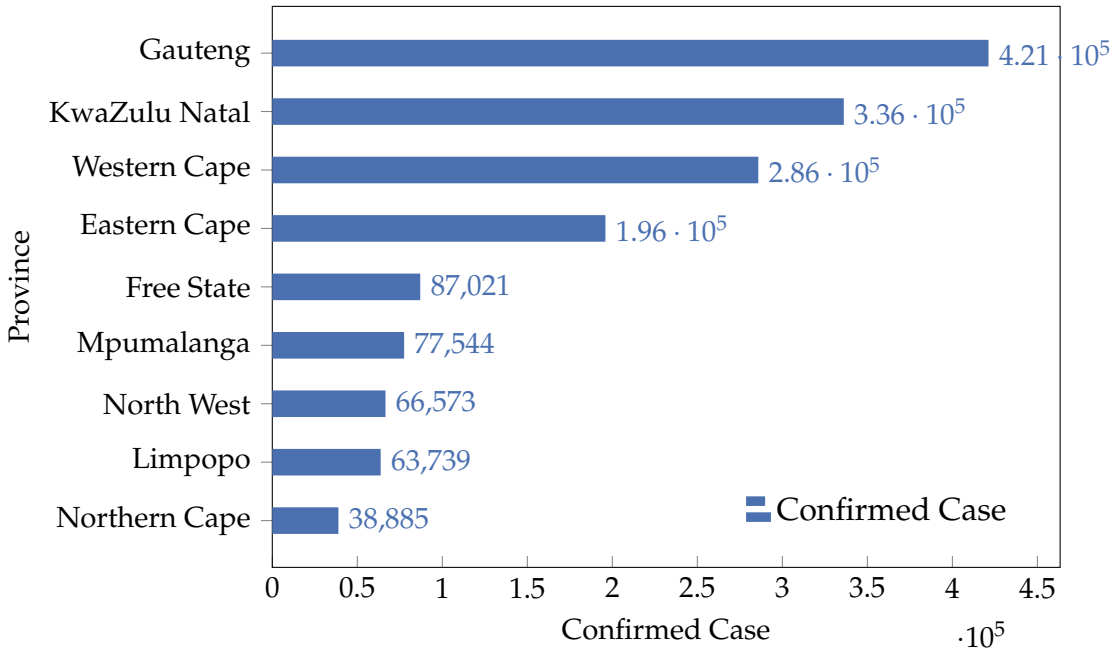


FIGURE 4.17: Cumulative Confirmed Cases by Province in South Africa (March, 2020 - April 2021) [82]

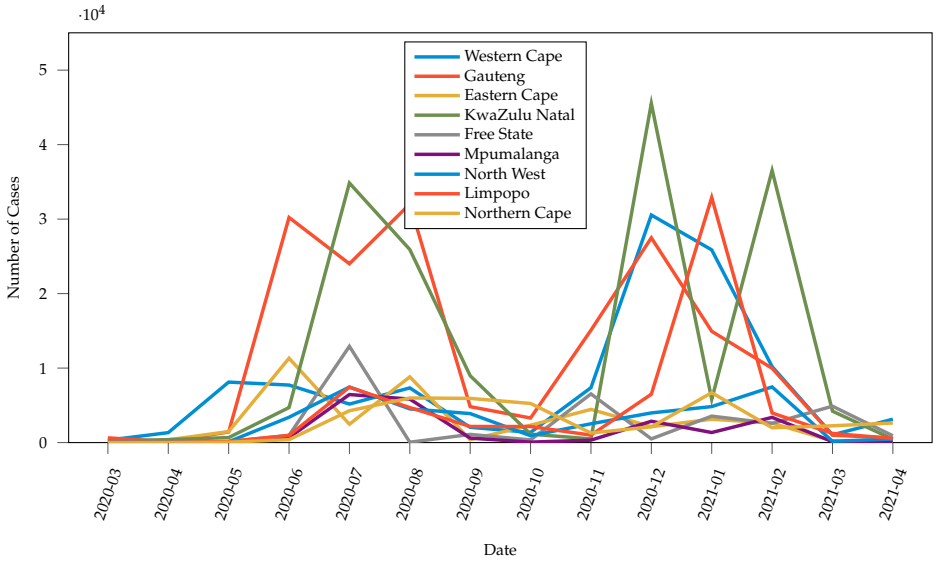


FIGURE 4.18: Monthly Observation of COVID-19 Recovered Case Across the Provinces in South Africa (March, 2020 - April 2021) [82]

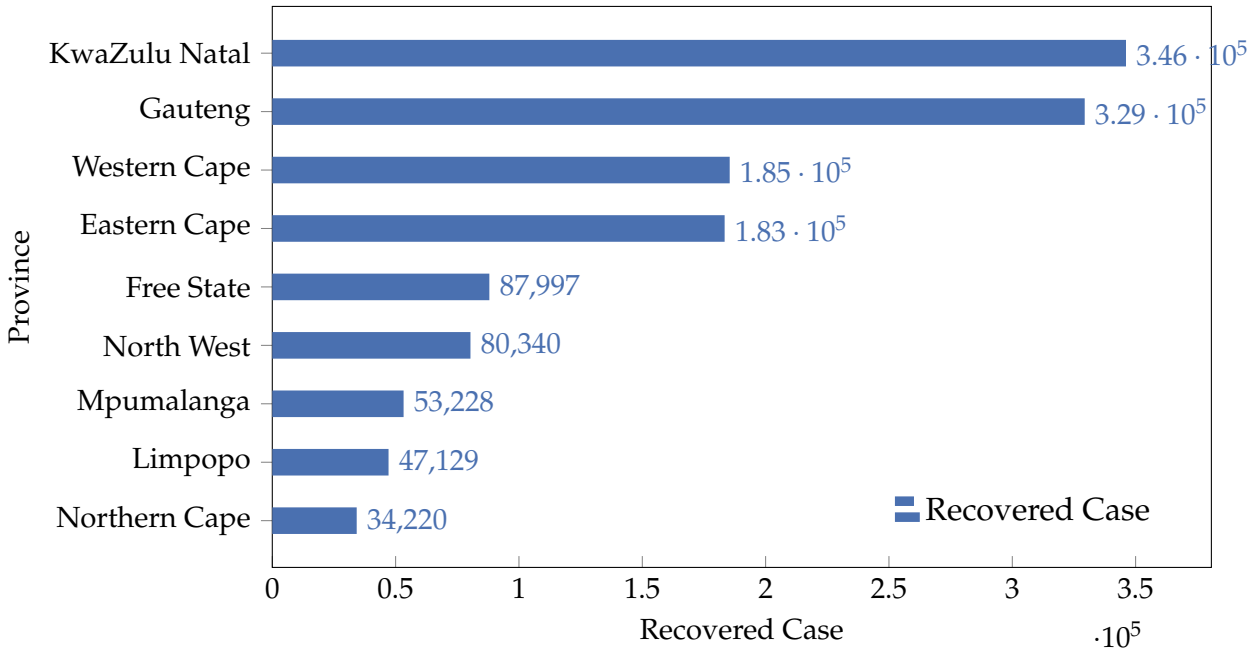


FIGURE 4.19: Cumulative Recovered Cases by Province in South Africa (March, 2020 - April 2021) [82]

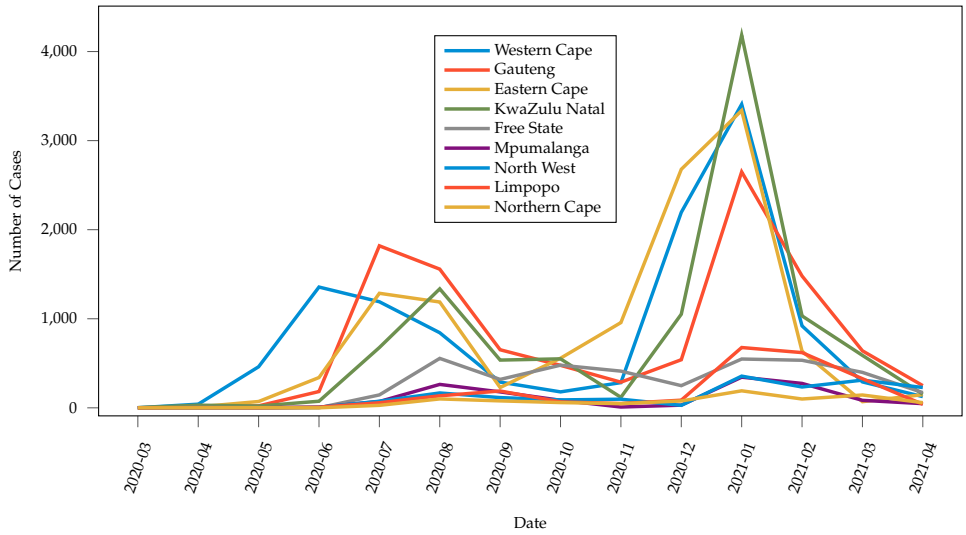


FIGURE 4.20: Monthly Observation of COVID-19 Death Case Across the Provinces in South Africa (March, 2020 - April 2021) [82]

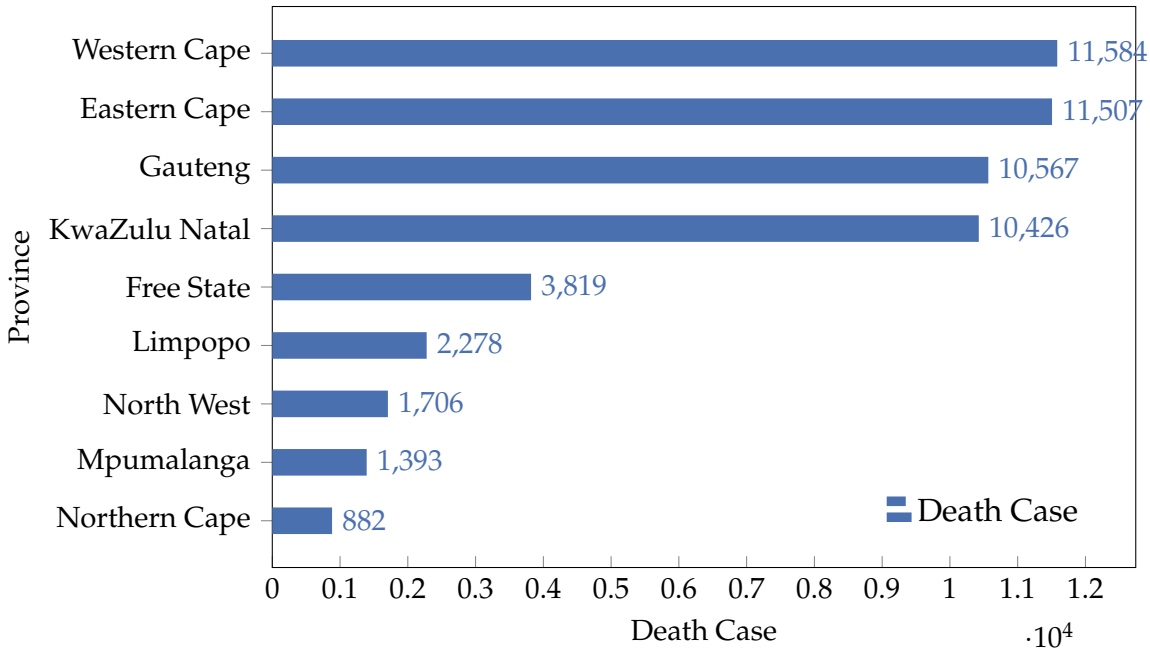


FIGURE 4.21: Cumulative Death Cases by Province in South Africa (March, 2020 - April 2021) [82]

4.4.3.1 Risk Index Evaluation for South African Provinces

Figure 4.22 shows the analysis of the COVID-19 risk index for the nine (9) provinces of South Africa. These values were established by a team of South African researchers [82].

4.4.3.2 Why Did We Examine the Risk Index?

Considering the impact of measuring the risk index in managing COVID-19 assists in predicting hospital resource demand which informs hospital capacity planning. In our study, the provincial risk analysis was undertaken to offer a computational perspective and influence of each province to the dynamics of COVID-19's third-wave in South Africa. The scope of measuring the risk index in this study does not include predicting the social vulnerability impact of COVID-19 (for example, the effect of COVID-19 in evaluating conflict risk within the population) and does not consider the mechanism that influences the secondary impact of the pandemic [83].

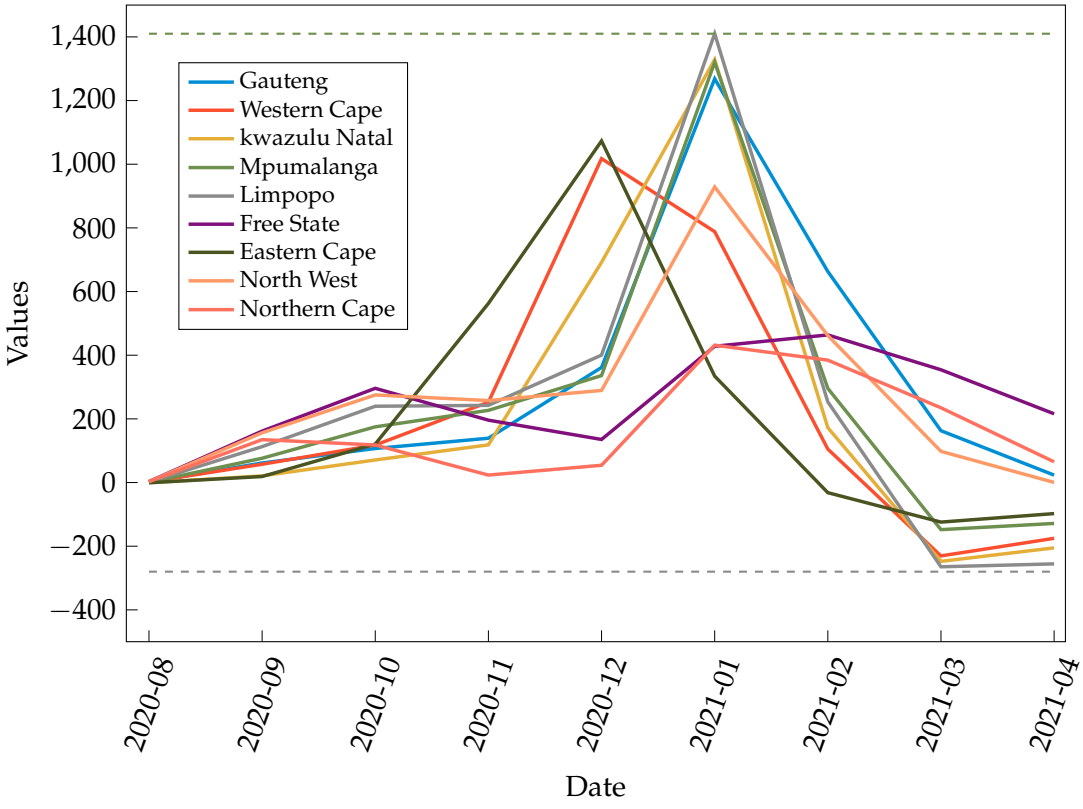


FIGURE 4.22: Risk Index of South Africa Provinces (August 2020 - April 2021) [82]

Figure 4.22 captures the various risk index distribution of COVID-19 across South African provinces. We observed an exponential rise of risk index corresponding to the confirmed cases recorded within the specified period.

TABLE 4.8: MAPE Computation for SIIMCF CRD Dataset

Country	Cases	MAPE	Accuracy (100 - MAPE)
South Africa	Confirmed	9.7%	90.3%
	Recovered	8.6%	91.4%
	Death	12.2%	67.8%
Italy	Confirmed	11.1%	90.3%
	Recovered	13.6%	86.4%
	Death	10.4%	89.6%
India	Confirmed	5.9%	94.1%
	Recovered	13.1%	86.9%
	Death	9.4%	90.6%
Mexico	Confirmed	8.7%	91.3%
	Recovered	13.3%	86.7%
	Death	12.2%	87.8%
Colombia	Confirmed	12.8%	87.2%
	Recovered	13.1%	86.9%
	Death	6.9%	93.1%
Fiji	Confirmed	2.1%	97.9%
	Recovered	6.9%	93.1%
	Death	3.2%	96.8%

We analyzed the risk projection based on current social and economic factors that contributed to disease spread. According to Figure 4.22, five provinces have similar high-risk index patterns, followed by other provinces with different risk level. Lastly, from Figure 4.22, the COVID-19's risk index was high when the national lockdown in South Africa was eased to a minimum level and low when the lockdown was enforced.

4.4.4 MAPE Computation and Result

As discussed in Section 3.8.2 of Chapter 3, we applied MAPE to evaluate accuracy of dataset simulated by the SEIR model and DES against real-world data. We used 15% as the maximum MAPE error that can be accepted to ensure valid forecast of the surrogate. Whenever we compute a MAPE error greater than 15%, we reconfigured the SEIR and DES parameters, and re-simulate the models until we achieve a MAPE less than or equal to 15% as illustrated in Figure 4.30. We presented the estimated MAPE computation for the SIIMCF countries in Table 4.8.

4.5 Hyperparameter Selection

Our study focused on optimization by analyzing and selecting the best surrogate hyperparameters using an optimization algorithm before fitting them to datasets. We adjusted the surrogate using the selected hyperparameters to get satisfactory

results. Table 4.9 presents the surrogate hyperparameter selection for South Africa CRD datasets, whereas Appendix B provides the surrogate hyperparameter selection for other SIIMCF countries.

4.6 Experimental Result

In this section, we show the result of the model fusion investigated in our study. We presented the CRD forecast graphs for the SIIMCF countries. In addition, in relation to our project's goals, we proposed the ideal period for policymakers to be strict or reduce lockdown action. Furthermore, our findings provide hospitals with a prediction (based on real-world datasets) of the peak time when hospital resources are likely to be underused or overwhelmed as shown in Table 4.10 and Table 4.11.

Similarly, the future predictions showing the estimate of COVID-19 CRD's peak period computed with the simulated datasets (S-SIMUL), and hybrid of simulated and real-world datasets (S-SIREAL) are presented in Appendix H of this dissertation. The forecasted day range (e.g 10 ± 5) is applied to allow for the model's flexibility of the predicted days.

4.6.1 Results Presentation

The findings of the ADRIANA model fusion is provided in this section, which comprises the SEIR model, discrete event simulation and surrogate result. The model implementation summary is presented in Section 3.9.1. Additionally, the key features of the ADRIANA system is shown in Figure 4.23.

4.6.2 SEIR Compartmental Model Result

Our study demonstrated the implementation of the SEIR compartmental model by grouping populations into COVID-19 epidemiological compartments for the SIIMCF countries. The model parameters were initialized at a realistic value for valid model simulation. The exposed compartment $E(t)$ of the model was initialized with population value $E(t) = 1$, infectious compartment $I(t)$ was initialized with population value of $I(t) = 1$, and recovered compartment was initialized with population value $R(t) = 0$.

The total population N was initialized at 150,000. The susceptible compartment was initialized with population $S(t) = N - (E(t) + I(t) + R(t))$, the infection rate ϵ was initialized as 1.38, the incubation rate $1/\epsilon$ initialized as 0.19, reproduction number R_0 initialized as 4, recovery rate γ was initialized as 0.34, and number of days initialized as 150 days. Each of the parameters can be adjusted to decode the effect of modifying the model's internal parameters to attain the desired result. Figure 4.24 shows the overview of the SEIR simulation output after being initialized

TABLE 4.9: Optimal Model Hyperparameter Selection for South Africa

		S-REAL			S-SIMUL			S-SIREAL		
Model	**Hyperparameter**	**C**	**R**	**D**	**C**	**R**	**D**	**C**	**R**	**D**
LSTM,	Neurons	58	51	56	49	42	54	46	66	35
GRU	Window Size	9	11	6	5	6	8	10	11	9
	Learning Rate	0.01	0.01	0.01	0.04	0.04	0.04	0.03	0.03	0.03
MLP	Batch Size	64	53	55	61	62	71	59	54	44
	Unit	32	24	16	18	16	22	21	12	14
	Learning Rate	0.01	0.01	0.03	0.01	0.03	0.02	0.01	0.06	0.01
	Solver	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam
XGBR	Learning Rate	0.05	0.05	0.05	0.02	0.03	0.02	0.01	0.01	0.01
	Max Depth	3	3	3	3	3	3	3	3	3
	Max features	20	20	20	20	20	20	20	20	20
	Estimators	25	25	25	25	25	25	25	25	25
RFR	Estimator	20	20	20	25	25	25	15	15	15
	Max features	20	25	25	20	20	20	20	20	20
	Max Depth	6	6	4	6	6	6	4	6	4
	Min Split	3	3	3	3	3	3	3	3	3
	Min Leaf	7	6	7	8	7	7	7	7	8
SVR	Epsilon	0.6	0.6	0.2	0.4	0.6	0.4	0.4	0.3	0.6
	Kernel	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear
LR	Coefficient(β)	0.91	0.56	0.41	0.37	0.54	1.24	0.26	2.17	3.03
	Intercept(α)	11.41	13.33	23.41	22.22	43.02	12.34	17.40	13.89	18.73
DTR	Max Depth	6	6	6	6	6	6	6	6	6
	Min Split	9	7	9	7	9	7	9	7	9
	Min Leaf	5	5	5	5	5	5	5	5	5
ARIMA	(p,d,q)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)
	p-value	0.90	0.96	0.94	0.64	0.86	0.65	0.82	0.61	0.63
	Min AIC	5087.7	4714.1	2708.2	3133.9	2128.7	1428.5	2032.5	1249.2	1023.4
GA	Population Size	10	10	10	10	10	10	10	10	10
	Generation	45	45	45	45	45	45	45	45	45
	Gene Length	10	10	10	10	10	10	10	10	10
	Crossover	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
	Mutation	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

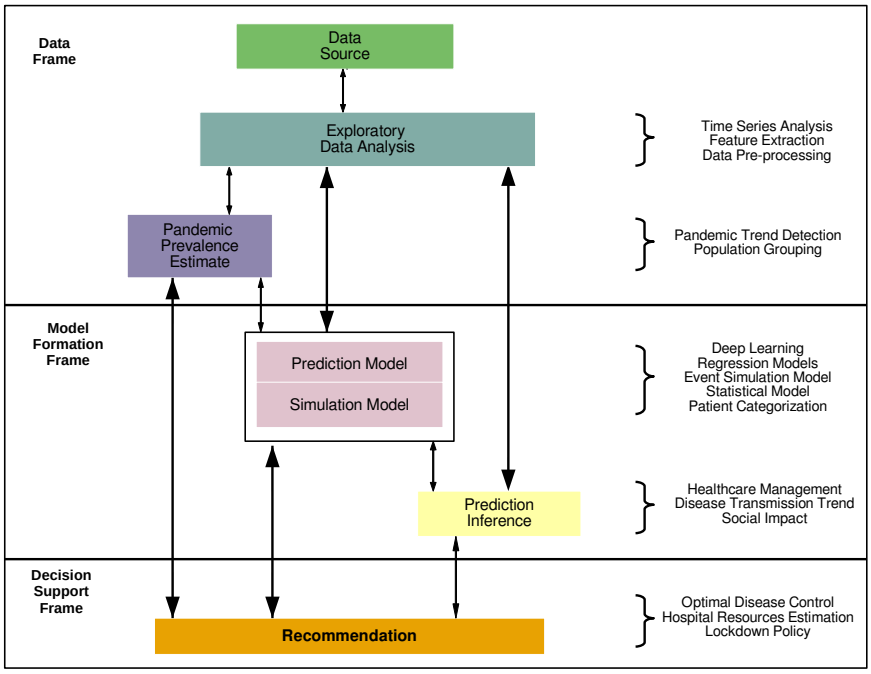


FIGURE 4.23: ADRIANA System Features

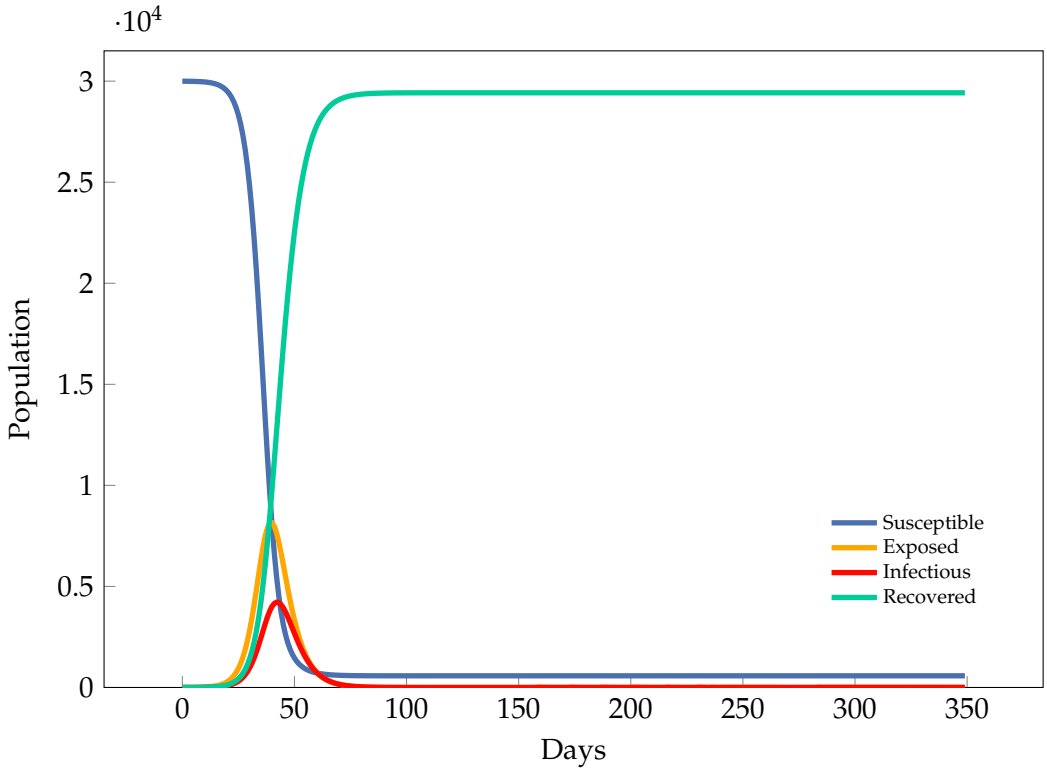


FIGURE 4.24: SEIR Simulation Model for COVID-19 in South Africa

with the aforementioned parameter configurations. The initial conditions of the compartmental model at time $t = 0$ are defined as $S(0) \geq 0, E(0) \geq 0, I(0) \geq 0$, and $R(0) \geq 0$. Appendix A presents the visual controls used to calibrate the SEIR

model. The model calibration process follows a standard representation that was applied to all the SIIMCF countries.

From Figure 4.24, we discovered that, when the population in the susceptible compartment (S) declines, the population in the exposed compartment (E) begins to decrease whenever S intercepts E at an equilibrium. Similarly, if the population in the susceptible compartment (S) declines, the population in the infectious compartments (I) begins to decrease whenever S intercepts I at equilibrium.

We presented the SEIR framework developed in this study as a modular design that could recreate the dynamics of epidemiology compartments in the SIIMCF countries.

4.6.3 Discrete Event Simulation Result (DES)

The DES simulates the COVID-19 event regarding the resources required to treat COVID-19 patients in the hospital. The DES obtained initial population from the output presented as the number of infectious individuals in the SEIR epidemiology. The DES simulates the entire hospital experience of the patient from admission through exit. The simulation time and parameter settings corresponds to the observations reported by COVID-19 monitoring platforms, such as the World Health Organization [8], National Institute for Communicable Diseases in South Africa [5], and South Africa Ministry of Health [7].

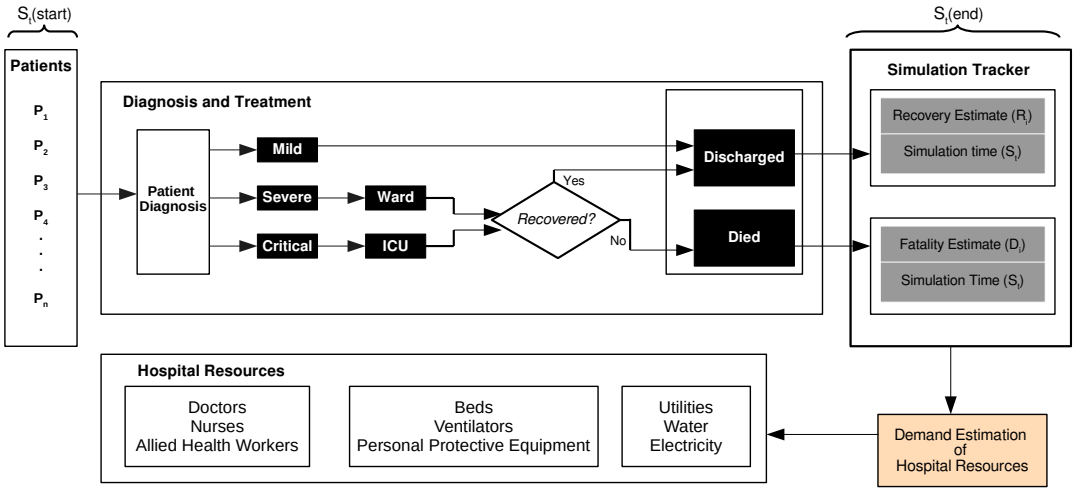


FIGURE 4.25: Discrete Event Simulation Procedure for Patient Hospitalization

The DES parameters are population (series of infectious population from the SEIR model), time of patient arrival and exit from hospital, and patients' risk factor

(such as age, immune threshold, infection history). The result of the DES is a data series estimating number of recoveries or fatalities that is expected from COVID-19 infectious population. DES categorizes patients as mild, severe, or critical. In addition, DES estimates the number of COVID-19 patients that will be admitted for either ICU or ward treatment, and predicts the peak demand time of hospital resources. DES simulation procedure explored in this study is illustrated by Figure 4.25.

4.6.4 Presentation of Surrogate Result

We presented the surrogate result in this section. As described in Section 4.1, we fitted the surrgate with the three types of datasets and compared the results. We observed that the surrogate was more accurate on the real-world datasets (S-REAL), followed by the simulated datasets (S-SIMUL), and hybrid datasets (S-SIREAL). The surrogate results for each of these datasets are presented in Figures 4.27, 4.28, and 4.29.

In addition, the CRD's surrogate forecast extracted for the S-REAL, S-SIMUL, and S-SIREAL for South Africa is presented the Figure 4.31. The evaluation metrics used to estimate the surrogate accuracy with the three categories of datasets is presented by Table 4.12 and Appendix G of this dissertation.

To describe the surrogate result, first, we plotted the actual CRD values versus the surrogate's predictions. Next, we show the future transmission trend of the CRD cases in South Africa and SIIMCF countries. The model evaluation results evaluated for the CRD cases in South Africa are presented in Table 4.12, while Appendix G shows the model evaluation results for the SIIMCF countries.

4.6.4.1 Actual Values Versus Surrogate Prediction

In Figure 4.26, we presented the graphical result of the surrogate predictions plotted against actual values for South Africa's CRD cases, while those of other SIIMCF countries are presented in Appendix C of this dissertation.

4.6.4.2 Schematics of Actual Values, Surrogate Prediction and Surrogate Forecast

Figures 4.27, 4.28, and 4.29 shows the graphical presentation of the actual values plotted against surrogate predictions and surrogate's forecast for South Africa CRD cases for S-REAL, S-SIMUL, and S-SIREAL. Similarly, the surrogate's forecast projection for SIIMCF countries are presented in Appendix D of this dissertation.

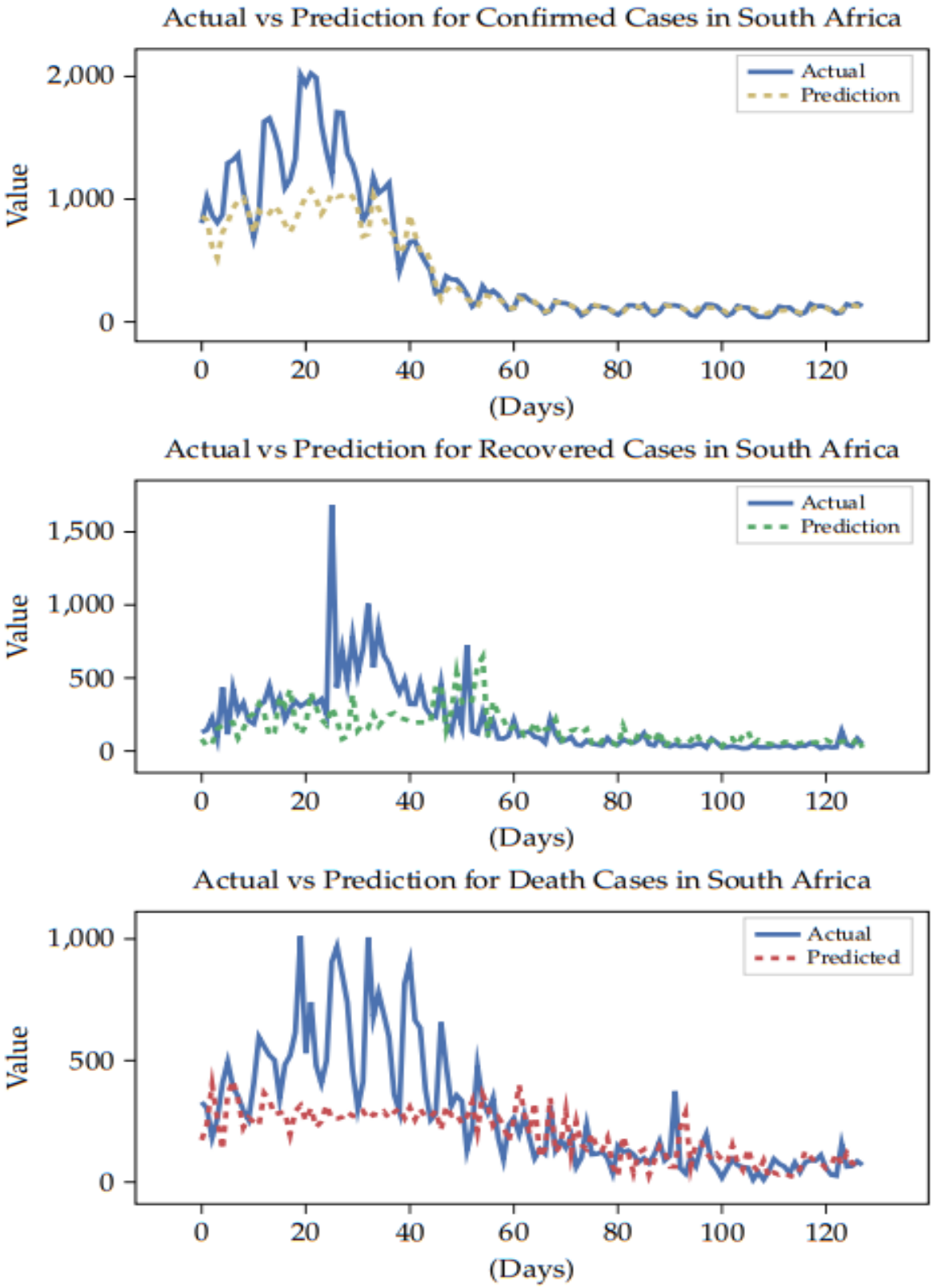


FIGURE 4.26: CRD Prediction Plot for South Africa

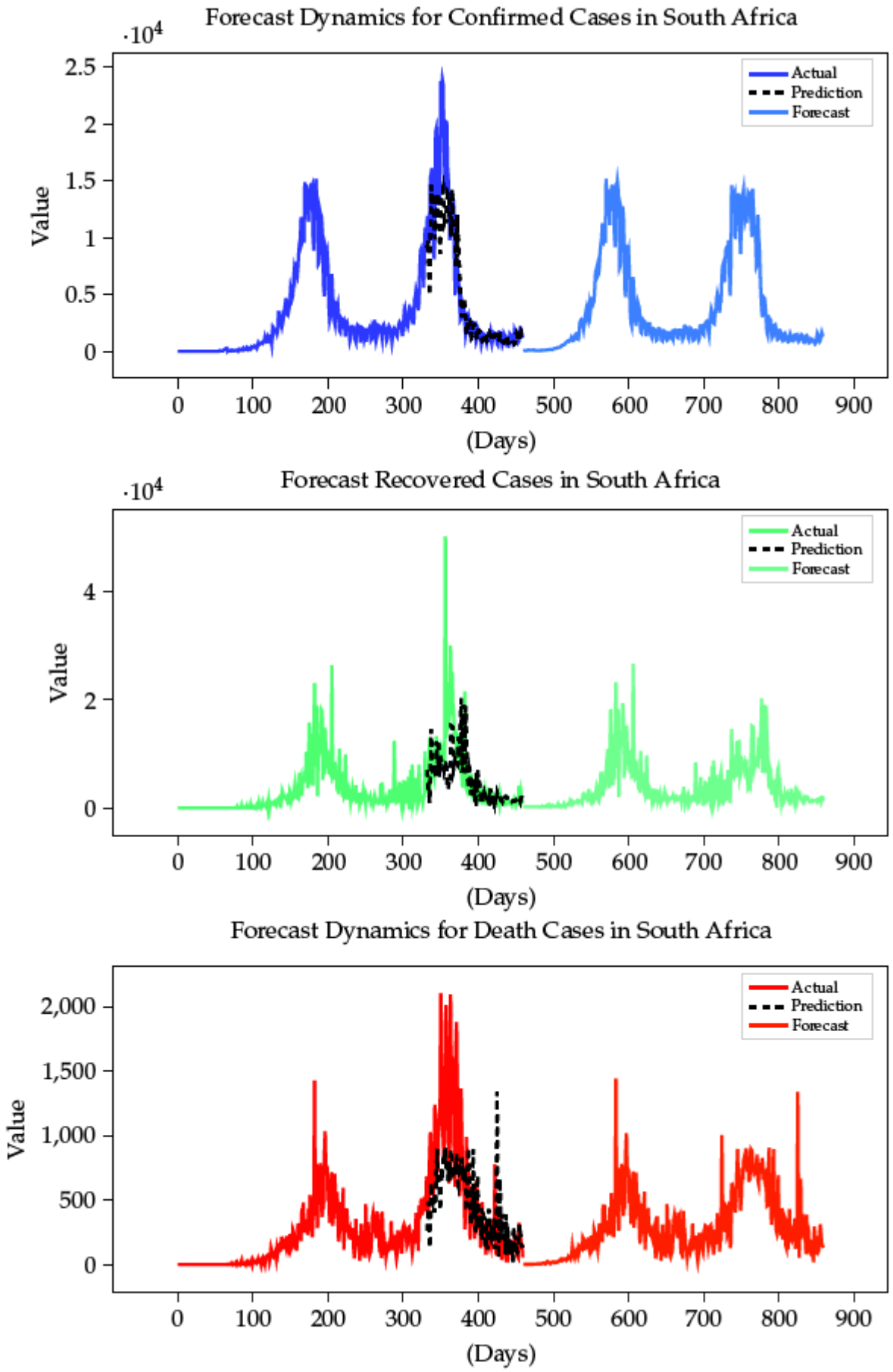


FIGURE 4.27: Prediction and Forecast Plot for South Africa

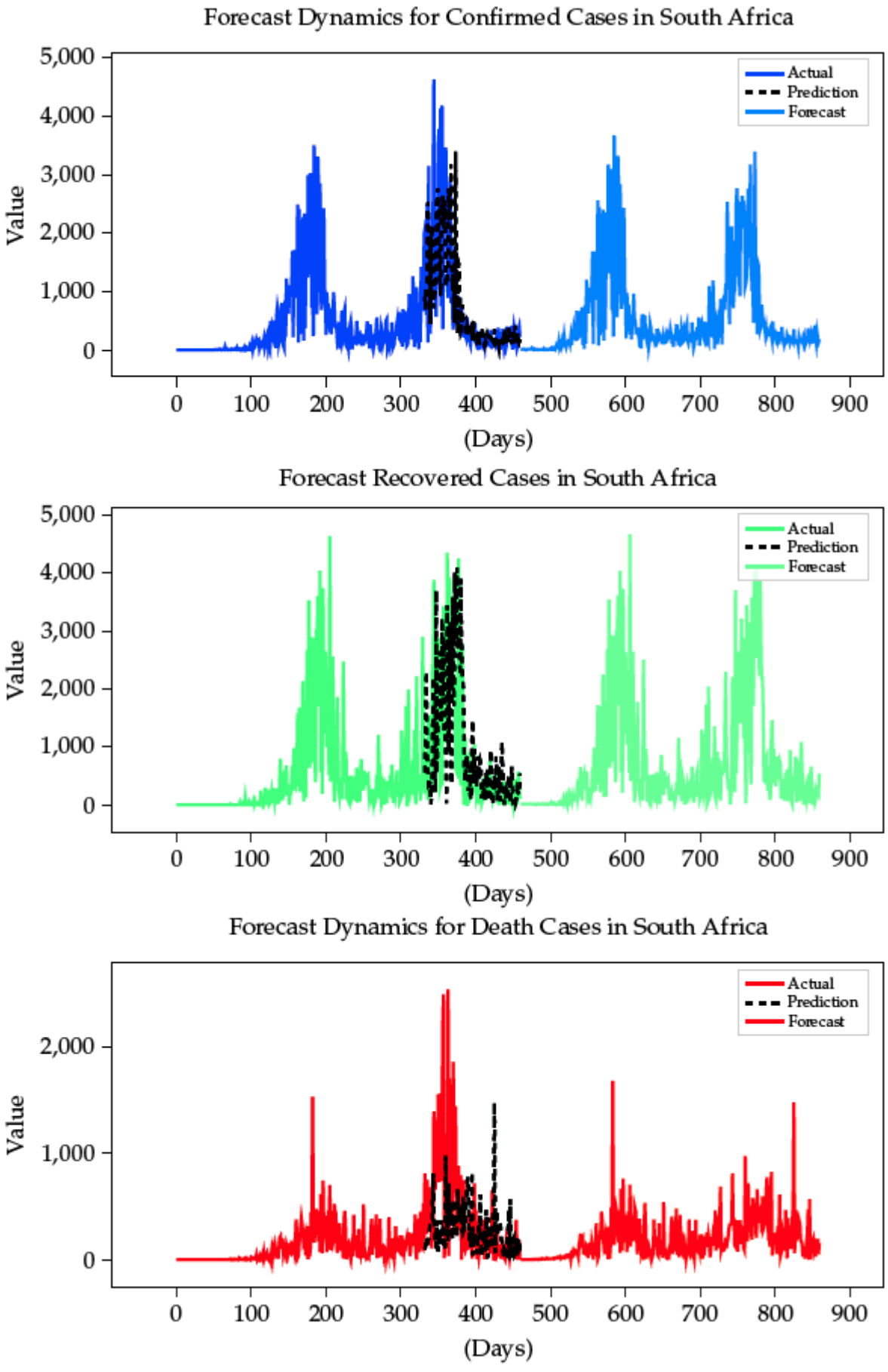


FIGURE 4.28: Prediction and Forecast Plot for South Africa (S-SIMUL)

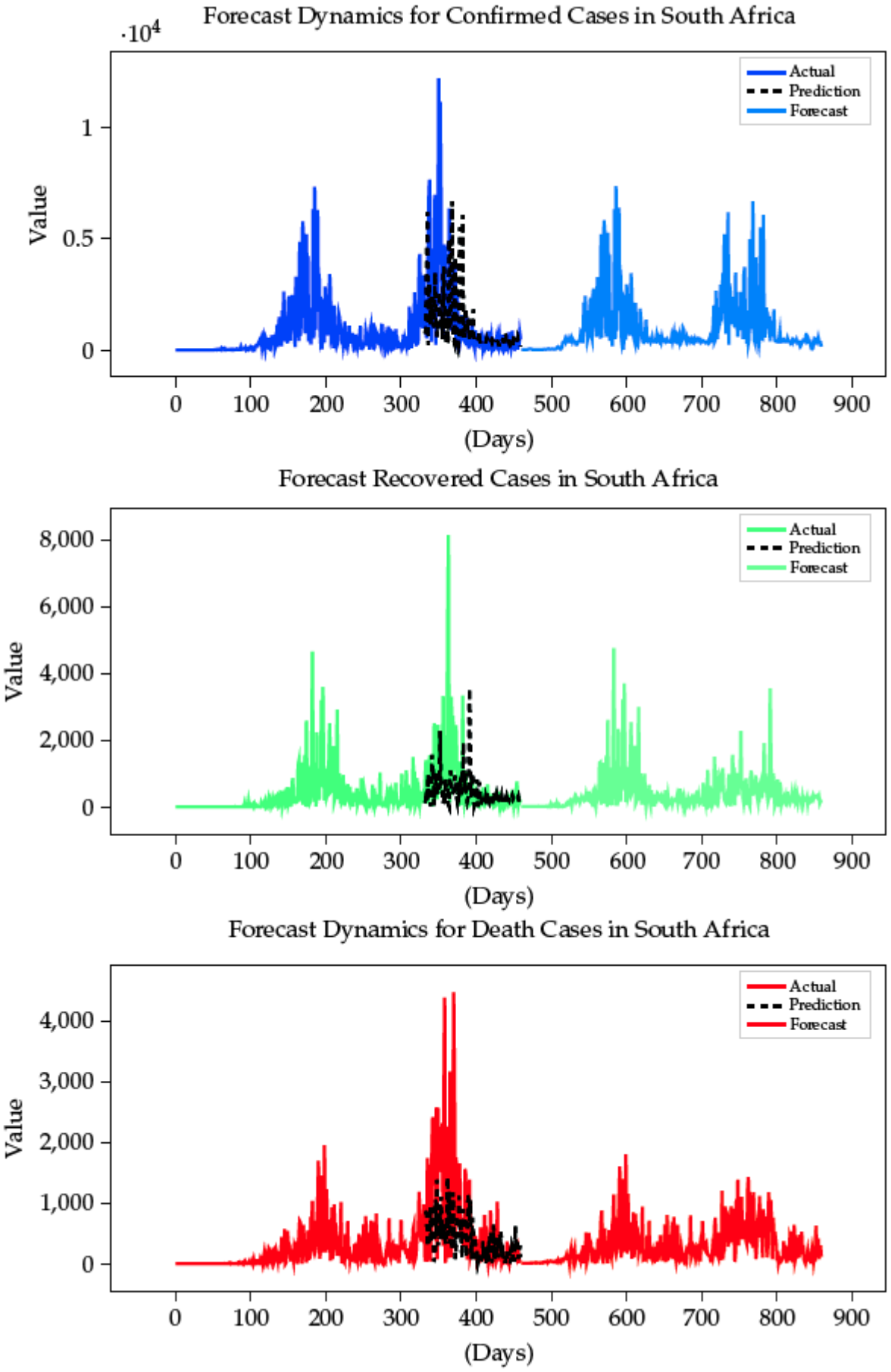


FIGURE 4.29: Prediction and Forecast Plot for South Africa (S-SIREAL)

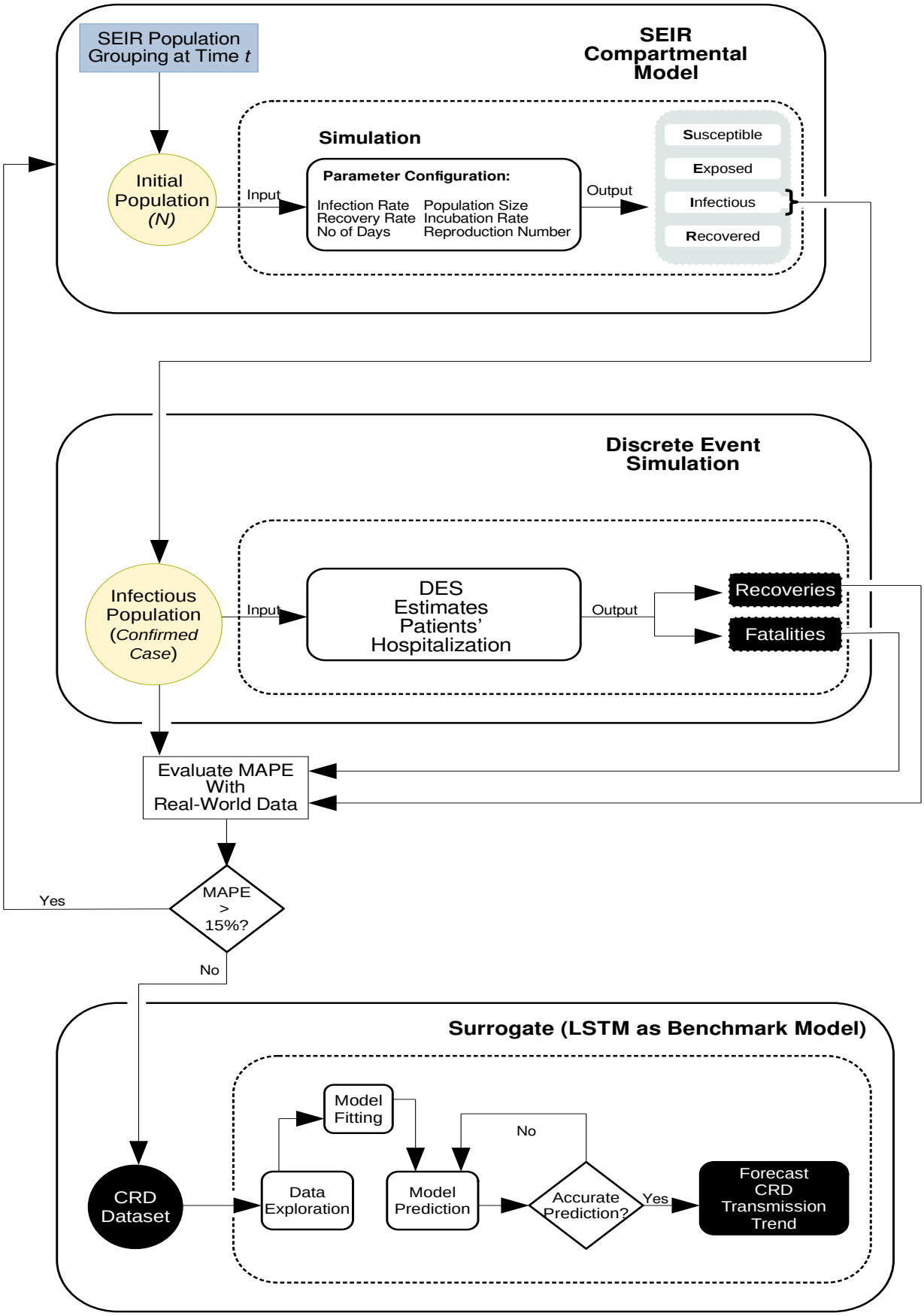


FIGURE 4.30: ADRIANA System Architecture

4.6.4.3 Combining the CRD Forecast for SIIMCF Countries

In Figure 4.31, we presented the CRD transmission trend forecasted by the surrogate for South Africa, while those of SIIMCF countries are presented in Appendix E of this dissertation.

4.6.4.4 Prediction Models

This section presents ADRIANA's CRD forecast for the SIIMCF countries. We show the result of stacked LSTM, the benchmark model of this study in Figure 4.32, while the results of other models are presented in Appendix F of this dissertation.

Interpretation of data representation in Table 4.10 and Table 4.11:

1. 150 ± 5 represents day observation within the range [145,155].
2. $150\pm5 - 200\pm6$ represents day observation from range [145,150] to [194,206].

Illustration: The confirmed cases (C) for South Africa for GRU model in Table 4.10 with values $100\pm5 - 150\pm6$ indicate GRU's forecast of an increase in confirmed cases between the day range [95,105] and [144,156]. The interpretation is consistent with other values in Tables 4.10, 4.11, H.1, H.2, H.3, and H.4.

Note: Fiji has blank prevalence day range because the country's CRD dataset is consistently between zero (0) and one (1). Therefore, surrogates cannot fit the data properly, as a linear trend of zero(0) or one (1) was observed in the predictions and forecasts as presented in Tables 4.10, 4.11, H.1, H.2, H.3, and H.4.

4.7 Model Evaluation Metrics

During model development, it is important that model quality is accurately measured. Therefore, it is vital to establish a good model evaluation strategy to measure model accuracy. This section outline the evaluation metrics for the ten (10) prediction models explored in this study. We used three model evaluation metrics, namely, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) as described in Section 3.8 of Chapter 3.

We presented the evaluation metrics for South Africa simulation experiment (with the three categories of datasets) in Figure 4.12. The model with the least RMSE,

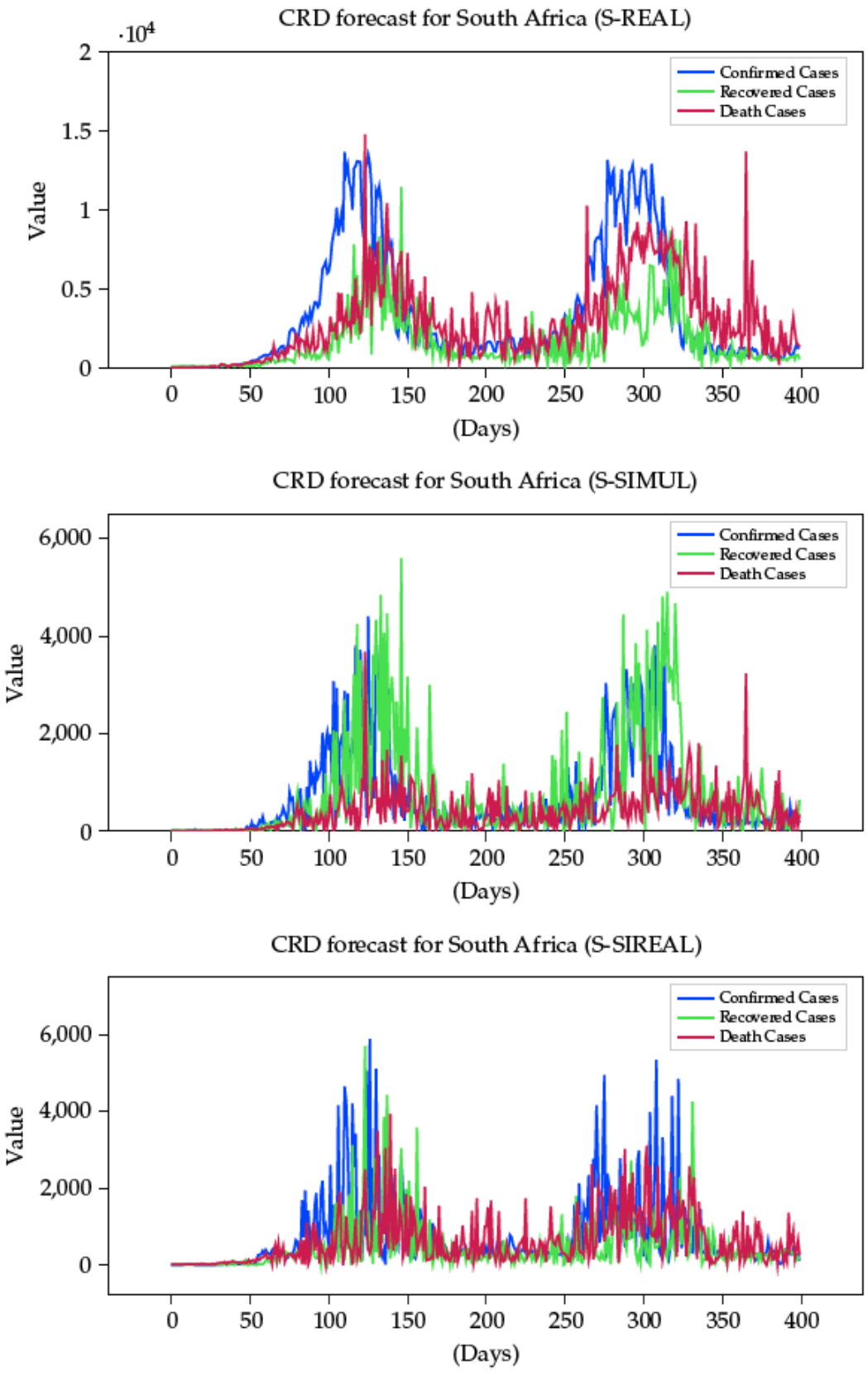


FIGURE 4.31: CRD Forecast for South Africa

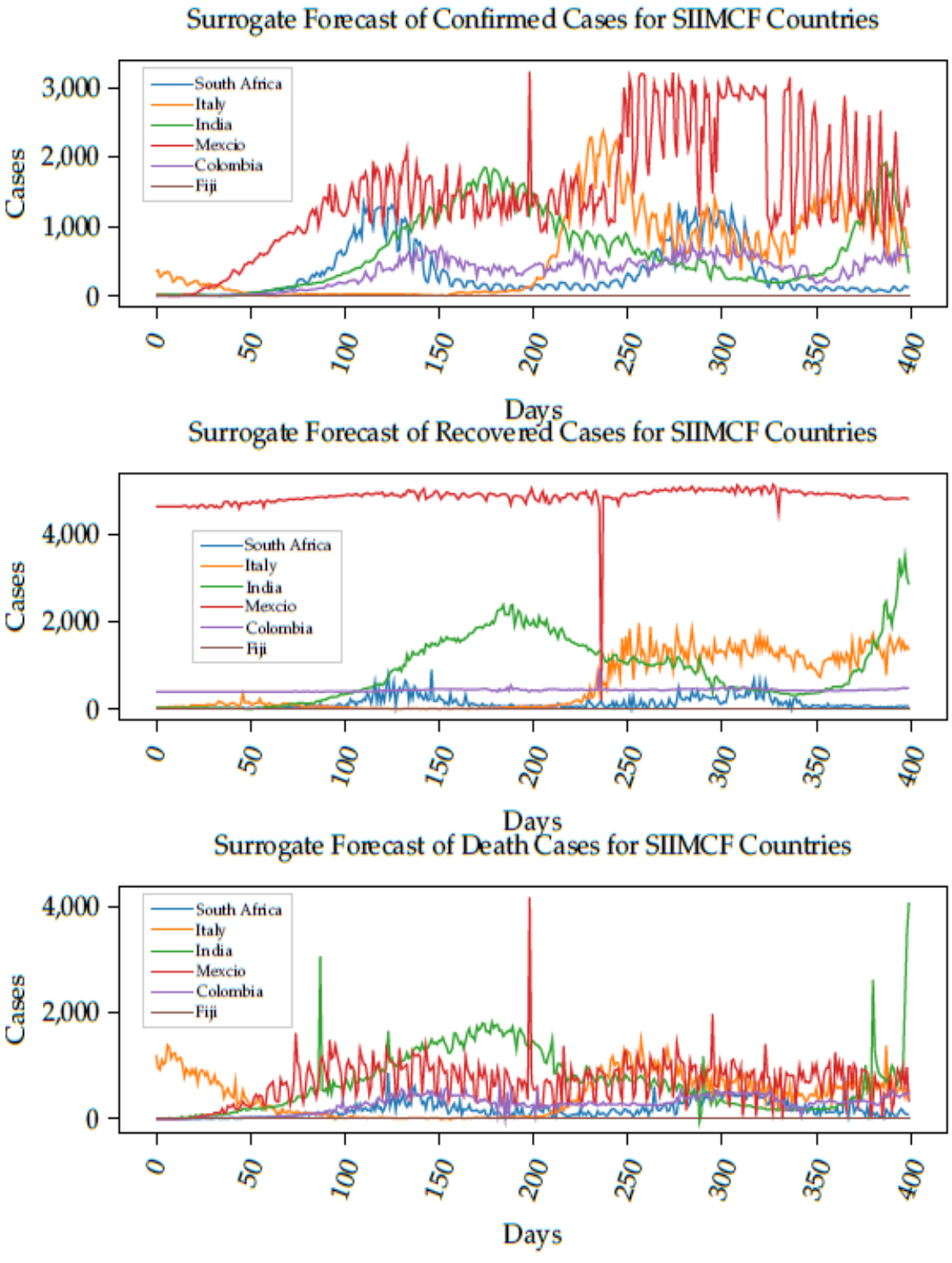


FIGURE 4.32: Stacked LSTM CRD Forecast for the SIIMCF Countries

TABLE 4.10: Surrogate Disease Prevalence Day Forecast

Country	**Case**	**LSTM**	**GRU**	**MLP**	**XGBR**	**SVR**
South Africa	C	70±5 – 150±8 270±5 – 320±3	100±5 – 150±6 270±6 – 320±5	95±5 – 150±9 270±8 – 330±7	90±10 – 150±6 260±5 – 320±2	90±10 – 150±6 230±10 – 330±5
	R	270±5 – 150±3 250±5 – 340±7	110±2 – 150±5 260±3 – 320±9	110±3 – 160±7 260±4 – 350±3	110±5 – 150±10 260±5 – 350±9	110±10 – 150±5 270±2 – 320±4
	D	120±2 – 150±5	120±5 – 150±5	120±5 – 160±5	120±3 – 150±10	110±6 – 150±2
Italy	C	220±8 350±3 – 390±4	220±5 – 380±5 375±4	210±4 360±4 – 390±3	220±6 365±5	210±8 350±2 – 390±4
	R	270±3 – 340±7 385±6	263±8 – 375±4	250±8 – 310±4 380±8	240±8 – 380±3	250±4 – 300±5 360±9
	D	42±5 – 315±6 392±5	43±6 – 390±5 395±4	44±5 – 286±10 375±8	45±5 245±6 – 320±8 375±10	35±8 219±8 – 320±2 379±2
India	C	120±5 – 235±4 355±6 – 390±5	125±3 – 249±5 350±6 – 390±6	110±5 – 230±8 353±2 – 390±4	200±5 – 245±9 340±5 – 370±3	120±5 – 230±7 360±6
	R	240±3 – 340±6 375±7	242±8 – 320±4 380±4	240±5 – 29±5 360±3 – 385±4	120±6 – 240±3	230±8 – 300±3 365±6
	D	10±3 – 210±5 360±5	20±9 – 210±3 375±7	12±6 – 230±7 370±4	15±5 – 250±8 360±7	12±4 – 260±7 375±2
Mexico	C	250±7 – 330±4 350±4 – 380±3	260±5 – 300±8 365±5	265±6 – 310±8	90±5 – 240±4 255±5 – 340±4	255±6 – 310±5
	R	145±4 – 320±7	150±5 – 340±2	155±4 – 350±7	140±8 – 345±9	130±6 – 360±5
	D	195±4	200±9 310±5	55±6 – 200±6 210±4 – 350±6	70±4 – 200±4 380±10	60±4 – 190±3 250±5 – 350±3
Colombia	C	110±4 – 160±6 270±6 – 310±7	13±8 – 160±4 220±4 – 320±9	125±6 – 190±5 290±2 – 310±4	120±4 – 160±5 250±6 – 320±4	130±7 – 160±4 260±5 – 295±6
	R	248±6 320±4	249±6 310±5	150±4 – 190±6 240±4 – 260±8	245±5	150±4 – 250±4
	D	120±6 – 160±3 260±5 – 340±5 375±6	125±4 – 180±6 265±8 – 330±2 380±3	120±5 – 190±6 290±6 – 340±4 388±5	90±4 – 180±6 200±2 – 340±2 375±10	120±5 – 140±8 300±5 – 340±4 378±7
Fiji	C	–	–	–	–	–
	R	–	–	–	–	–
	D	–	–	–	–	–

TABLE 4.11: Surrogate Disease Prevalence Day Forecast (continued...)

Country	**Case**	**RFR**	**DTR**	**LR**	**AR**	**ARIMA**
South Africa	C	90±5 – 150±6 230±7 – 320±3	100±5 – 150±3 260±4 – 330±9	90±7 – 150±4 260±4 – 320±8	40±20 100±5 – 200±4	150±10 200±5
	R	100±6 – 150±3 250±4 – 320±5	100±8 – 150±2 270±5 – 320±6	110±4 – 150±5 260±8 – 320±5	120±8 – 180±5 120±5 – 170±3	220±10 250±4
	D	110±4 – 150±9	120±5 – 160±5	120±6 – 150±5	50±10	175±5
Italy	C	215±10 375±8	205±6 370±6	210±5 – 240±6 370±5	180±4	150±6
	R	260±3 – 380±7	240±2 – 340±2 380±10	240±6 – 300±5 330±9	300±2 376±5	365±2
	D	25±5 – 36±4 375±3	15±5 380±5	12±6 – 26±4 383±4	13±2	36±5
India	C	120±5 – 247±4 355±5	125±6 – 245±3 375±6	130±5 – 255±8 370±6	120±6 120±3 – 350±4	350±4 150±10
	R	230±4 – 290±3 360±2 – 380±3	25±8 – 320±6 335±6 – 390±3	230±6 – 260±3 350±6	290±2 – 360±3 255±6 – 340±3	365±5
	D	130±3 – 230±6 360±2 – 385±4	125±3 – 220±9 360±6	140±5 – 200±6 355±6	145±20	150±10
Mexico	C	100±6 – 170±3 250±3 – 350±8	200±10 250±6 – 350±3	250±4 – 340±9	140±10	150±5
	R	100±6 – 270±2	120±3 – 360±6	135±4 – 370±6	120±5	180±10
	D	55±3 – 195±7 260±5 – 370±3	60±4 – 180±4 290±3 – 350±8	90±8 – 170±4 250±4 – 340±6	50±20	95±9
Colombia	C	120±6 – 160±7 250±6 – 320±4 380±5	130±4 – 155±8 255±2 – 320±8 376±8	120±6 – 190±6 260±3 – 290±6 370±8	135±8	150±6
	R	120±5 – 260±3 280±3 – 320±8	210±8 – 250±5	240±5 – 260±8	140±7	195±10
	D	120±4 – 155±5 260±3 – 330±2 375±6	120±6 – 160±3 300±2 – 340±3 389±5	125±3 – 145±4 285±3 – 340±2 385±4	146±6	270±3
Fiji	C	–	–	–	–	–
	R	–	–	–	–	–
	D	–	–	–	–	–

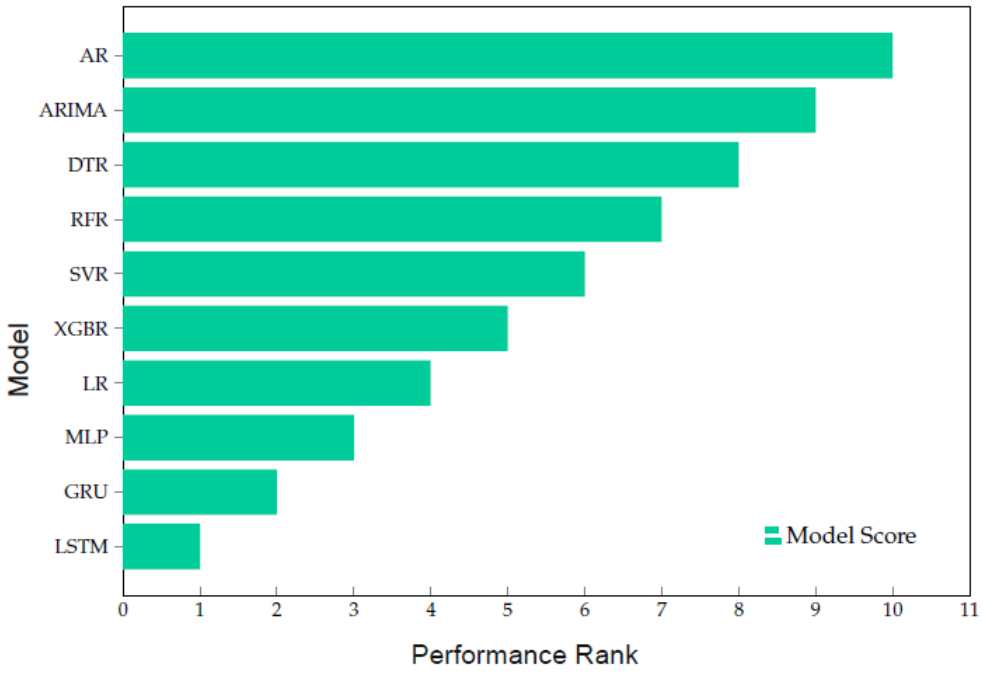


FIGURE 4.33: Model Performance Hierarchy

least MAE, and highest R^2 indicate the best performing model for each CRD case and are presented and in bold. The model evaluation for the SIIMCF countries are presented in Appendix G of this dissertation.

4.7.1 Model Performance Analysis

Overall, the models explored in this study performed well since they could reflect the goal of our study in forecasting COVID-19 transmission patterns in SIIMCF countries. As shown in Figure 4.33, we rated the performance of the models based on the evaluation metrics presented in Table 4.12 and Appendix G of this dissertation. The hierarchy of the model in Figure 4.33 is on a ranking position of 1 to 10 signifying an increasing ranking order. The model performance rank is consistent for the three categories of datasets explored in this study.

4.8 Conclusion

We discussed the core functionality and the results achieved from the sub-models that formulated the ADRIANA system. Also, as shown in Table 4.10 and Table 4.11, we presented the period when there would be an increase or decrease in CRD cases for each SIIMCF country as predicted by the models. Lastly, we presented the results (visually and theoretically) of our study both in this chapter and appendix section of this dissertation.

TABLE 4.12: Model Evaluation Metrics for South Africa

		S-REAL			S-SIMUL			S-SIREAL		
Model	**Cases**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**
LSTM	Confirmed	**31.33**	41.19	**0.96**	**38.01**	55.91	**0.89**	**42.76**	39.88	**0.76**
	Recovered	144.22	98.44	**0.89**	**151.36**	120.65	**0.75**	**169.02**	98.86	0.63
	Death	**138.82**	**87.06**	0.75	**41.34**	156.41	0.75	231.33	184.76	**0.69**
GRU	Confirmed	34.34	**27.74**	0.84	43.51	**28.79**	0.73	58.53	**29.94**	0.69
	Recovered	**143.10**	**94.92**	0.41	162.48	**99.26**	0.63	235.40	**97.65**	0.72
	Death	189.54	93.71	**0.85**	183.49	112.42	**0.78**	**195.59**	125.23	0.61
MLP	Confirmed	47.14	47.45	0.92	52.59	59.46	0.69	74.85	59.31	0.61
	Recovered	189.02	114.48	0.45	212.24	146.21	0.67	317.42	164.48	**0.83**
	Death	215.16	175.69	0.69	123.57	98.63	0.64	186.61	104.69	0.58
XGBR	Confirmed	677.53	431.95	0.71	621.34	691.13	0.70	872.13	748.85	0.58
	Recovered	8351.45	5382.05	0.60	8490.38	6852.55	0.45	7332.02	6392.51	0.39
	Death	272.81	196.66	0.64	312.49	201.47	0.62	305.86	289.21	0.52
RFR	Confirmed	2478.96	1397.71	0.91	2587.41	1577.90	0.72	3834.91	2489.17	0.66
	Recovered	5099.74	2731.33	0.80	6241.93	2892.39	0.75	6029.45	2939.43	0.34
	Death	193.22	253.66	0.69	99.35	**68.46**	0.62	134.52	**101.36**	0.52
SVR	Confirmed	142.76	248.83	0.92	169.64	282.45	0.74	195.60	384.36	0.38
	Recovered	354.62	258.98	0.53	365.82	271.52	0.31	388.26	408.81	0.29
	Death	117.65	175.38	0.64	126.68	94.82	0.62	271.42	89.58	0.34
LR	Confirmed	137.79	958.98	0.92	156.54	979.32	0.65	167.63	1058.08	0.82
	Recovered	5226.33	2529.19	0.42	6386.23	2893.28	0.40	6763.43	2939.19	0.33
	Death	120.57	176.39	0.64	122.71	78.49	0.62	215.39	145.44	0.44
DTR	Confirmed	570.13	1383.40	0.91	688.42	1541.77	0.84	760.31	1583.84	0.72
	Recovered	6153.67	3352.54	0.61	6351.72	4814.50	0.49	6452.72	3632.81	0.21
	Death	169.84	108.81	0.63	216.52	118.45	0.62	329.54	157.26	0.47
AR	Confirmed	6705.23	4737.39	0.38	6603.52	4898.93	0.34	7405.89	4396.35	0.28
	Recovered	6864.37	4052.41	0.41	6592.73	4914.61	0.30	7293.46	4826.43	0.26
	Death	230.28	153.73	0.24	249.84	166.33	0.22	282.24	198.63	0.19
ARIMA	Confirmed	4745.20	2969.01	0.82	4849.30	2930.68	0.45	4282.30	3101.01	0.23
	Recovered	7107.04	4154.26	0.29	6407.45	4293.32	0.19	7426.33	4784.62	0.19
	Death	255.06	169.06	0.37	286.68	1719.68	0.32	306.61	191.42	0.29

The next chapter provides the summary of this study, and a computational solution-guide to the COVID-19 event. The recommendation that can be adopted by hospital management for administrative control and the government regarding lockdown policy formulation.

Chapter 5

Conclusion

5.1 Introduction

This chapter concludes our study. Section 5.4 presents the general study summary, the contribution and relevance of this work to the research community is presented in Section 5.3. We made suggestions for future work in Section 5.5, and we highlighted the recommendations for government and hospital management in Section 5.5.1.

Our study explored 180 scenarios for each category of the datasets explored, 18 surrogate experiments, one (1) SEIR compartmental framework, and one (1) discrete event simulation (DES) model. The breakdown of the scenerios are ten (10) models, six (6) countries, and three (3) cases, (confirmed, recovered, and death cases). We had one (1) SEIR and one (1) DES because they were developed as a blueprint that can be adjusted to the COVID-19 event simulations of the SIIMCF countries. Our study showed how an optimized machine learning model outperforms statistical methods, as shown by model performance metrics in Chapter 4. Furthermore, our analysis was able to estimate the peak duration of hospital resources demand needed to treat infected individuals. Lastly, our findings can guide governments regarding lockdown policy by estimating the period of high transmission of COVID-19. Therefore, we have achieved the research goals and objectives.

5.2 General Inferences and Observations

In this subsection, we highlight key observations and interpretations deduced from our findings.

- Empirically, our results showed that fatality cases for SIIMCF countries are on the lower margin (compared to the confirmed and recovered cases) for both observed and forecast values. At the time of writing, the case fatality rate in South Africa was 2.95%, Italy was 2.87%, Colombia was 2.54%, and India was 1.34%, according to a report by WHO [1] and John Hopkins research [2], [85].

- The simulated and real-world dataset for Fiji, which we evaluated for Oceania continent, had a confined range of values for the daily CRD cases, ranging between zero (0) and one (1). The model simulation for Fiji resulted in unusual and flat modeling results for SEIR, DES, and surrogate experiment as presented in Tables 4.10, 4.11, H.1, H.2, H.3, and H.4.

- Our forecast indicated that South Africa, the benchmark country of this study, would experience an increase in the confirmed cases at different periods, as presented in Tables 4.10, 4.11, H.1, H.2, H.3, and H.4. However, the predicted recovery rate was higher than the mortality rate, implying that more individuals with confirmed COVID-19 cases would recover from COVID-19, which is consistent to the SIIMCF countries.

- For 400 days (starting from 24^{th}, *April* 2021), there is no single peak period for CRD cases in the SIIMCF countries. Tables 4.10, 4.11, H.1, H.2, H.3, and H.4 outlines the different CRD peak periods predicted by ADRIANA system for the SIIMCF countries.

- There was no correlation between the CRD cases of the SIIMCF countries, implying that an increasing or decreasing trend in a country's CRD case does not influence other SIIMCF countries. Each SIIMCF country's analysis of COVID-19 is unique and examined as a distinct entity.

- Investigating the future trend of COVID-19's CRD cases for the SIIMCF countries, we present the following;

  Confirmed Cases: According to Tables 4.10 and 4.11, and Figure 4.32, Mexico will experience the highest increase of confirmed cases, followed by India, Italy, South Africa, Colombia, and Fiji. Likewise, the confirmed cases for the SIIMCF countries showed similar transmission patterns, but in different magnitude. The implication of the similar transmission pattern is that preventative measures should be severe and properly adhered to by population. Similarly, hospitals in these countries should prepare to accommodate more patients during the predicted uptick period of COVID-19's confirmed cases. Furthermore, our analyses predicted that South Africa and Colombia will experience a minimal downward trend of confirmed cases between early February 2022 and March 2022.

  Recovered Cases: According to the analysis of recovered cases presented in Table 4.10 and Table 4.11, Mexico will experience the highest recovery rate, followed by India, South Africa, Italy, Colombia, and Fiji . Additionally, our findings reported that the recovery pattern in India would be on the increase by April 2022. In addition, the average ratio of confirmed cases to recovery

cases for the SIIMCF countries would be approximately 6:2, indicating that there will be more active cases than recovered cases.

Death Cases: According to the results presented in Tables 4.10, 4.11, H.1, H.2, H.3, and H.4, Mexico exhibit a wavy death pattern. Similarly, India will experience a spike in fatalities at four different periods from January, 2022. However, depending on the population's adherence to preventative measures, India's uptick in death cases can be minimized. Our findings predicted the mortality cases in the SIIMCF countries to decrease between November 2021 and March 2022.

We would like to report the following observations:

(i) Overall, there is a high probability that COVID-19 would be prevalent globally until April 2022. Our study findings evidenced this claim and also, COVID-19 publication by World Health Organization [8].

(ii) The predictions of the CRD cases is limited to the six SIIMCF countries considered in this study, and not all the countries in the world.

5.3 Contribution

Our research study offers three significant contributions to the field of research.

- First, we examined the CRD cases in the SIIMCF countries, spanning the global continent. Next, we evaluated the risk index (Section 4.4.3) and analyzed the reported CRD cases for the provinces in South Africa as the study's benchmark country. The risk index evaluation provided us with helpful insight about the disease's transmission dynamics in each province.

- Second, as stated in Section 4.6.2 of Chapter 4, we implemented a blueprint of an SEIR compartment model and a DES. The SEIR model is capable of simulating the number of exposed, infectious, and recovered (or death) cases in any population based on defined local parameters. The infectious population in the SEIR compartment was used as initial population for the discrete simulator. Likewise, DES simulates the treatment activities of infected COVID-19 patients in hospitals.

  Furthermore, our models were trained with the COVID-19's CRD dataset for each SIIMCF country. The daily 400-day disease transmission pattern was generated from 24^{th} April, 2021 to 22^{nd} May, 2022. Additionally, we were able to detect the trend in the SIIMCF's CRD forecasts as shown in Figure 4.32. The

CRD forecast presents a global perspective about the future transmission of COVID-19.

- Third, our study can ensure that government and hospitals are ready for the peak scenarios to provide optimal preventive measures and disease-risk control. The surrogate forecast can guide the government in its policy-making process regarding lockdown measures. In addition hospital management can use the SEIR result, event simulation, and surrogate forecast to guide preparation towards hospital resources demand required in the future to accommodate the treatment of infected individuals conveniently.

Our study observed that SIIMCF countries experienced different levels of social impact and considered lockdown as a major disease-spread preventive strategy. Therefore, our work is acknowledged and relevant in predicting the future dynamics of COVID-19 globally.

5.4 Conclusion

The COVID-19 pandemic caused significant threat to humanity, and its negative impact is inevitable [1], [4], [8]. some COVID-19's negative impact includes loss of lives, loss of jobs, and health challenges [5], [8]. One of the notable health challenges of COVID-19 is severe lung damage, resulting in death of several individuals globally. The global death toll of 3,582,724 was recorded as of 23^{rd} April, 2021 [2], [7]. Many healthcare and scientific modeling research studies have offered innovative and novel ideas as computational solution to the COVID-19 event. Different computational studies about COVID-19 pandemic evolved from Operation Research (OR) and Artificial Intelligence (AI). These approaches have significantly contributed to policy-regulation solutions, economic guide, and healthcare management centered around COVID-19 pandemic.

Recently, using deep learning framework for COVID-19's prediction has gained much attention. In the past, deep learning techniques were commonly recognized as only effective at dealing with non-linear computational tasks [86]. Our study's benchmark model utilizes deep learning, and is able to forecast both long-term and short-term COVID-19 trend. Additionally, we expanded the existing literature (highlighted in Chapter 2) by integrating three models (described in Chapter 1 and detailed in Chapter 3) into a single system named ADRIANA for prediction.

Using evolutionary approach to optimize surrogate models for prediction is advantageous for modeling task [45] by using Genetic Algorithm (GA). The Genetic Algorithm is an efficient tool for optimizing forecast models. GA operates by encoding data using chromosome representation [26], [47]. It is worth noting that, because

GAs are modular and can easily adapt to different model structure, GA optimization process is less complex. Therefore, GA method can be adjusted to fit different hypothetical modeling scenarios [26].

From theoretical perspective, hospital administrators should: (a) understand COVID-19 patients' trend, available treatment resources, and economic implications in hospital management, (b) be able to adjust resource allocation model parameters (for example, available beds in the wards or ICU, and patients' length of stay) of the discrete event simulator for further analyses, and (c) formulate implementation plan regarding observed changes in (b).

Similarly, policymakers should: (a) observe the adherence level of population to lockdown measures, (b) ensure that lockdown measures are not economically burdensome for citizens to adopt. Lockdown not properly managed can lead to different social vices, and (c) provide basic socio-economic support which are beneficial to citizens.

Our findings indicated that COVID-19 would be globally dominant by minimum of 230-days from 24^{th} April, 2021. Empirical study predicts that countries, such as South Africa and France, already anticipated the third wave of the disease [1], [4]. We captured this scenario in our simulation experiment and we presented the results and explanations in Tables 4.10, 4.11, H.1, H.2, H.3, and H.4 in Section 3.9.1 of Chapter 4.

The models investigated in this study used the COVID-19 transmission trend in past 400 days to forecast transmission trend for next 400 days. The learning function of COVID-19's past transmission trend to predict future trend was consistent for SIIMCF country's forecast experiment. Therefore, we are able to adequately predict disease peak duration. The predictions are beneficial to hospital management and government to inform policy, planning, and intervention action to control COVID-19 spread.

We strongly recommended that hospital administration and policymakers identify confirmed case trend versus recovered and death cases trend to guide policy formulation. The recommendation is given because the number of confirmed cases determines key intervention and policy planning structure for COVID-19 transmission. We observed the peak period of CRD cases for the SIIMCF countries as presented in Tables 4.10, 4.11, H.1, H.2, H.3, and H.4.

For example, the confirmed cases (C) for South Africa in Table 4.10 with value $70\pm5 - 150\pm8$ represents the stacked LSTM. The forecast estimated an upsurge in the confirmed cases between the day range [65,75] to [142,158]. Therefore, it implies, between days 65 and 158, (starting 24^{th} April 2021), hospital management should expect and prepare treatment for more infected patients. Similarly, the estimate recommends that government should raise the level of lockdown for these

day range. The computation and recommendation is applicable and consistent for SIIMCF countries.

5.5 Recommendation for Future Work

Our study proved the ability of machine learning to predict COVID-19's future transmission trend globally. We have identified four key areas of further study in this research.

Recommendation 1: Considering external factors, ADRIANA system can be programmed to dynamically adapt to different pandemic cases. In addition, evaluate and integrate the factors that influence pandemic transmission trend (such as the atmospheric conditions, geolocation, population response to government policy, and age bracket).

Recommendation 2: Novel GA models, such as hybrid neural networks-GAs, can be developed to suit specific geolocation. Hence, primary "What-if" model calibration could be expanded to include broader hypothetical scenarios.

Recommendation 3: Using a multi-objective optimization function to configure the surrogate hyperparameters. Especially by creating a CRD data-aggregation module to substitute each CRD modeling.

Recommendation 4: Our simulation did not consider the distribution mechanism of vaccines to the population, and distribution of personal protective equipment (PPE) to healthcare personnel. Further research can be explored to integrate these components into a single system, through a visual-control-based system.

5.5.1 COVID-19's Policy Recommendation

Several measures have been undertaken to reduce the increasing transmission rate of COVID-19 in South Africa and SIIMCF countries. Therefore, policymakers can use data-driven machine learning approaches to enhance their effort to reduce the high transmission rate of COVID-19. The recommended policy includes:

Discretion in Model Applications: Although, models explored in this research presented excellent forecasts, they should be applied with discretion for practical use, and should not be entirely relied on. Different factors can influence the accuracy of the model. The factors can be data alteration used for simulation, and different external factors that could modify COVID-19 CRD cases in geolocation.

Lockdown Regulations: It is vital to categorize government's policy on lockdown measures. The categorization will guide access to specific infrastructure and public settings. In general, total lockdown of population is discouraged, since it could lead to unexpected crises. Therefore, the lockdown policy should be flexible and firmly adhered by citizens.

Engage More Medical Professionals: Based on the future projection of the ADRIANA system, especially for confirmed cases, hospital management should engage more competent medical professionals to treat various COVID-19 cases. In addition, based on the number of confirmed cases for hospital admission, the ratio of medical personnel to the infected patient should be about 4:3 [15]. The ratio estimate will allow patients to access supervised medical treatment. By so doing, COVID-19's transmission can be reduced. Likewise, well-being of healthcare professionals must be considered with utmost importance by providing sufficient personal protective equipment (PPE) when administering treatments to patients.

Economic support to citizens: One of the major intervention strategies for minimizing COVID-19 transmission is the lockdown strategy. Government's inability to manage lockdown policy efficiently could lead to severe crises, as some population fragments may experience social difficulty. Individuals may fail to meet their fundamental needs, economic demands, and other responsibilities as a result of the aforementioned flaws, which may lead to anti-social behavior in the society.

A recommendation for government is to create a sufficient welfare management scheme that can help citizens meet basic needs during the lockdown. The welfare scheme can be: reduced cost of items, relief funds, housing allowances, free healthcare services, and loans. In addition, the welfare scheme will strengthen citizen's credibility, demonstrating government's concern for their well-being.

5.6 Disclosure of Competing Interest

We declare no commercial interests or personal connections that may appear to have motivated our study. Our primary goal is to provide computational solution to reduce the increasing trend of COVID-19 in the SIIMCF countries.

Appendix A

SEIR Simulation Graph

The Figure A.1 represents the design and controls for tuning the parameters of the SEIR model. The parameters architecture are consistent and can be applied to all the SIIMCF countries. The implementation details and results interpretation of the SEIR model are described in Chapter 4, Section 4.6.2 of this dissertation.

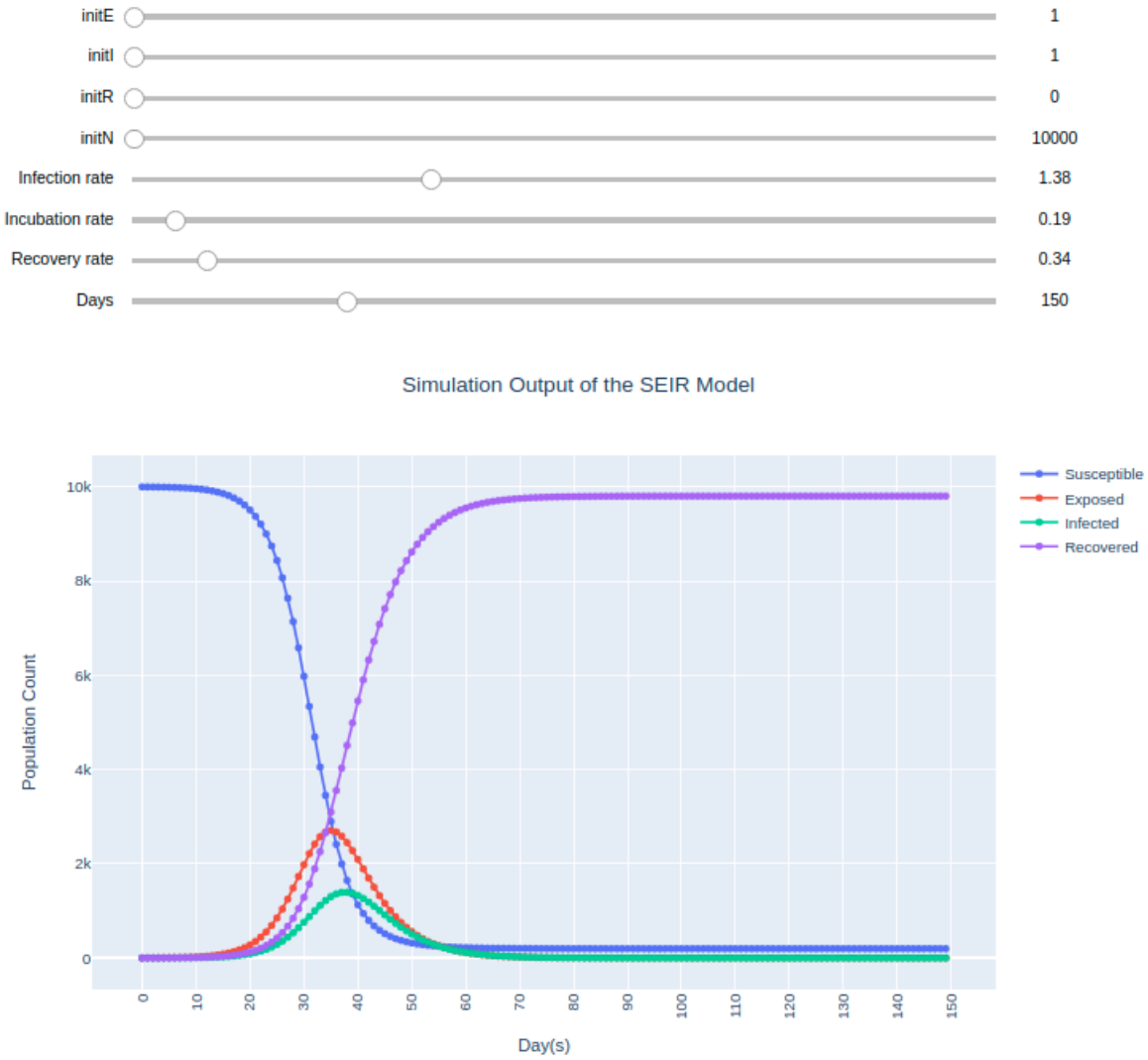


FIGURE A.1: SEIR Simulation Based on Configured Parameters

Appendix B

Surrogate Hyperparameter Selection

This section delves into the hyperparameter selection of optimal surrogates for each of the countries investigated in this study. The surrogate is modified using the selected hyperparameters to get appropriate results.

TABLE B.1: Optimal Model Hyperparameter Selection for Italy

		S-REAL			S-SIMUL			S-SIREAL		
Model	**Hyperparameter**	**C**	**R**	**D**	**C**	**R**	**D**	**C**	**R**	**D**
LSTM,	Neurons	62	56	52	48	42	53	61	64	68
GRU	Window Size	10	12	9	10	6	11	8	6	5
	Learning Rate	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
MLP	Batch Size	50	40	43	46	56	51	58	47	56
	Unit	16	12	18	11	18	10	14	16	9
	Learning Rate	0.01	0.04	0.03	0.01	0.02	0.02	0.01	0.06	0.01
	Solver	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam
XGBR	Learning Rate	0.005	0.005	0.005	0.01	0.03	0.02	0.004	0.04	0.03
	Max Depth	3	3	3	3	3	3	3	3	3
	Max features	15	15	15	15	15	15	15	15	15
	Estimators	20	20	20	20	20	20	20	20	20
RFR	Estimator	20	20	20	25	20	20	10	15	10
	Max features	15	25	25	15	25	15	15	15	20
	Max Depth	4	4	4	4	4	6	4	4	4
	Min Split	3	2	4	3	4	3	4	2	3
	Min Leaf	7	7	6	8	6	6	6	7	6
SVR	Epsilon	0.8	0.4	0.6	0.2	0.2	0.4	0.12	0.8	0.4
	Kernel	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear
LR	Coefficient(β)	0.63	0.57	0.81	1.25	0.34	3.01	0.33	2.14	1.48
	Intercept(α)	6.30	23.78	18.52	51.12	23.62	31.14	27.40	16.24	17.03
DTR	Max Depth	8	8	8	6	4	8	6	8	6
	Min Split	4	4	6	6	12	6	6	6	6
	Min Leaf	3	3	3	2	4	3	2	3	3
ARIMA	(p,d,q)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)
	p-value	0.92	0.56	0.63	0.91	0.72	0.57	0.93	0.64	0.72
	Min AIC	5087.7	4714.1	2708.2	2147.2	31498.5	1082.4	4030.1	1932.4	20127.2
GA	Population Size	10	10	10	10	10	10	10	10	10
	Generation	45	45	45	45	45	45	45	45	45
	Gene Length	10	10	10	10	10	10	10	10	10
	Crossover	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
	Mutation	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

TABLE B.2: Optimal Model Hyperparameter Selection for India

		S-REAL			S-SIMUL			S-SIREAL		
Model	**Hyperparameter**	**C**	**R**	**D**	**C**	**R**	**D**	**C**	**R**	**D**
LSTM,	Neurons	64	54	64	72	48	40	58	67	62
GRU	Window Size	8	12	9	5	7	6	5	6	7
	Learning Rate	0.01	0.01	0.01	0.03	0.03	0.03	0.05	0.05	0.05
MLP	Batch Size	48	52	44	51	53	57	68	46	53
	Unit	28	16	23	16	18	24	15	19	21
	Learning Rate	0.03	0.02	0.02	0.01	0.01	0.02	0.01	0.06	0.01
	Solver	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam
XGBR	Learning Rate	0.01	0.01	0.01	0.05	0.05	0.05	0.02	0.02	0.02
	Max Depth	3	3	3	3	3	3	3	3	3
	Max features	10	10	10	10	10	10	10	10	10
	Estimators	15	15	15	15	15	15	15	15	15
RFR	Estimator	10	10	10	10	10	10	10	10	10
	Max features	15	15	15	12	12	12	13	13	13
	Max Depth	4	4	4	4	4	4	4	4	4
	Min Split	3	3	3	3	3	3	3	3	3
	Min Leaf	8	6	8	8	6	8	8	6	8
SVR	Epsilon	0.2	0.4	0.2	0.6	0.6	0.6	0.3	0.3	0.3
	Kernel	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear
LR	Coefficient(β)	1.21	0.32	9.27	2.67	0.41	2.27	10.76	4.28	5.43
	Intercept(α)	19.07	19.98	33.53	21.20	23.08	32.52	12.52	17.12	18.75
DTR	Max Depth	4	4	4	4	4	4	4	4	4
	Min Split	7	7	7	7	7	7	7	7	7
	Min Leaf	3	3	3	3	3	3	3	3	3
ARIMA	(p,d,q)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)
	p-value	0.94	0.91	0.92	0.64	0.56	0.91	0.84	0.67	0.94
	Min AIC	2057.4	6744.5	1767.8	1037.3	2827.2	2376.4	2051.5	5232.5	1008.5
GA	Population Size	10	10	10	10	10	10	10	10	10
	Generation	45	45	45	45	45	45	45	45	45
	Gene Length	10	10	10	10	10	10	10	10	10
	Crossover	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
	Mutation	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

TABLE B.3: Optimal Model Hyperparameter Selection for Mexico

		S-REAL			S-SIMUL			S-SIREAL		
Model	**Hyperparameter**	**C**	**R**	**D**	**C**	**R**	**D**	**C**	**R**	**D**
LSTM,	Neurons	72	52	67	36	44	89	86	67	65
GRU	Window Size	12	14	8	11	9	12	10	14	8
	Learning Rate	0.01	0.01	0.01	0.08	0.08	0.08	0.04	0.04	0.04
MLP	Batch Size	62	51	42	44	61	49	53	52	46
	Unit	28	21	18	22	14	18	12	16	19
	Learning Rate	0.05	0.02	0.01	0.01	0.01	0.03	0.01	0.02	0.02
	Solver	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam
XGBR	Learning Rate	0.012	0.05	0.010	0.02	0.02	0.03	0.04	0.03	0.02
	Max Depth	3	3	3	3	3	3	3	3	3
	Max features	25	25	25	25	25	25	25	25	25
	Estimators	20	20	20	20	20	20	20	20	20
RFR	Estimator	30	30	30	25	20	25	25	15	25
	Max features	25	25	25	15	10	15	10	25	20
	Max Depth	4	3	4	4	4	6	4	4	4
	Min Split	5	2	5	3	4	6	4	8	2
	Min Leaf	3	3	3	6	2	2	8	8	8
SVR	Epsilon	0.2	0.7	0.3	0.1	0.7	0.5	0.16	0.18	0.2
	Kernel	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear
LR	Coefficient(β)	2.42	0.32	6.34	0.45	0.27	0.62	0.36	3.54	2.06
	Intercept(α)	1.07	15.34	12.02	18.14	33.42	23.19	24.72	18.49	15.22
DTR	Max Depth	6	6	6	4	4	6	4	6	6
	Min Split	5	5	5	9	2	4	6	12	4
	Min Leaf	3	2	3	3	5	2	4	4	2
ARIMA	(p,d,q)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)
	p-value	0.90	0.96	0.94	0.96	0.62	0.33	0.83	0.62	0.74
	Min AIC	7033.7	6581.6	7408.8	105.7	3981.5	1233.5	2353.1	1233.8	2012.5
GA	Population Size	10	10	10	10	10	10	10	10	10
	Generation	45	45	45	45	45	45	45	45	45
	Gene Length	10	10	10	10	10	10	10	10	10
	Crossover	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
	Mutation	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

TABLE B.4: Optimal Model Hyperparameter Selection for Colombia

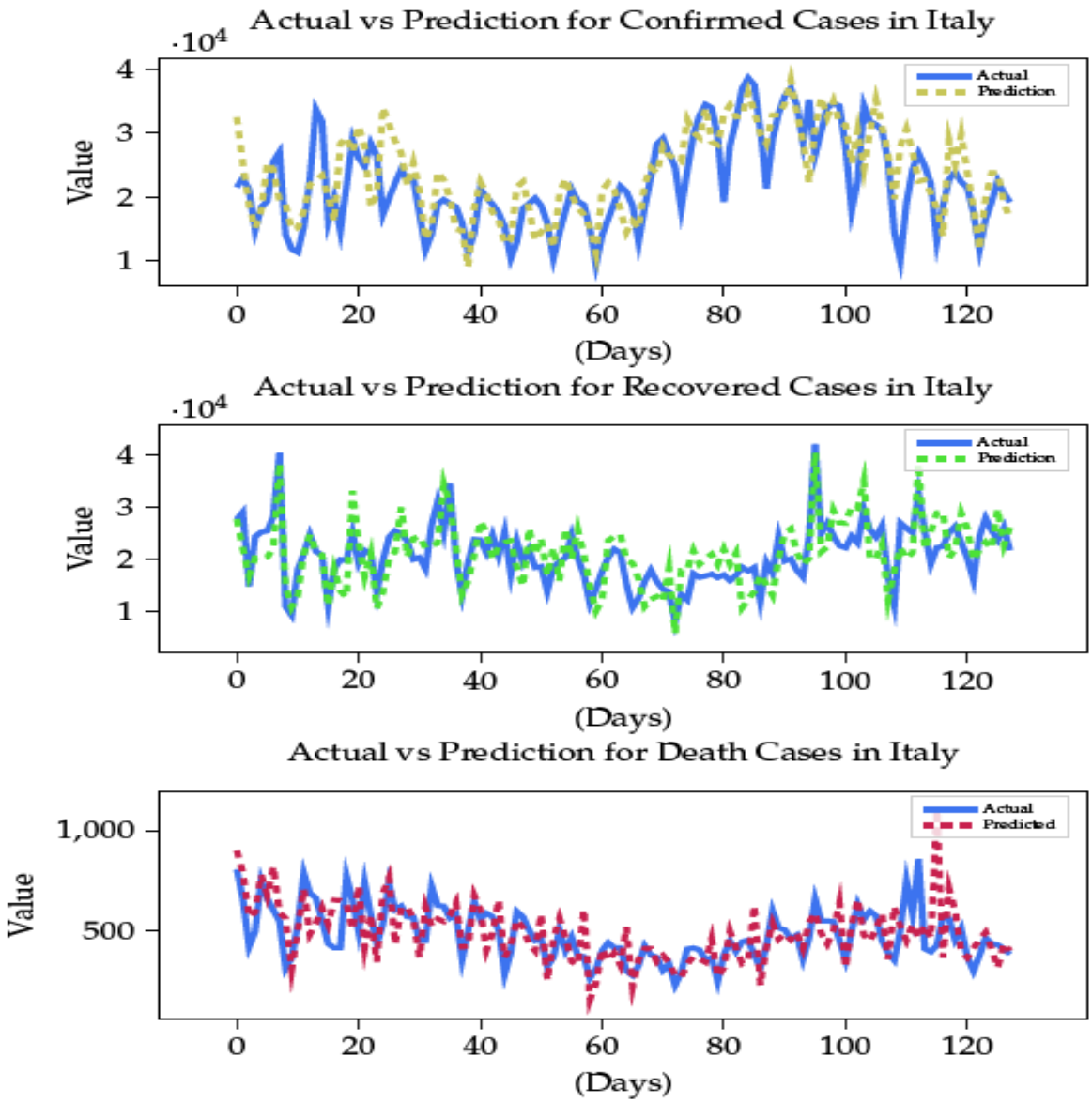
		S-REAL			S-SIMUL			S-SIREAL		
Model	**Hyperparameter**	**C**	**R**	**D**	**C**	**R**	**D**	**C**	**R**	**D**
LSTM,	Neurons	74	78	64	61	63	58	54	44	61
GRU	Window Size	16	6	12	6	8	5	16	14	14
	Learning Rate	0.03	0.03	0.03	0.02	0.03	0.01	0.03	0.07	0.07
MLP	Batch Size	52	46	51	62	68	63	36	52	56
	Unit	16	12	32	18	32	19	17	18	21
	Learning Rate	0.03	0.02	0.01	0.01	0.01	0.02	0.01	0.06	0.01
	Solver	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam
XGBR	Learning Rate	0.02	0.015	0.014	0.04	0.02	0.03	0.02	0.02	0.16
	Max Depth	3	3	3	3	3	3	3	3	3
	Max features	30	30	30	25	35	15	25	15	35
	Estimators	20	35	30	25	15	35	25	30	15
RFR	Estimator	15	15	15	20	20	15	15	25	20
	Max features	10	15	10	15	25	25	10	15	20
	Max Depth	6	6	8	7	5	4	3	6	8
	Min Split	4	2	5	5	4	6	4	8	2
	Min Leaf	7	5	5	6	4	4	4	3	3
SVR	Epsilon	0.2	0.6	0.4	0.4	0.3	0.4	0.8	0.2	0.14
	Kernel	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear
LR	Coefficient(β)	3.31	0.44	0.57	3.26	0.54	1.41	2.23	4.02	3.91
	Intercept(α)	12.21	19.30	6.72	8.12	3.24	15.52	26.83	11.19	13.63
DTR	Max Depth	8	8	8	4	6	8	8	8	3
	Min Split	3	3	3	6	9	2	4	7	4
	Min Leaf	3	3	5	5	3	3	3	3	3
ARIMA	(p,d,q)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)
	p-value	0.14	0.19	0.13	0.42	0.31	0.34	0.15	0.42	0.33
	Min AIC	7033.7	6581.6	7408.8	3105.1	4286.2	2821.4	8659.2	1322.4	60409.4
GA	Population Size	10	10	10	10	10	10	10	10	10
	Generation	45	45	45	45	45	45	45	45	45
	Gene Length	10	10	10	10	10	10	10	10	10
	Crossover	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
	Mutation	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

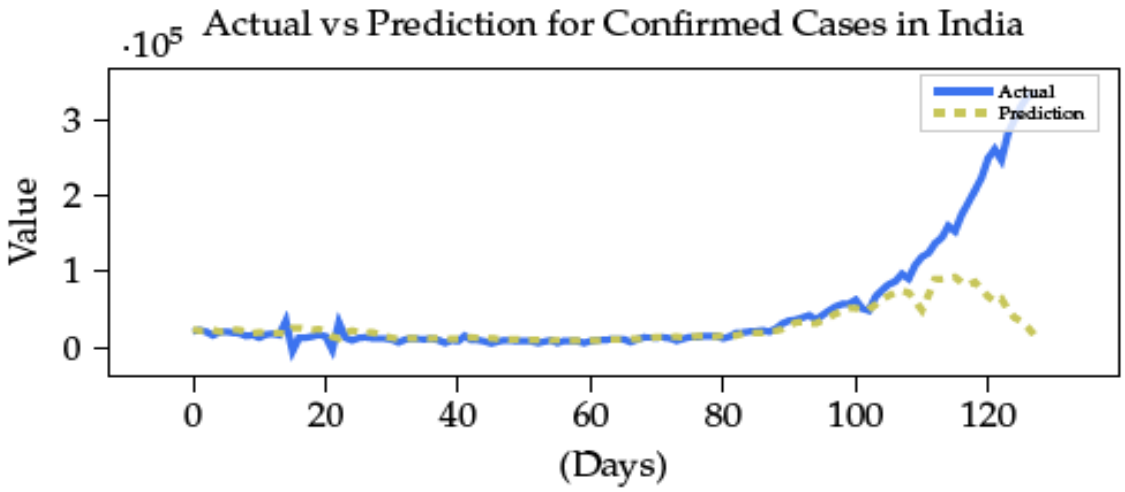
TABLE B.5: Optimal Model Hyperparameter Selection for Fiji

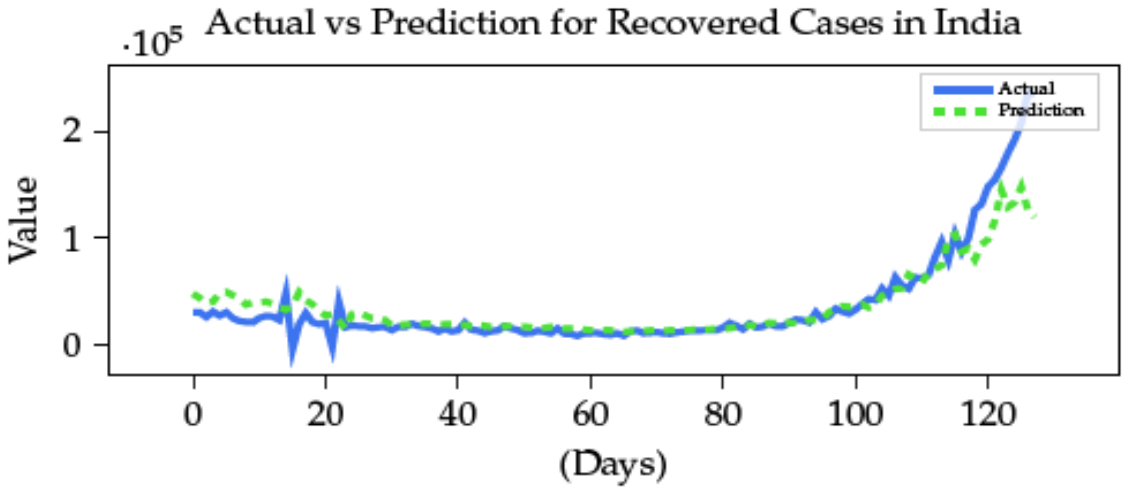
		S-REAL			S-SIMUL			S-SIREAL		
Model	**Hyperparameter**	**C**	**R**	**D**	**C**	**R**	**D**	**C**	**R**	**D**
LSTM,	Neurons	46	32	45	52	43	58	54	44	61
GRU	Window Size	5	6	4	10	8	5	11	9	4
	Learning Rate	0.03	0.03	0.03	0.02	0.03	0.01	0.03	0.07	0.07
MLP	Batch Size	42	36	43	52	58	34	31	52	50
	Unit	12	16	11	18	34	12	15	21	34
	Learning Rate	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.04	0.01
	Solver	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam
XGBR	Learning Rate	0.012	0.010	0.002	0.14	0.20	0.05	0.04	0.007	0.12
	Max Depth	3	3	3	3	3	3	3	3	3
	Max features	10	10	15	25	10	25	35	10	20
	Estimators	25	15	20	15	20	25	15	30	25
RFR	Estimator	30	15	30	25	35	10	10	15	25
	Max features	10	5	5	10	15	15	16	15	25
	Max Depth	4	6	8	7	3	4	5	5	8
	Min Split	8	5	5	4	7	6	2	4	3
	Min Leaf	5	4	8	6	3	4	7	3	5
SVR	Epsilon	0.2	0.8	0.16	0.4	0.3	0.2	0.8	0.4	0.12
	Kernel	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear
LR	Coefficient(β)	0.01	0.01	0.02	0.2	0.31	0.11	0.32	0.01	0.13
	Intercept(α)	1.31	1.20	1.27	1.21	1.42	1.25	1.38	1.91	1.36
DTR	Max Depth	6	2	2	4	6	2	6	2	5
	Min Split	4	5	3	6	2	7	5	4	5
	Min Leaf	4	7	6	5	4	2	8	4	5
ARIMA	(p,d,q)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)	(1,1,0)
	p-value	0.11	0.09	0.12	0.24	0.13	0.43	0.12	0.22	0.18
	Min AIC	34.7	83.6	25.5	55.9	32.65	21.5	19.4	17.3	12.32
GA	Population Size	10	10	10	10	10	10	10	10	10
	Generation	45	45	45	45	45	45	45	45	45
	Gene Length	10	10	10	10	10	10	10	10	10
	Crossover	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
	Mutation	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

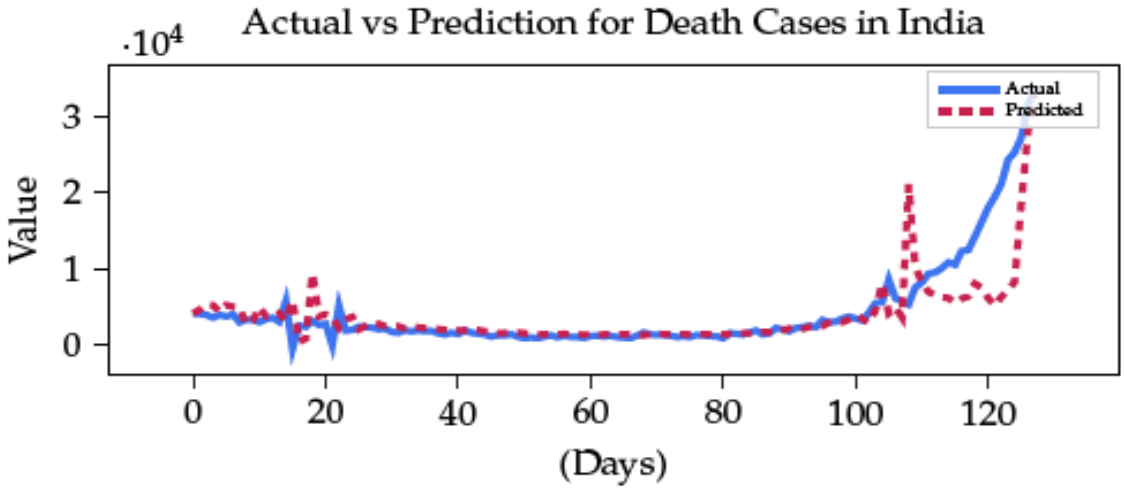
Appendix C

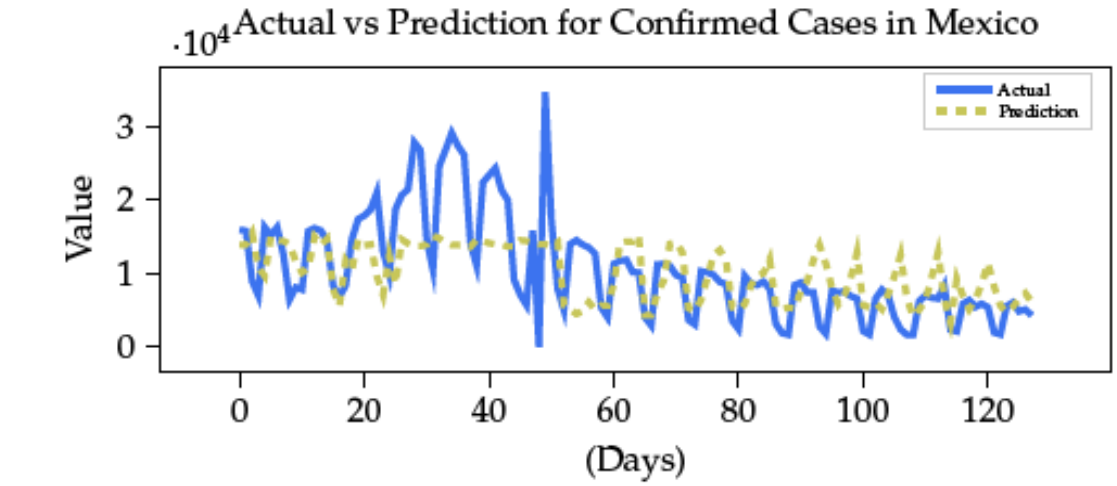
Schematics of Actual Values Versus Model Predictions

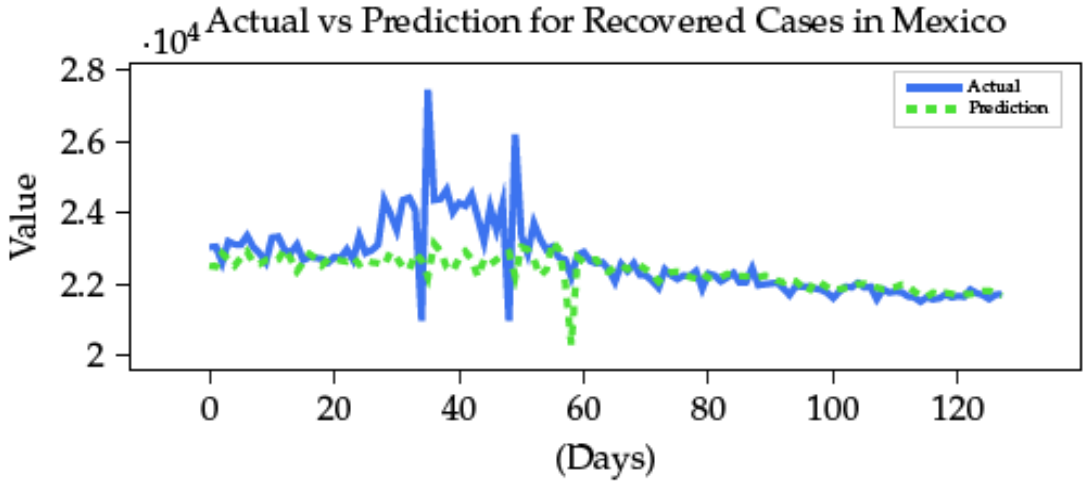


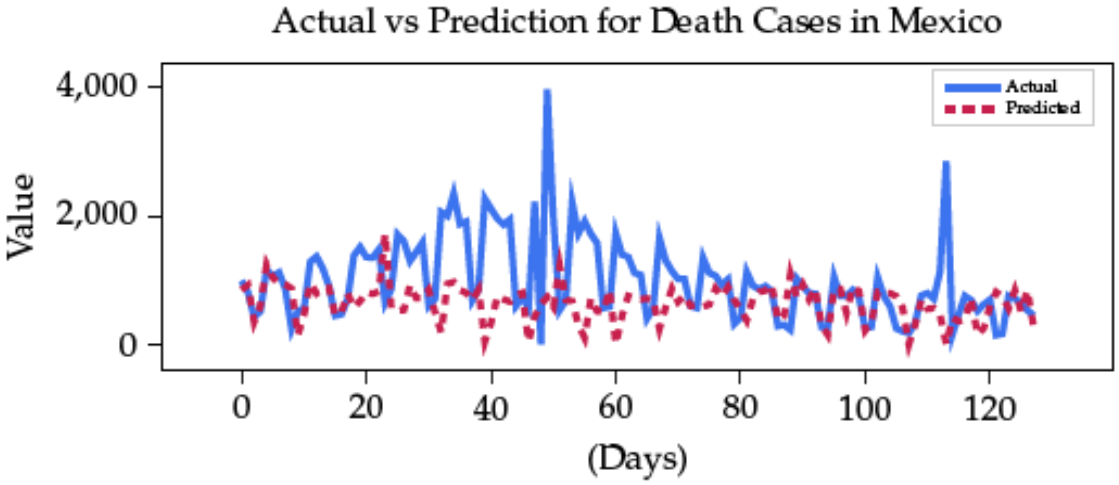

Actual vs Prediction for Confirmed Cases in India
·10^5
Value
(Days)
Actual
Prediction

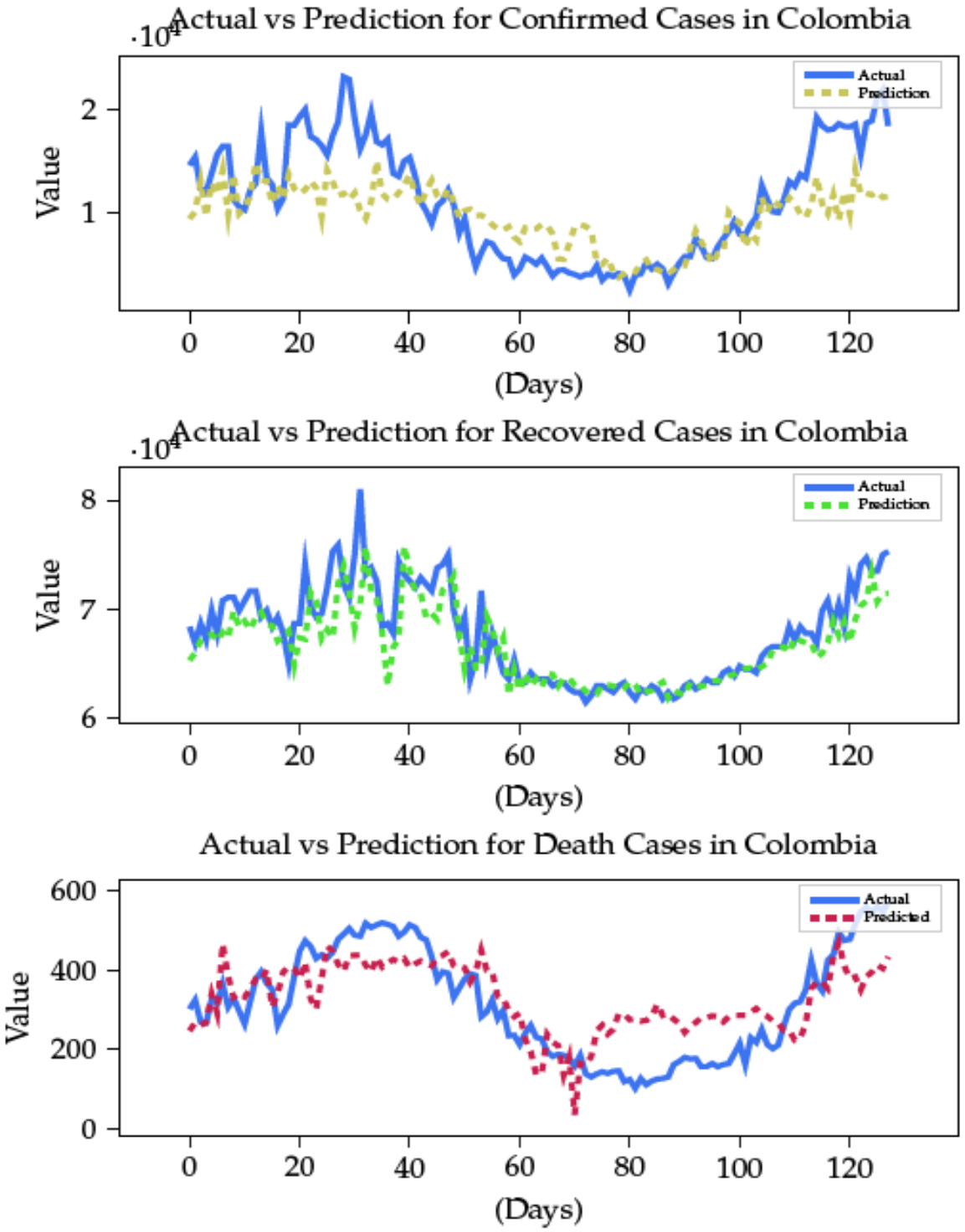

Actual vs Prediction for Recovered Cases in India
·10^5
Value
(Days)
Actual
Prediction

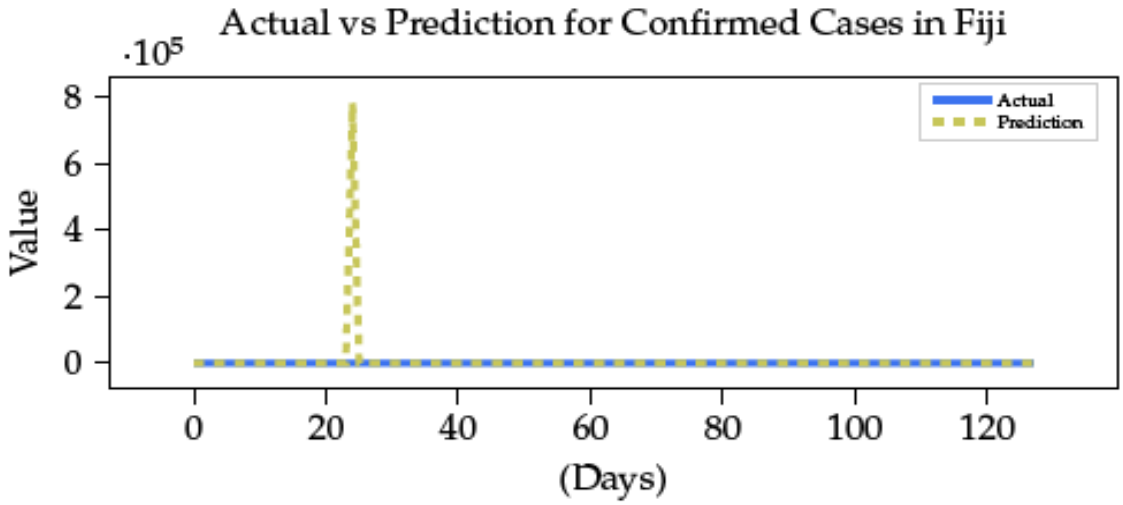

Actual vs Prediction for Death Cases in India
·10^4
Value
(Days)
Actual
Predicted

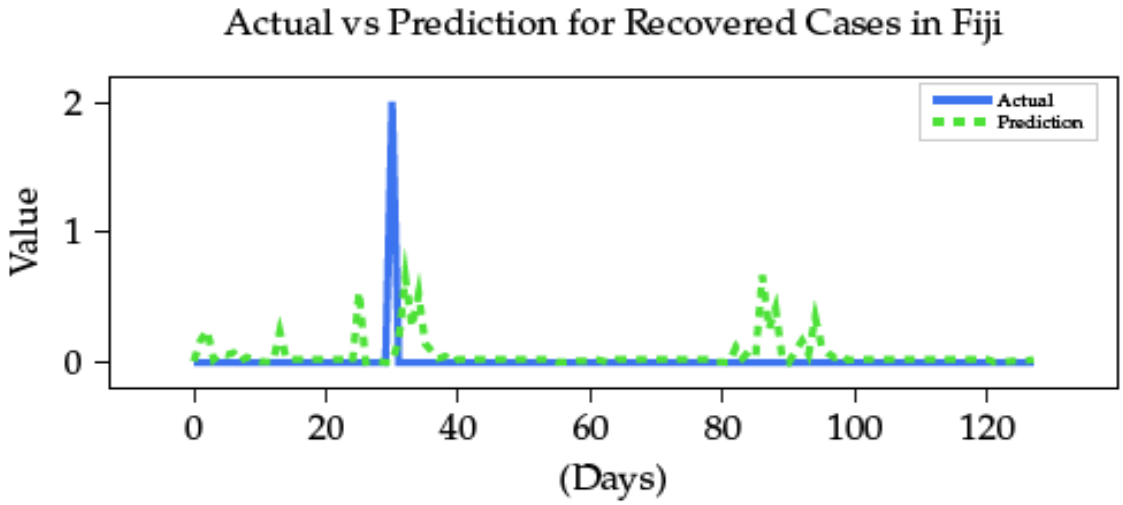

Actual vs Prediction for Confirmed Cases in Mexico
$\cdot 10^4$
Value
3
2
1
0
0
20
40
60
80
100
120
(Days)
Actual
Prediction

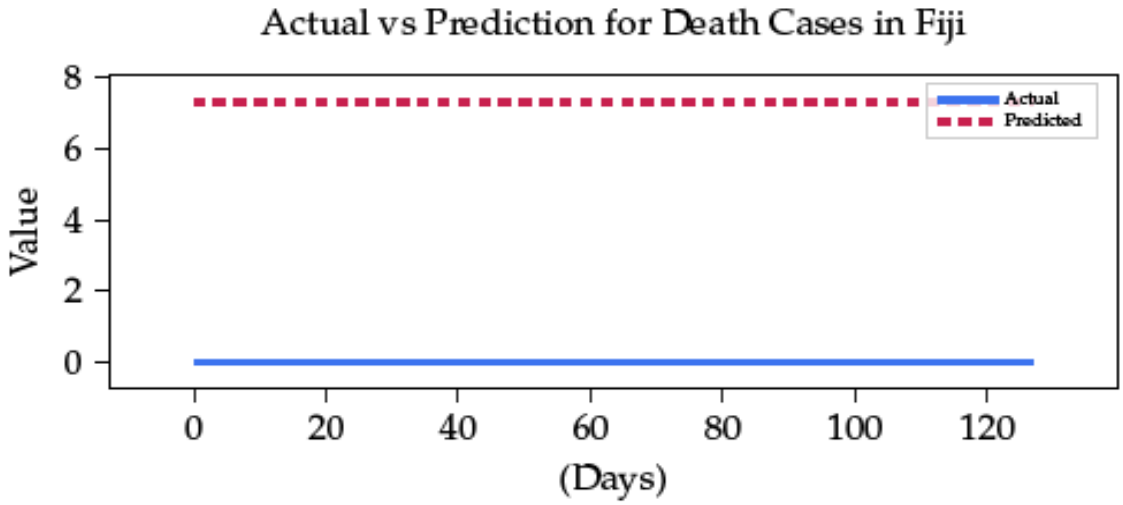

Actual vs Prediction for Recovered Cases in Mexico
$\cdot 10^4$
Value
2.8
2.6
2.4
2.2
2
0
20
40
60
80
100
120
(Days)
Actual
Prediction


Actual vs Prediction for Death Cases in Mexico
Value
4,000
2,000
0
0
20
40
60
80
100
120
(Days)
Actual
Predicted


Actual vs Prediction for Confirmed Cases in Colombia
·10^4
Value
2
1
Actual
Prediction
0
20
40
60
80
100
120
(Days)
Actual vs Prediction for Recovered Cases in Colombia
·10^4
Value
8
7
6
Actual
Prediction
0
20
40
60
80
100
120
(Days)
Actual vs Prediction for Death Cases in Colombia
Value
600
400
200
0
Actual
Predicted
0
20
40
60
80
100
120
(Days)

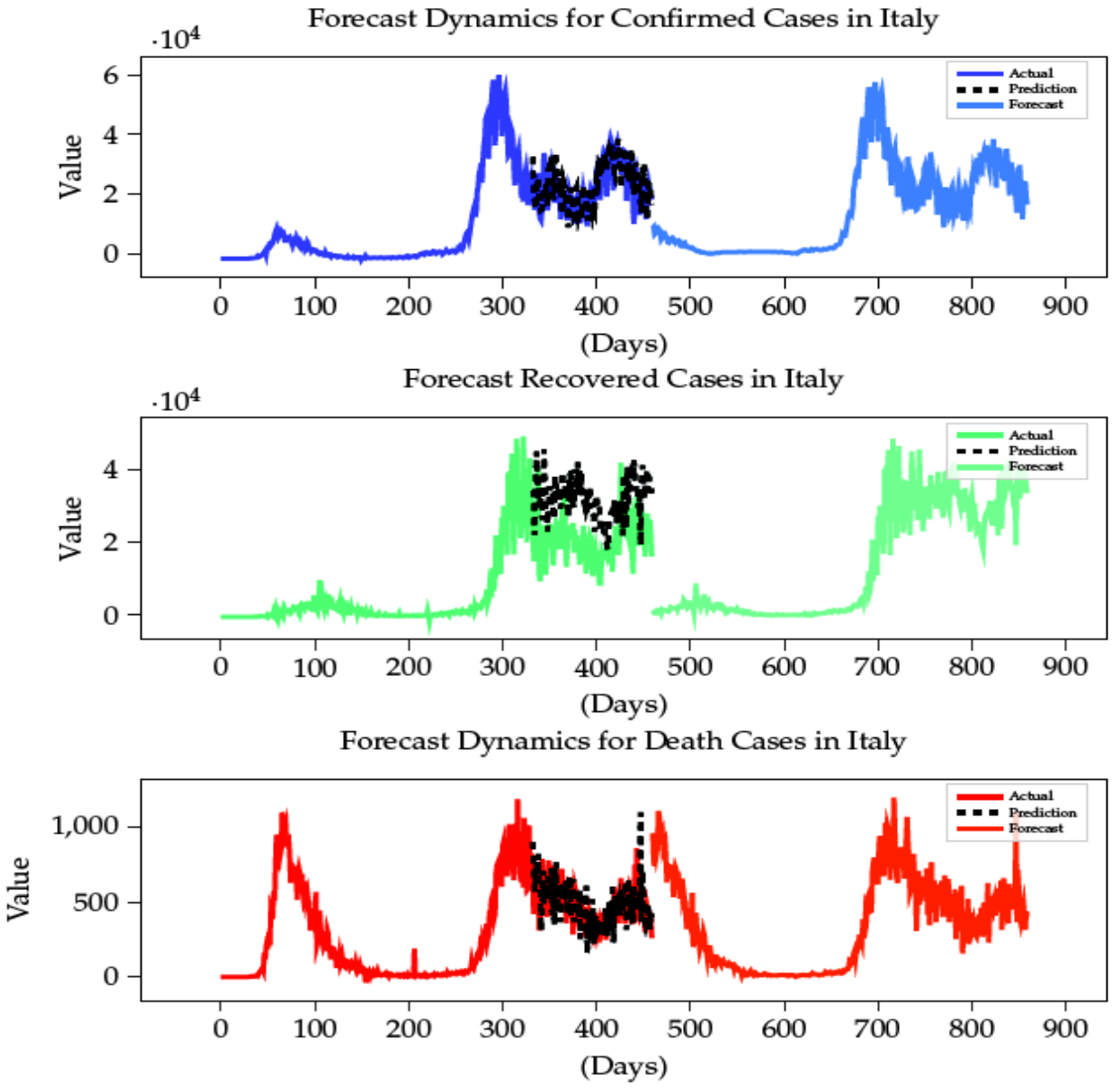

Actual vs Prediction for Confirmed Cases in Fiji
$\cdot 10^5$
Value
(Days)
Actual
Prediction

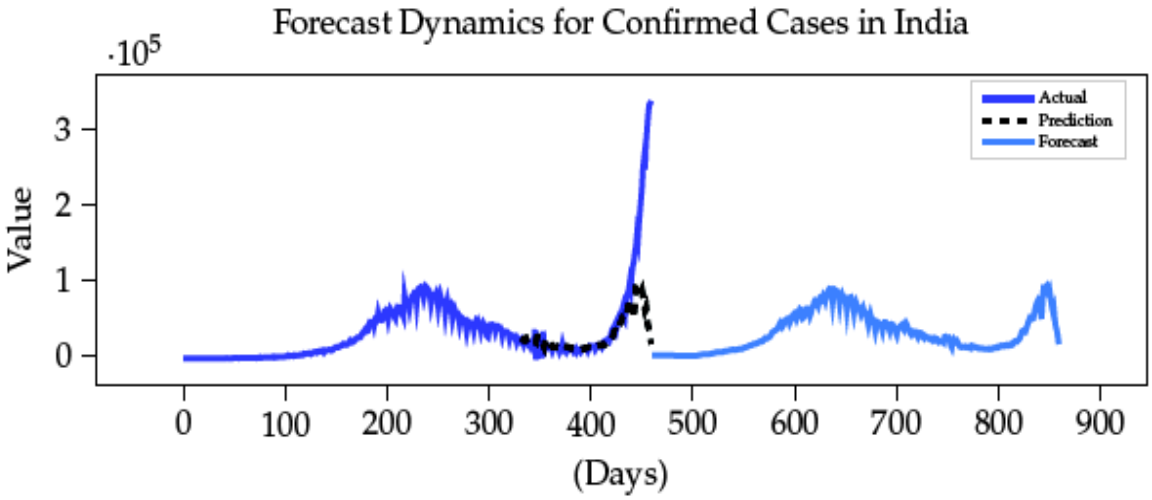

Actual vs Prediction for Recovered Cases in Fiji
Value
(Days)
Actual
Prediction

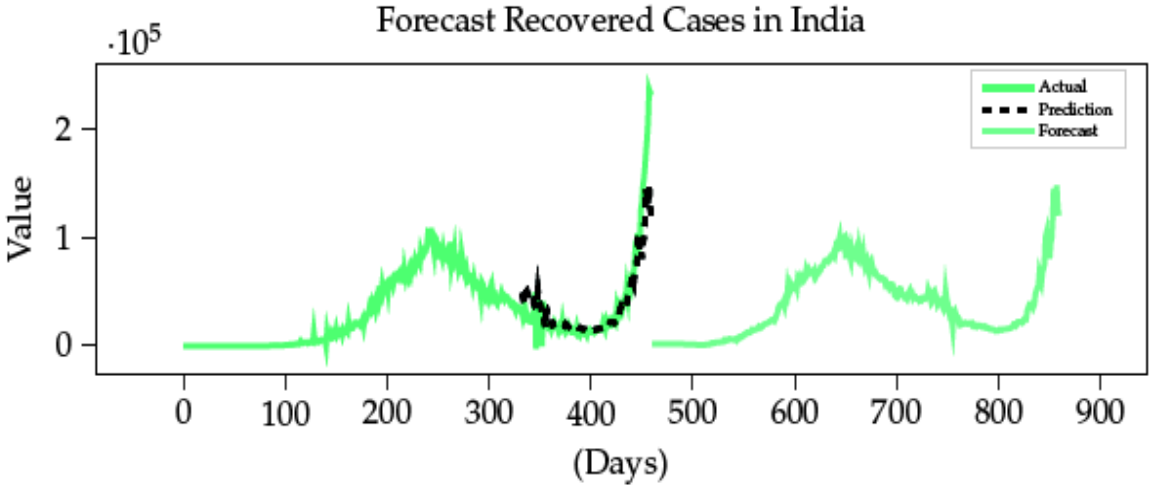

Actual vs Prediction for Death Cases in Fiji
Value
(Days)
Actual
Predicted

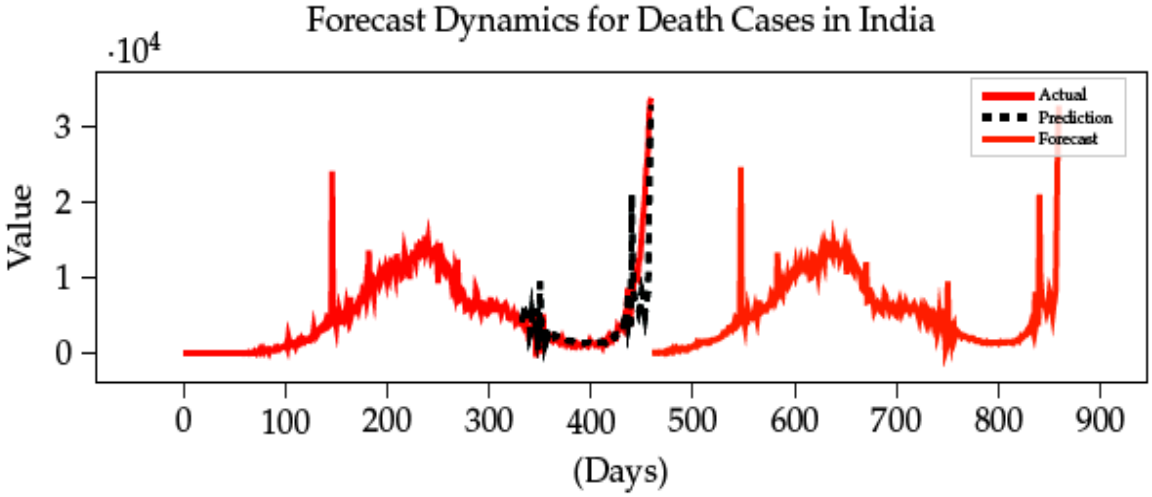
Appendix D

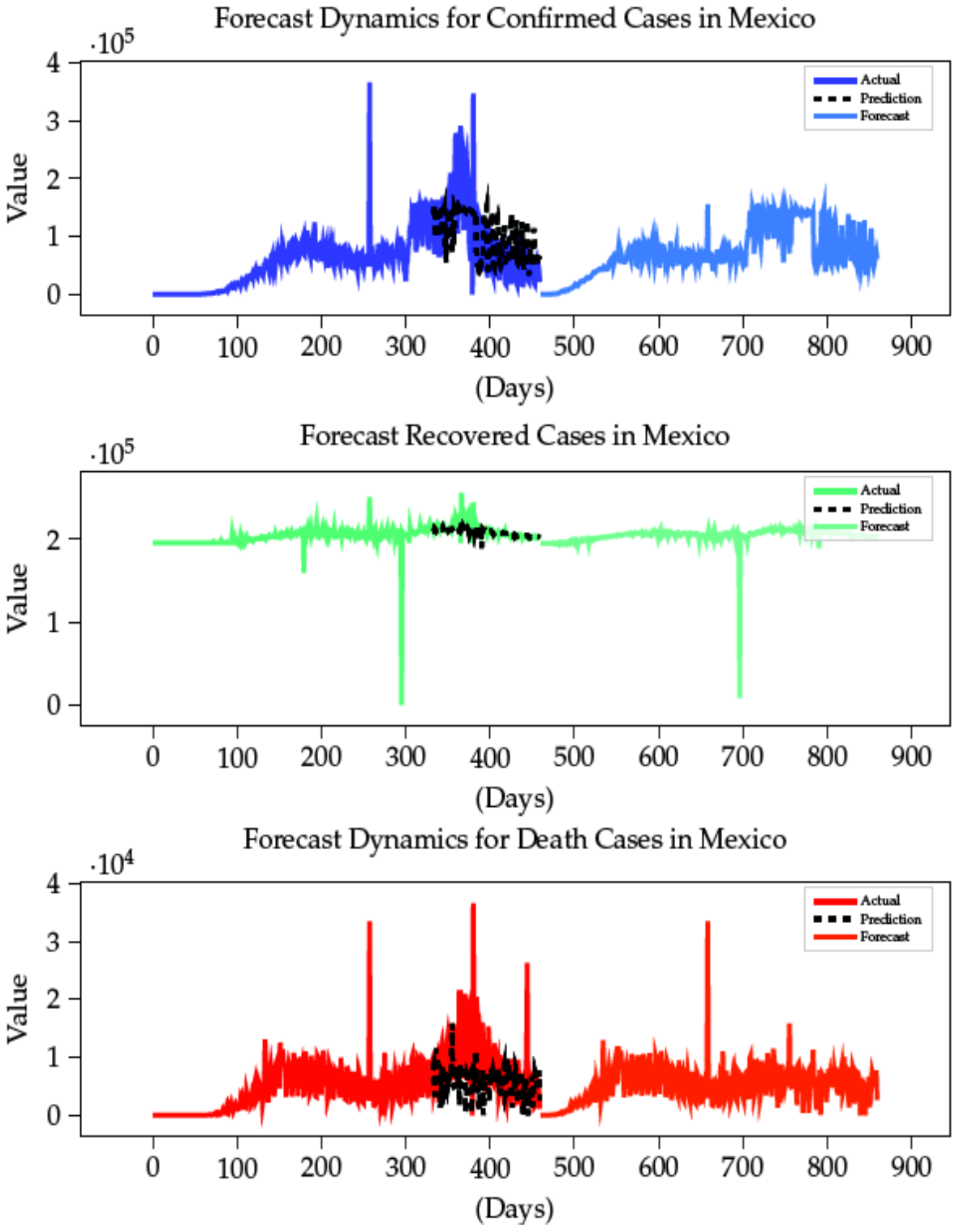
Schematics of Actual Values, Predictions, and Forecast

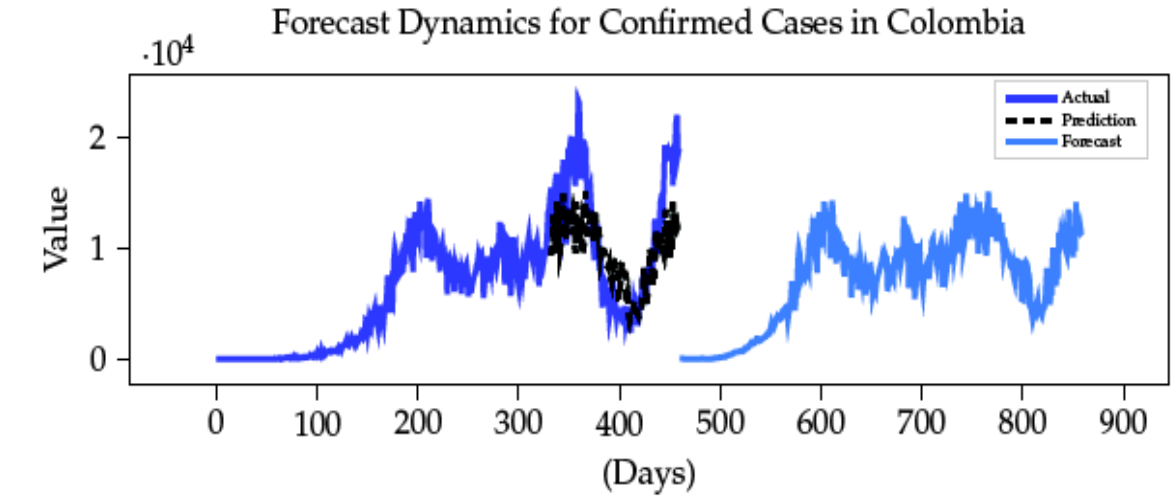


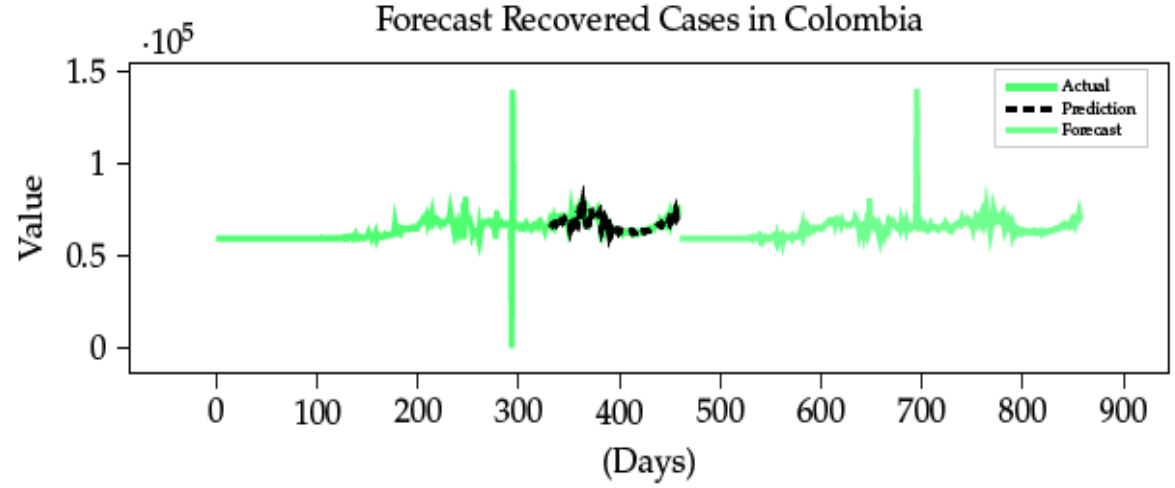

Forecast Dynamics for Confirmed Cases in India
$\cdot 10^5$
Value
(Days)
Actual
Prediction
Forecast

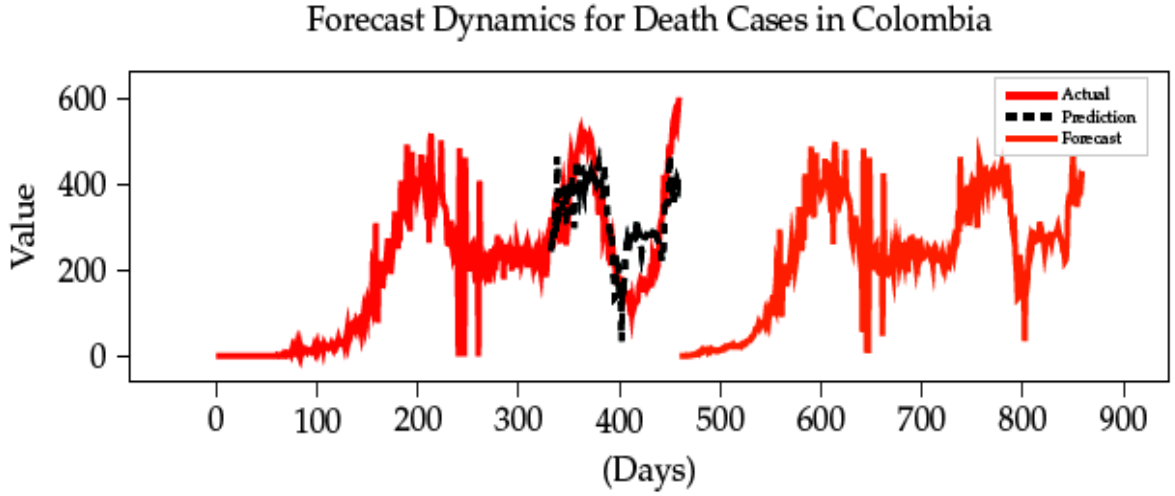

Forecast Recovered Cases in India
$\cdot 10^5$
Value
(Days)
Actual
Prediction
Forecast

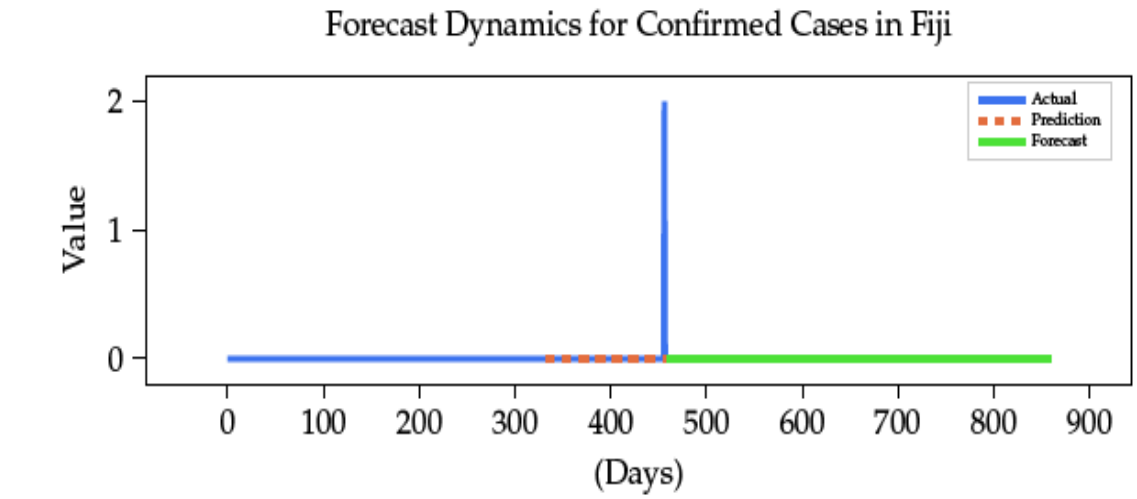

Forecast Dynamics for Death Cases in India
$\cdot 10^4$
Value
(Days)
Actual
Prediction
Forecast

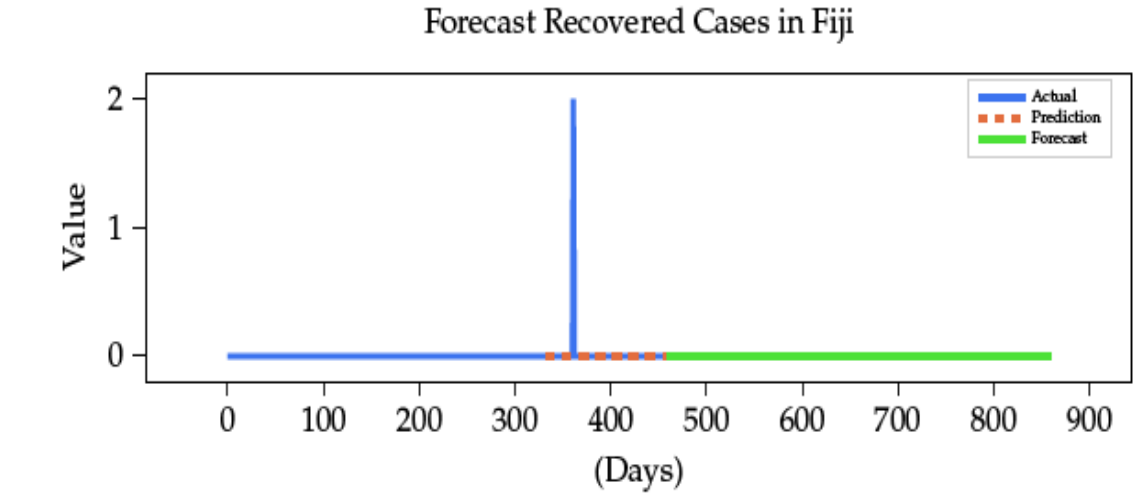

Forecast Dynamics for Confirmed Cases in Mexico
·10^5
Value
(Days)
Actual
Prediction
Forecast
Forecast Recovered Cases in Mexico
·10^5
Value
(Days)
Actual
Prediction
Forecast
Forecast Dynamics for Death Cases in Mexico
·10^4
Value
(Days)
Actual
Prediction
Forecast

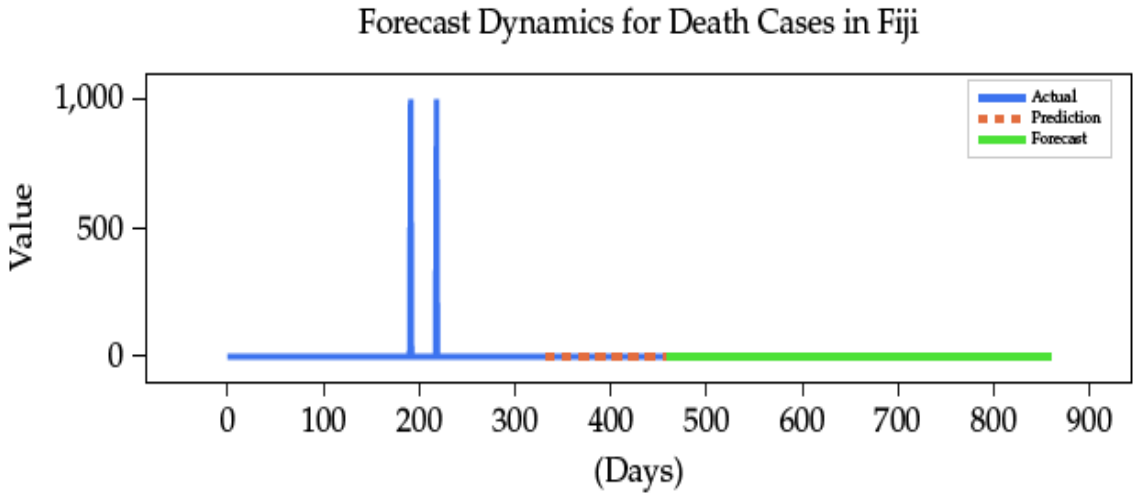

Forecast Dynamics for Confirmed Cases in Colombia
·10^4
Value
(Days)
Actual
Prediction
Forecast


Forecast Recovered Cases in Colombia
·10^5
Value
(Days)
Actual
Prediction
Forecast


Forecast Dynamics for Death Cases in Colombia
Value
(Days)
Actual
Prediction
Forecast

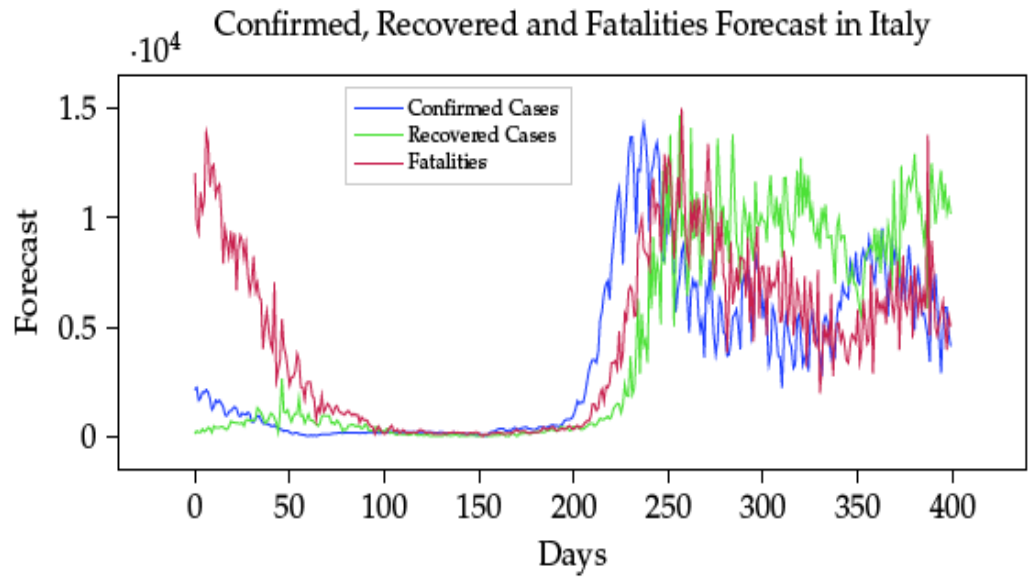

Forecast Dynamics for Confirmed Cases in Fiji
Actual
Prediction
Forecast
Value
2
1
0
0
100
200
300
400
500
600
700
800
900
(Days)

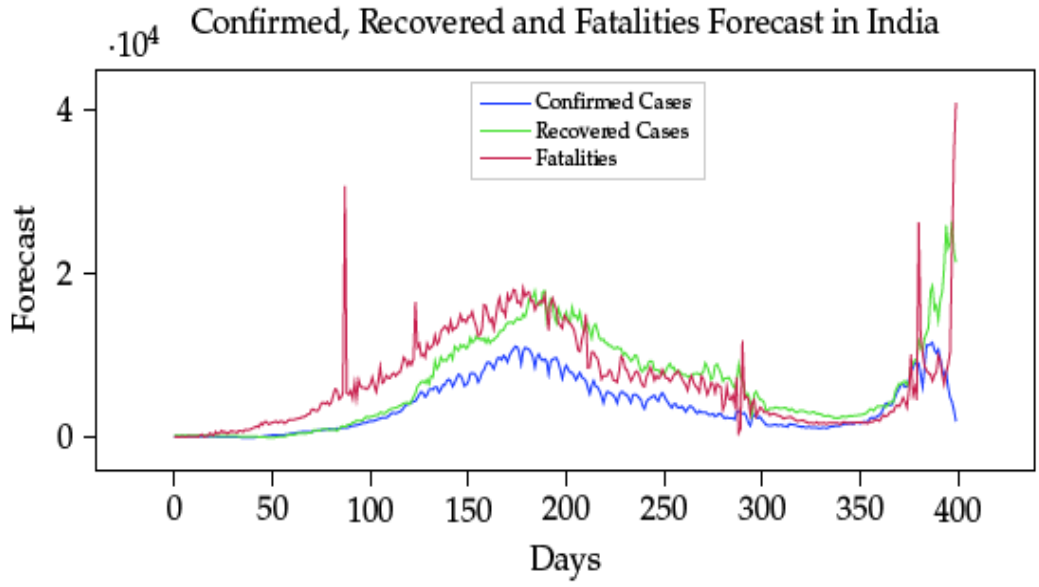

Forecast Recovered Cases in Fiji
Actual
Prediction
Forecast
Value
2
1
0
0
100
200
300
400
500
600
700
800
900
(Days)

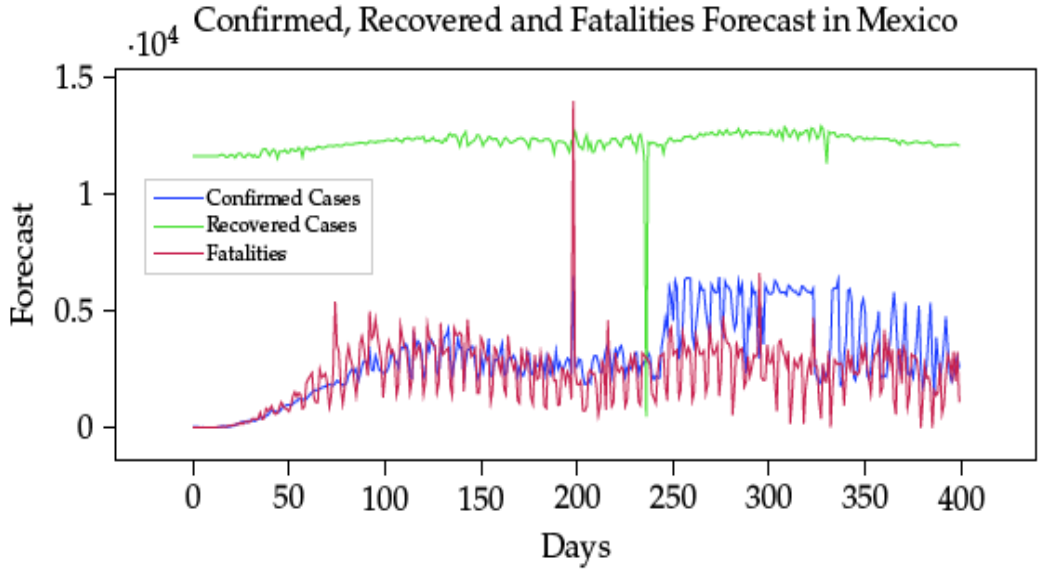

Forecast Dynamics for Death Cases in Fiji
Actual
Prediction
Forecast
Value
1,000
500
0
0
100
200
300
400
500
600
700
800
900
(Days)

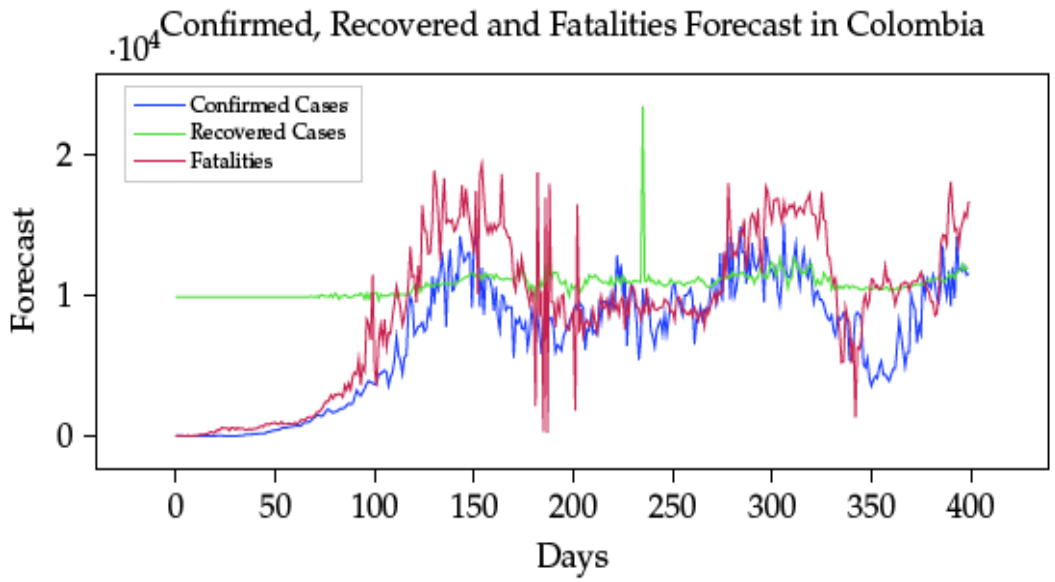
Appendix E

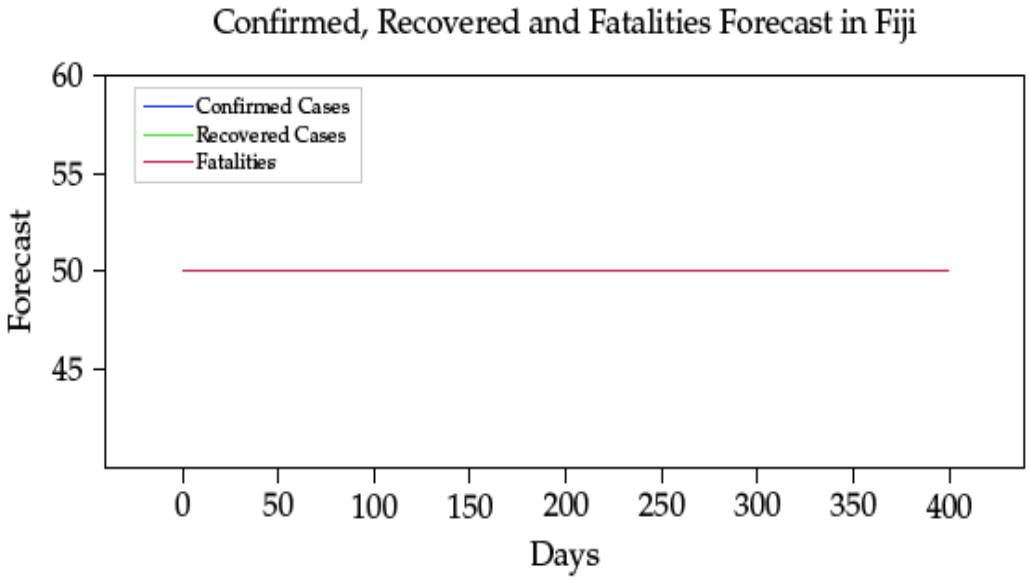
Schematics of CRD Forecast for SIIMCF Countries





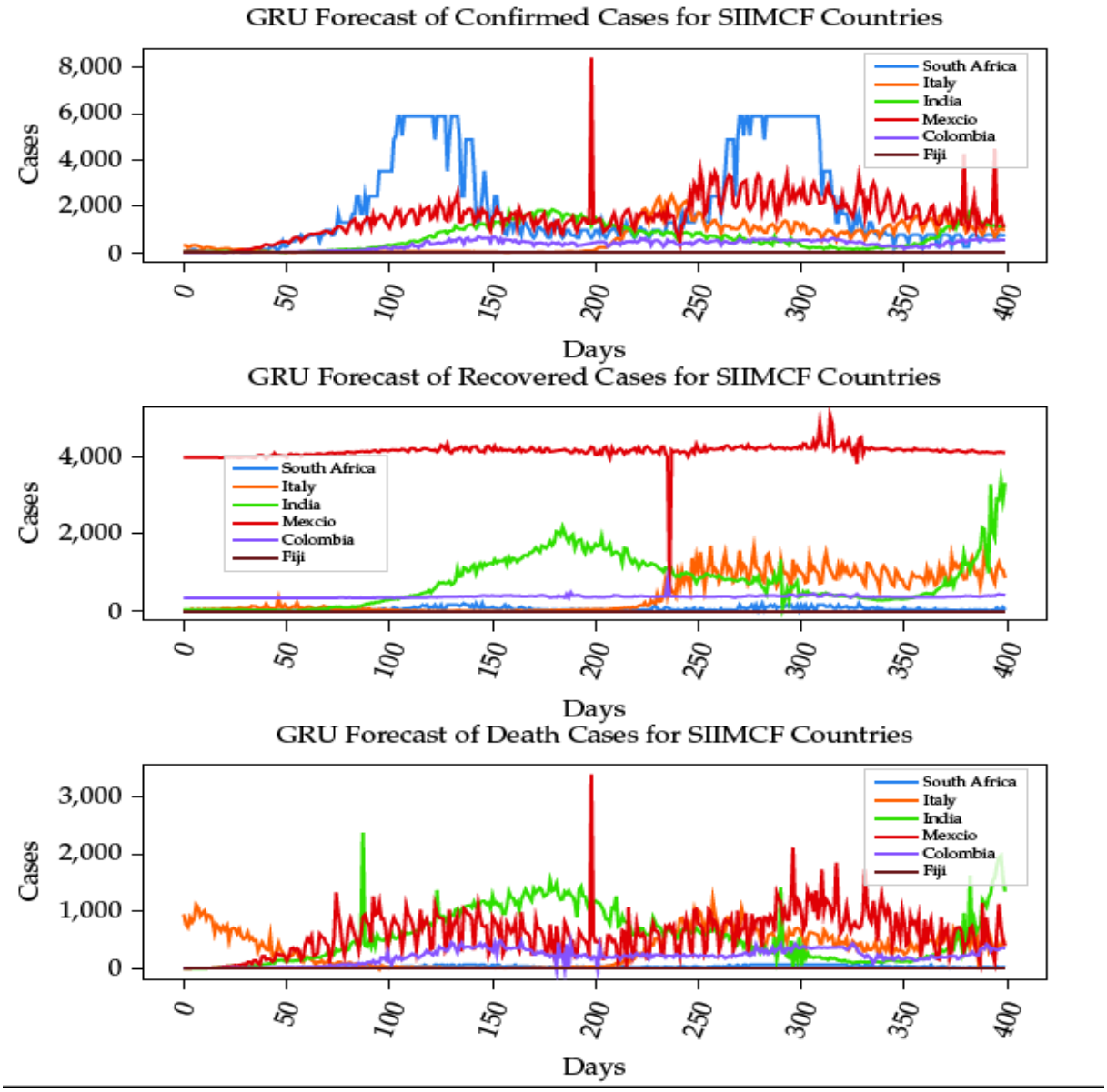

Confirmed, Recovered and Fatalities Forecast in Mexico
·10^4
1.5
1
0.5
0
Forecast
Confirmed Cases
Recovered Cases
Fatalities
0
50
100
150
200
250
300
350
400
Days

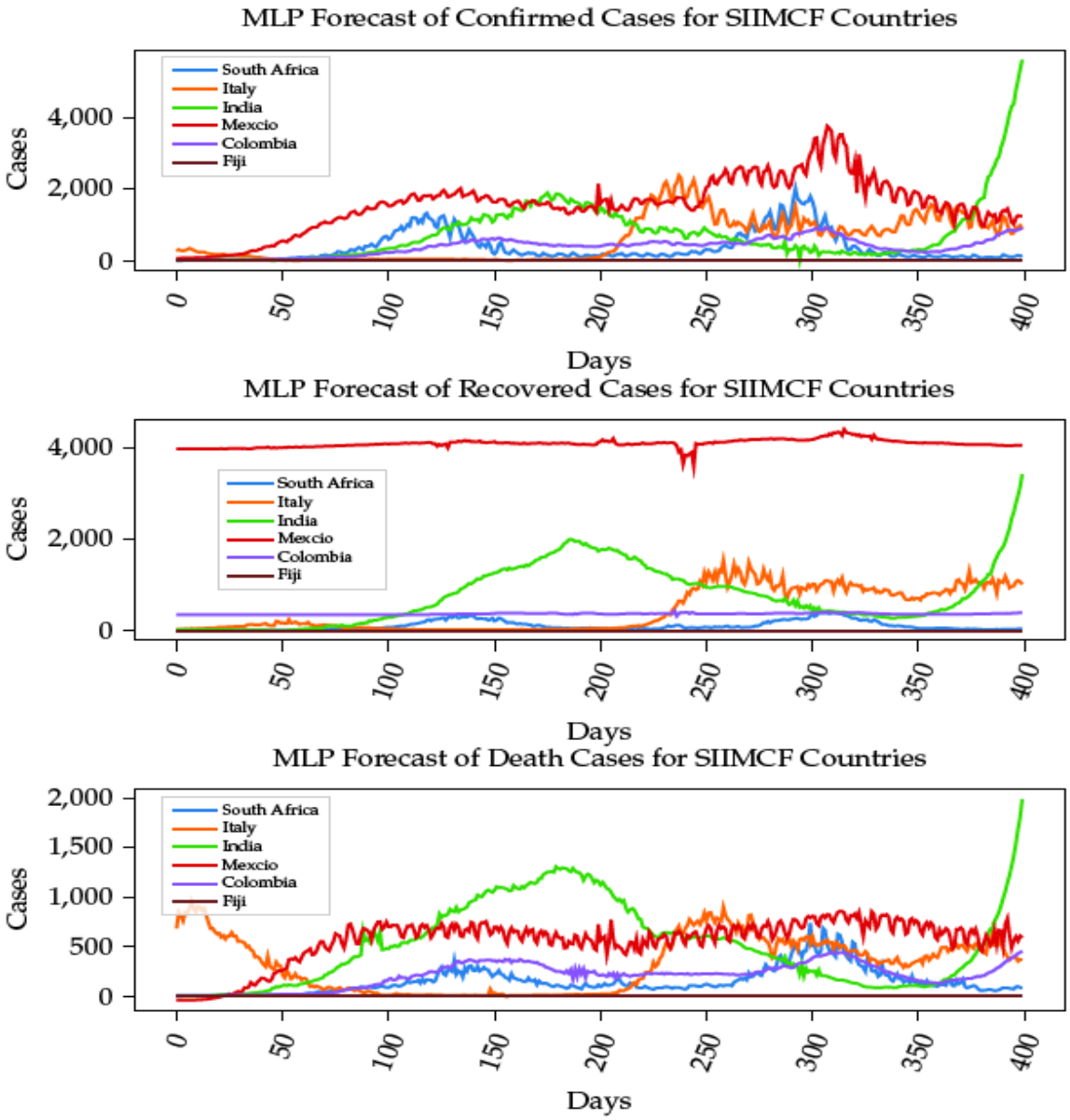

Confirmed, Recovered and Fatalities Forecast in Colombia
·10^4
2
1
0
Forecast
Confirmed Cases
Recovered Cases
Fatalities
0
50
100
150
200
250
300
350
400
Days

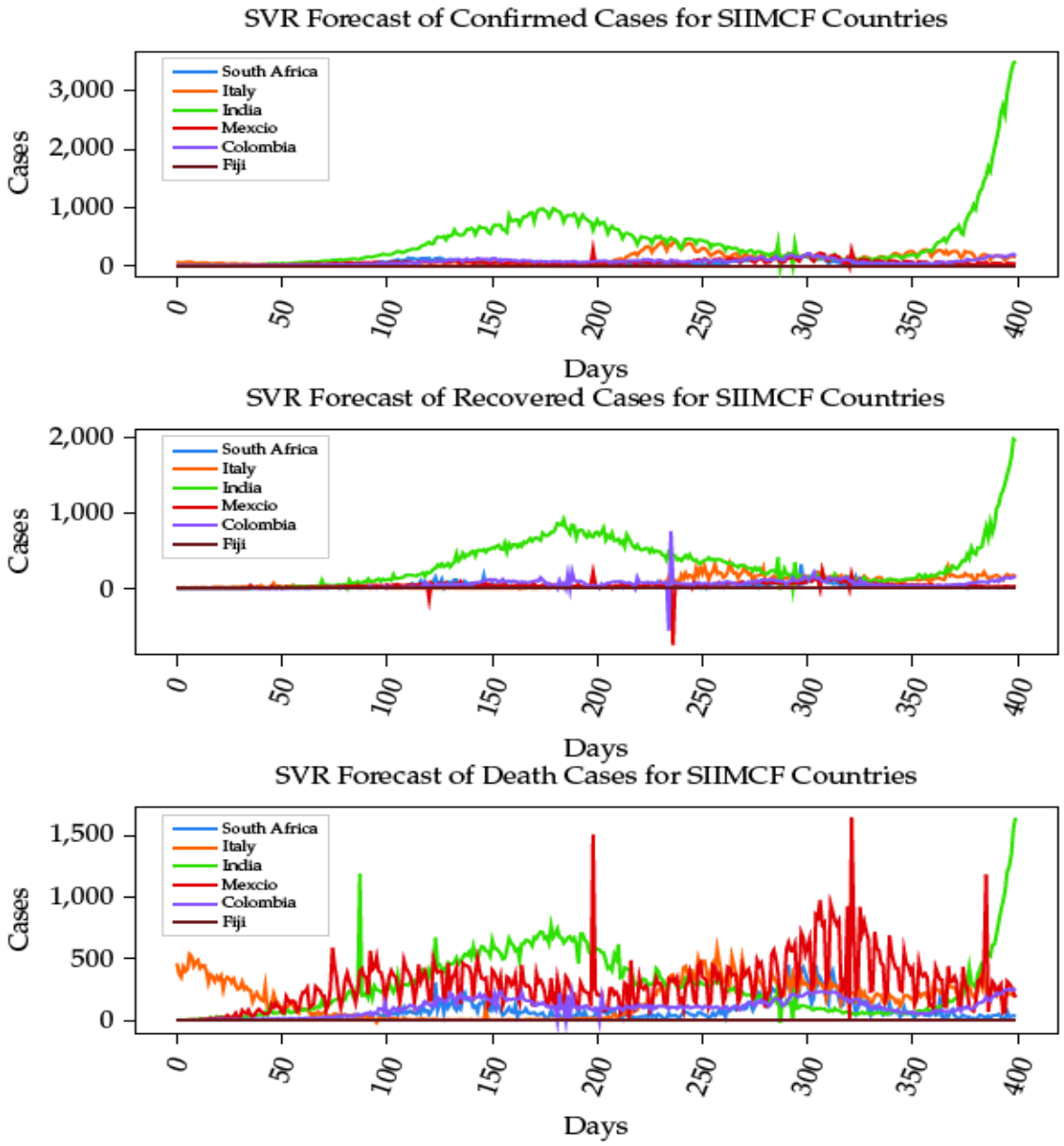

Confirmed, Recovered and Fatalities Forecast in Fiji
Confirmed Cases
Recovered Cases
Fatalities
Forecast
60
55
50
45
0
50
100
150
200
250
300
350
400
Days

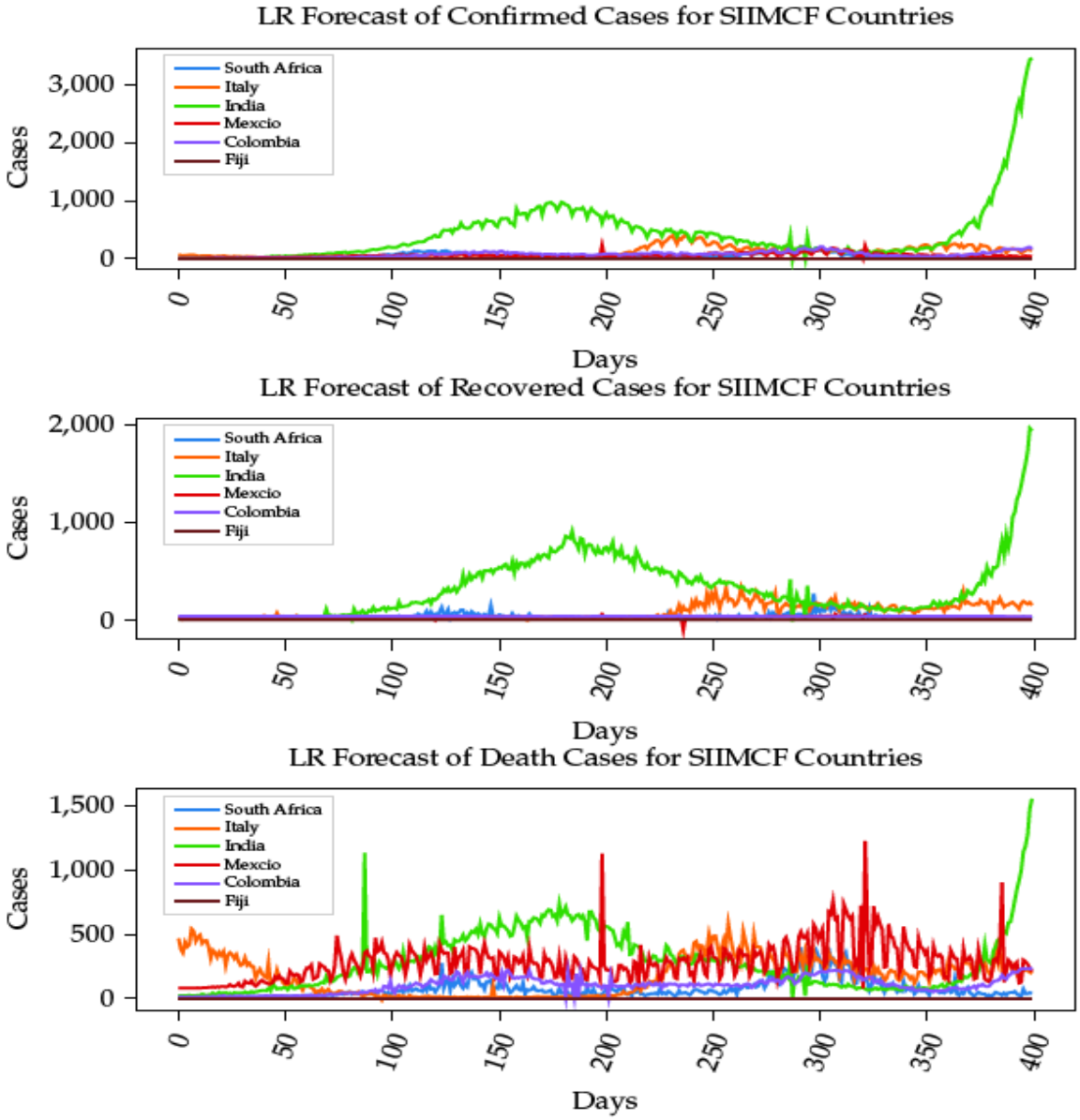
Appendix F

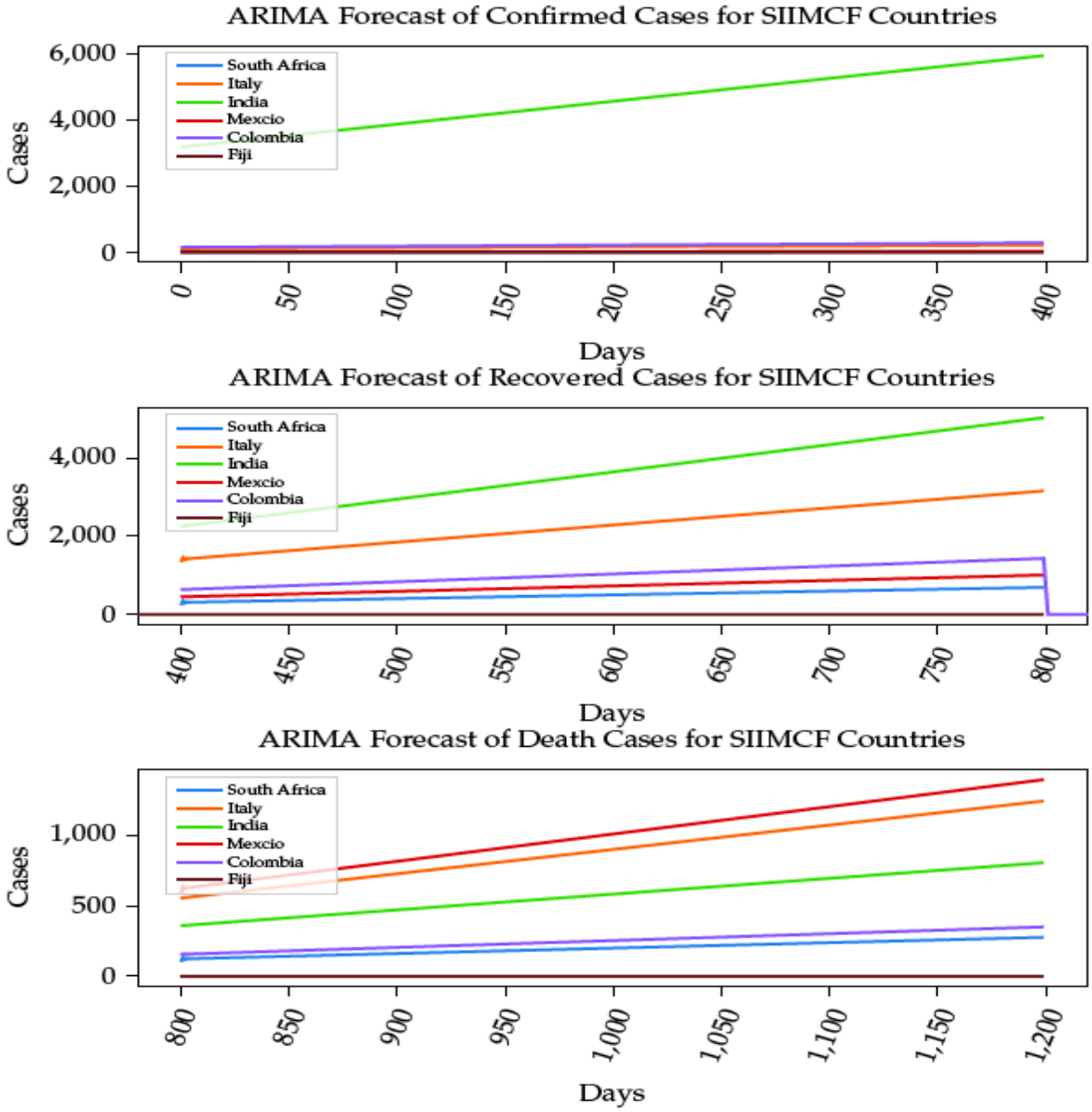
C, R and D Models Forecast

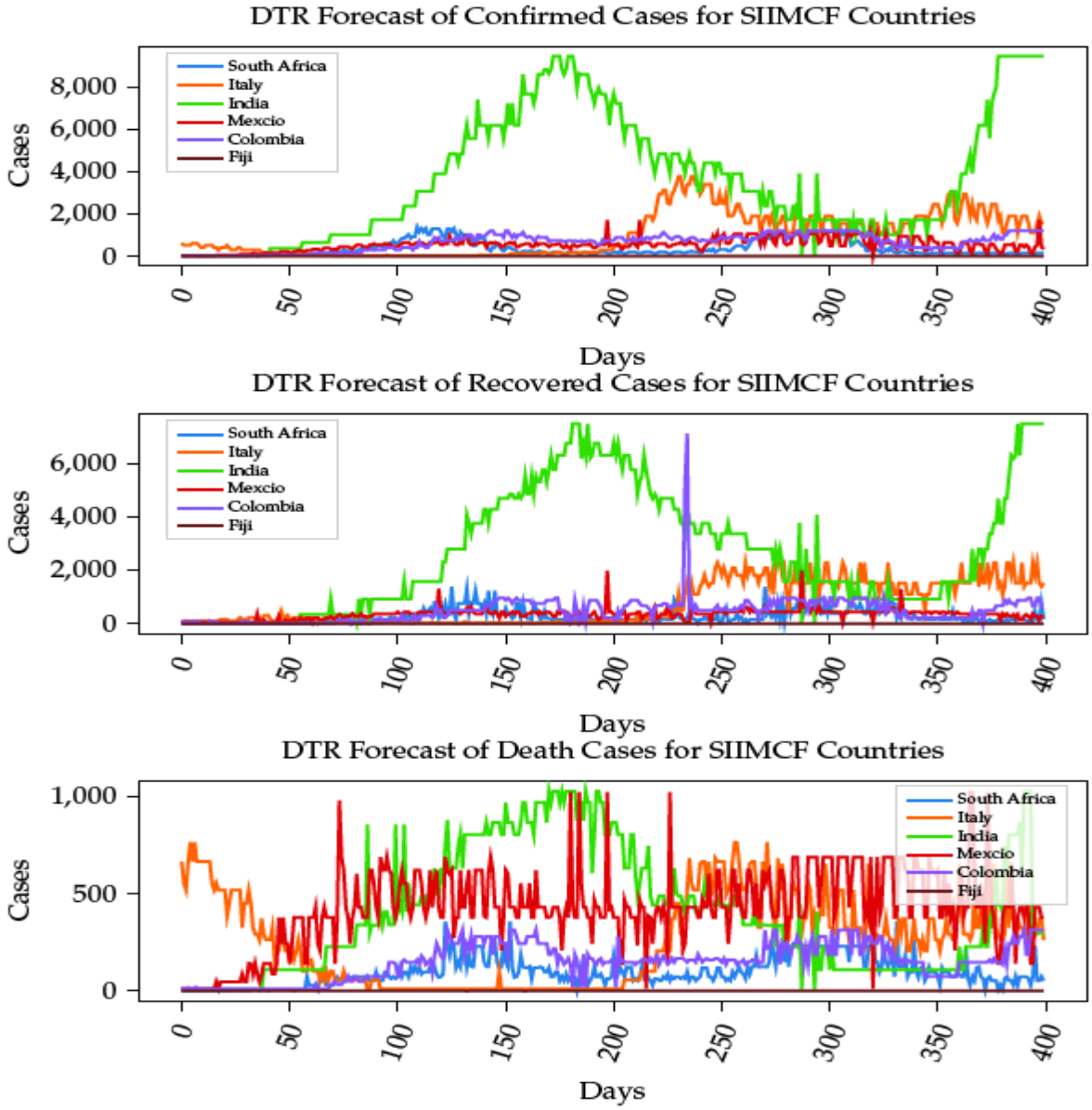


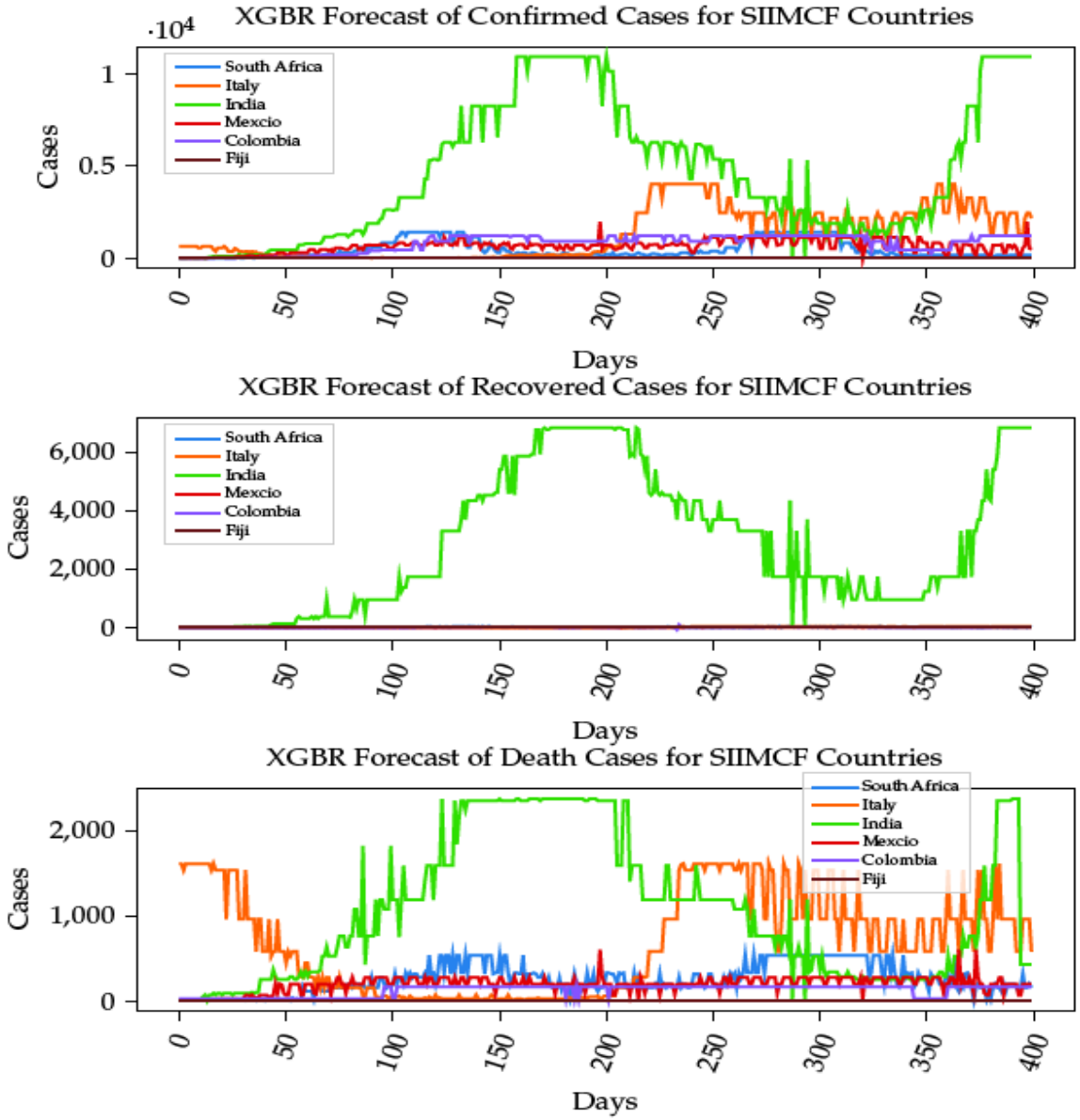

MLP Forecast of Confirmed Cases for SIIMCF Countries
South Africa
Italy
India
Mexcio
Colombia
Fiji
Cases
4,000
2,000
0
0
50
100
150
200
250
300
350
400
Days
MLP Forecast of Recovered Cases for SIIMCF Countries
South Africa
Italy
India
Mexcio
Colombia
Fiji
Cases
4,000
2,000
0
0
50
100
150
200
250
300
350
400
Days
MLP Forecast of Death Cases for SIIMCF Countries
South Africa
Italy
India
Mexcio
Colombia
Fiji
Cases
2,000
1,500
1,000
500
0
0
50
100
150
200
250
300
350
400
Days

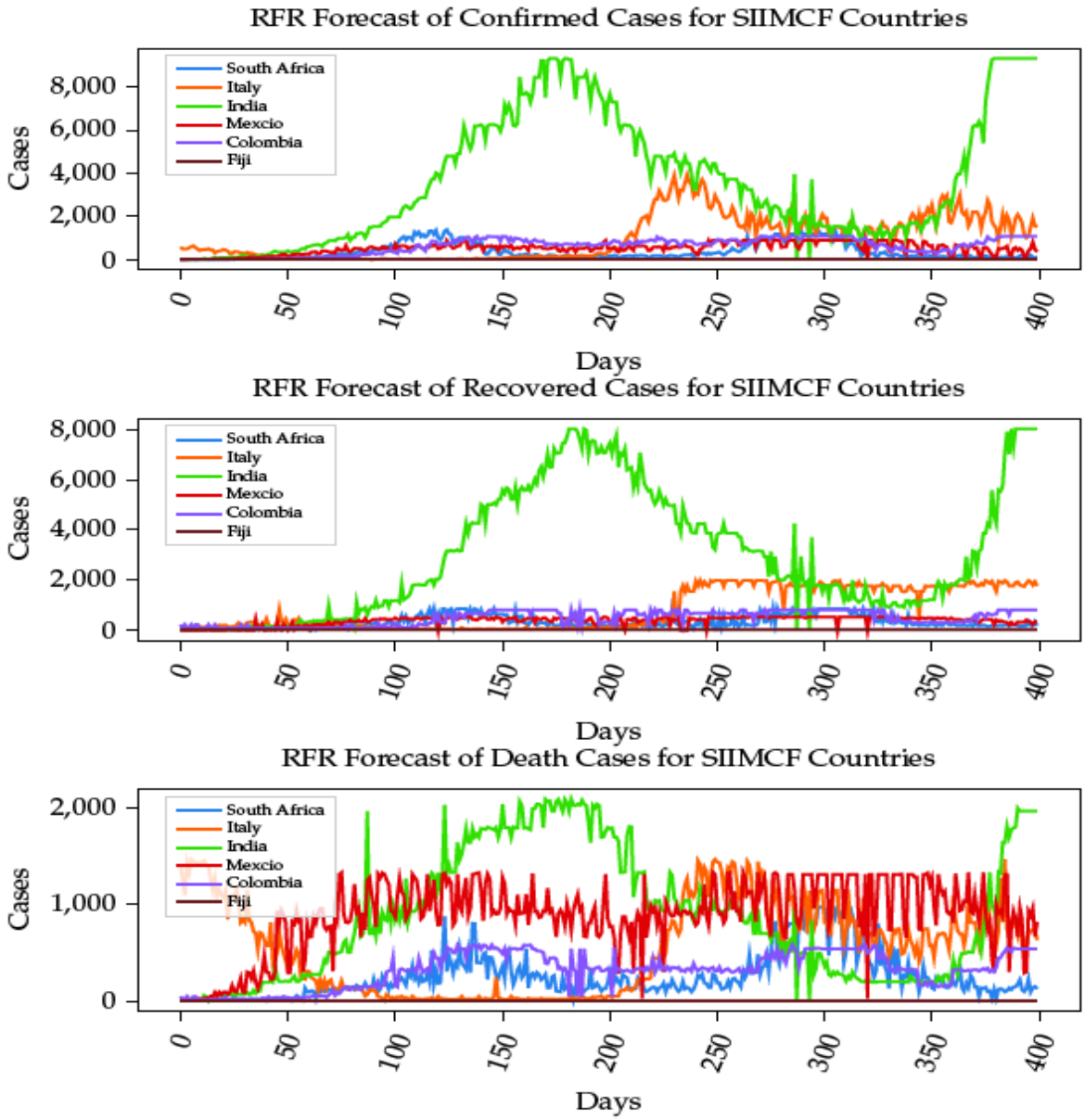

SVR Forecast of Confirmed Cases for SIIMCF Countries
South Africa
Italy
India
Mexcio
Colombia
Fiji
Cases
3,000
2,000
1,000
0
0
50
100
150
200
250
300
350
400
Days
SVR Forecast of Recovered Cases for SIIMCF Countries
2,000
1,000
0
Days
SVR Forecast of Death Cases for SIIMCF Countries
1,500
1,000
500
0
Days


LR Forecast of Confirmed Cases for SIIMCF Countries
South Africa
Italy
India
Mexcio
Colombia
Fiji
Cases
3,000
2,000
1,000
0
0
50
100
150
200
250
300
350
400
Days
LR Forecast of Recovered Cases for SIIMCF Countries
South Africa
Italy
India
Mexcio
Colombia
Fiji
Cases
2,000
1,000
0
0
50
100
150
200
250
300
350
400
Days
LR Forecast of Death Cases for SIIMCF Countries
South Africa
Italy
India
Mexcio
Colombia
Fiji
Cases
1,500
1,000
500
0
0
50
100
150
200
250
300
350
400
Days


ARIMA Forecast of Confirmed Cases for SIIMCF Countries
Cases
Days
South Africa
Italy
India
Mexcio
Colombia
Fiji
ARIMA Forecast of Recovered Cases for SIIMCF Countries
Cases
Days
South Africa
Italy
India
Mexcio
Colombia
Fiji
ARIMA Forecast of Death Cases for SIIMCF Countries
Cases
Days
South Africa
Italy
India
Mexcio
Colombia
Fiji


DTR Forecast of Confirmed Cases for SIIMCF Countries
South Africa
Italy
India
Mexcio
Colombia
Fiji
Cases
8,000
6,000
4,000
2,000
0
0
50
100
150
200
250
300
350
400
Days
DTR Forecast of Recovered Cases for SIIMCF Countries
South Africa
Italy
India
Mexcio
Colombia
Fiji
Cases
6,000
4,000
2,000
0
0
50
100
150
200
250
300
350
400
Days
DTR Forecast of Death Cases for SIIMCF Countries
South Africa
Italy
India
Mexcio
Colombia
Fiji
Cases
1,000
500
0
0
50
100
150
200
250
300
350
400
Days


XGBR Forecast of Confirmed Cases for SIIMCF Countries
·10^4
Cases
Days
South Africa
Italy
India
Mexcio
Colombia
Fiji
XGBR Forecast of Recovered Cases for SIIMCF Countries
Cases
Days
XGBR Forecast of Death Cases for SIIMCF Countries
Cases
Days


RFR Forecast of Confirmed Cases for SIIMCF Countries
South Africa
Italy
India
Mexcio
Colombia
Fiji
Cases
8,000
6,000
4,000
2,000
0
0
50
100
150
200
250
300
350
400
Days
RFR Forecast of Recovered Cases for SIIMCF Countries
South Africa
Italy
India
Mexcio
Colombia
Fiji
Cases
8,000
6,000
4,000
2,000
0
0
50
100
150
200
250
300
350
400
Days
RFR Forecast of Death Cases for SIIMCF Countries
South Africa
Italy
India
Mexcio
Colombia
Fiji
Cases
2,000
1,000
0
0
50
100
150
200
250
300
350
400
Days

Appendix G

SIIMCF Model Evaluation Metrics

The model evaluation metrics for the SIIMCF countries are presented in the tables below as defined in Section 3.8 of Chapter 3.

TABLE G.1: Model Evaluation Metrics for Italy

		S-REAL			S-SIMUL			S-SIREAL		
Model	**Cases**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**
LSTM	Confirmed	**44.09**	**27.64**	**0.91**	**48.91**	**37.73**	**0.88**	**51.49**	**36.24**	**0.72**
	Recovered	152.54	125.40	**0.82**	178.37	148.46	**0.73**	193.41	185.99	**0.65**
	Death	**10.15**	**78.77**	0.79	**45.25**	198.37	0.57	94.65	214.52	0.46
GRU	Confirmed	94.28	74.91	0.88	95.13	88.12	0.69	124.34	88.48	0.66
	Recovered	**134.95**	109.02	0.42	**124.05**	128.33	0.41	181.85	182.65	0.36
	Death	88.83	92.79	0.63	85.45	73.89	0.55	**90.48**	82.90	0.52
MLP	Confirmed	88.19	65.15	0.87	93.48	74.25	0.58	99.39	79.01	0.68
	Recovered	139.02	**104.28**	0.75	142.34	**111.88**	0.71	**179.08**	**134.56**	0.61
	Death	175.12	85.69	**0.82**	195.6	**48.39**	**0.72**	194.26	**32.87**	**0.70**
XGBR	Confirmed	232.51	462.14	0.75	267.52	4882.23	0.72	334.59	498.04	0.55
	Recovered	346.39	327.05	0.56	366.49	341.53	0.34	415.93	390.45	0.26
	Death	777.21	173.08	0.50	793.58	183.18	0.45	983.42	183.04	0.44
RFR	Confirmed	3612.41	2903.74	0.82	3921.14	3101.54	0.74	38322.31	2991.59	0.62
	Recovered	5164.32	4059.69	0.42	6129.24	4895.77	0.39	6242.36	4399.39	0.52
	Death	116.41	87.67	0.76	128.39	98.64	0.66	146.49	83.62	0.45
SVR	Confirmed	3400.86	2716.27	0.84	3459.26	2839.74	0.72	3496.39	239.39	0.48
	Recovered	4922.01	3597.85	0.66	4950.45	3392.53	0.71	5231.41	3956.33	0.52
	Death	3128.31	2198.01	0.72	3243.19	2378.91	0.71	3592.32	2859.41	0.56
LR	Confirmed	3253.21	2625.14	0.49	3485.81	2734.24	0.42	3893.54	2729.44	0.44
	Recovered	4859.93	3528.81	0.69	45102.39	3688.50	0.64	4921.33	35902.12	0.53
	Death	110.13	81.33	0.62	160.42	88.38	0.61	190.39	86.26	0.59
DTR	Confirmed	3514.31	2785.78	0.49	3612.45	2892.18	0.42	4390.43	2902.17	0.38
	Recovered	6108.21	4840.52	0.12	6215.18	4937.34	0.45	7046.59	4930.24	0.33
	Death	133.46	101.17	0.76	138.64	183.38	0.46	156.90	212.82	0.35
AR	Confirmed	7057.91	5415.09	0.70	7341.15	5628.74	0.72	9347.69	316.29	0.62
	Recovered	7033.22	5705.59	0.94	7054.49	5849.49	0.41	82093.38	5805.29	0.67
	Death	216.76	181.75	0.24	234.64	187.99	0.19	257.58	189.45	0.19
ARIMA	Confirmed	4591.77	3728.93	0.34	4728.49	3847.03	0209	4892.78	3739.43	0.21
	Recovered	12716.02	11704.15	0.25	12412.32	10550.45	0.23	19490.32	11924.33	0.22
	Death	3708.84	337.55	0.21	3790.34	349.67	0.21	3809.45	359.36	0.16

TABLE G.2: Model Evaluation Metrics for India

		S-REAL			S-SIMUL			S-SIREAL		
Model	**Cases**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**
LSTM	Confirmed	**505.05**	218.73	**0.90**	**575.38**	288.84	**0.87**	**582.91**	284.64	**0.73**
	Recovered	**218.11**	223.98	0.72	545.48	**56.39**	0.62	553.44	**83.54**	**0.68**
	Death	323.54	151.74	0.60	**102.39**	193.48	**0.91**	**166.67**	193.54	0.40
GRU	Confirmed	553.28	221.86	0.75	693.59	268.94	0.66	873.83	256.89	0.54
	Recovered	502.23	230.06	0.67	833.33	450.39	0.63	712.46	283.85	0.54
	Death	257.96	115.54	0.75	289.29	167.29	0.65	297.42	125.46	0.56
MLP	Confirmed	565.84	**80.08**	0.83	588.43	**89.83**	0.72	616.73	**89.84**	0.64
	Recovered	192.73	110.65	0.82	**167.99**	135.39	0.74	362.93	160.39	0.59
	Death	237.47	113.73	0.79	362.49	139.99	0.72	487.97	427.23	0.64
XGBR	Confirmed	768.22	859.61	0.58	892.38	902.31	0.52	768.22	1059.41	0.48
	Recovered	428.90	662.04	0.75	488.73	6390.38	0.65	892.48	362.34	0.98
	Death	609.89	330.50	**0.96**	699.52	389.38	0.53	749.34	412.67	0.51
RFR	Confirmed	1083.33	2353.08	0.89	1292.59	4330.73	0.83	3350.85	2568.43	0.66
	Recovered	305.08	1084.86	0.85	360.18	1893.59	0.55	409.39	343.36	0.25
	Death	325.25	121.53	0.96	376.22	136.85	0.68	452.89	148.89	**0.67**
SVR	Confirmed	9113.15	5376.44	0.88	9694.45	5630.59	0.59	9873.22	5585.98	0.74
	Recovered	8689.44	4969.36	0.75	85783.85	4983.93	**0.75**	8903.63	4895.95	0.59
	Death	**100.41**	**53.24**	0.96	167.94	**67.55**	0.86	228.69	39.44	0.56
LR	Confirmed	630.64	692.98	0.78	850.90	739.09	0.67	815.44	642.84	0.56
	Recovered	267.81	**42.33**	**0.95**	58.55	90.38	0.81	**101.43**	89.49	0.65
	Death	114.29	65.43	0.92	145.49	98.63	0.56	130.19	**59.95**	0.46
DTR	Confirmed	642.94	866.43	0.58	767.94	937.02	0.52	693.34	887.59	0.52
	Recovered	282.03	797.09	0.75	312.34	799.49	0.74	356.52	818.39	0.53
	Death	459.34	156.43	0.76	352.35	195.49	0.66	602.35	196.33	0.46
AR	Confirmed	565.21	541.77	0.32	595.64	660.09	0.51	395.59	901.33	0.21
	Recovered	962.82	538.04	0.44	603.66	937.24	0.34	354.35	3804.55	0.32
	Death	534.15	389.52	0.19	594.49	3453.67	0.14	589.95	349.78	0.32
ARIMA	Confirmed	7311.53	3513.36	0.44	2804.66	3994.23	0.16	4030.40	340.22	0.17
	Recovered	4266.06	3974.82	0.31	466.23	3054.20	0.41	440.45	34595.22	0.42
	Death	521.21	365.36	0.31	849.22	932.45	0.41	856.89	338.45	0.29

TABLE G.3: Model Evaluation Metrics for Mexico

		S-REAL			S-SIMUL			S-SIREAL		
Model	**Cases**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**
LSTM	Confirmed	**402.30**	369.12	**0.81**	456.15	386.22	**0.78**	**522.45**	390.21	**0.74**
	Recovered	**1699.25**	1676.70	0.58	1738.93	**392.31**	0.53	**210.45**	1982.89	0.42
	Death	598.01	391.52	0.53	698.44	457.92	0.61	892.48	420.56	0.44
GRU	Confirmed	519.25	361.68	0.72	549.03	392.59	0.49	782.48	934.44	0.42
	Recovered	1711.01	1693.34	**0.92**	1854.34	1963.33	**0.89**	3289.27	1384.55	**0.64**
	Death	490.98	337.09	0.70	**520.20**	382.34	0.69	691.10	357.80	**0.89**
MLP	Confirmed	488.58	307.71	0.48	593.44	**316.31**	0.58	741.98	456.40	0.45
	Recovered	1706.14	1692.86	0.53	2909.34	2940.03	0.63	1516.21	1490.03	0.59
	Death	**328.67**	327.91	0.72	942.95	930.33	0.42	**532.69**	359.58	0.42
XGBR	Confirmed	8566.16	6878.14	0.66	8593.23	6723.94	0.61	8893.09	6493.49	0.52
	Recovered	8078.18	7004.38	0.65	8903.28	7394.21	0.35	9301.13	7305.01	0.56
	Death	871.32	**132.55**	**0.79**	893.76	754.84	0.63	925.462	798.23	0.43
RFR	Confirmed	4869.19	3181.29	0.36	4489.32	3283.90	0.31	5230.35	3233.69	0.28
	Recovered	4518.61	2704.39	0.35	4734.16	2490.44	0.24	4902.76	2903.44	0.36
	Death	511.13	334.47	0.55	532.34	**393.56**	0.45	632.44	390.34	0.73
SVR	Confirmed	4171.68	2436.21	0.76	4276.42	2578.76	0.66	4490.72	3894.21	0.55
	Recovered	3851.08	1651.89	0.85	3739.30	1927.32	0.54	4390.82	1394.30	0.42
	Death	513.63	294.83	0.72	559.53	258.93	0.55	523.73	**384.94**	0.83
LR	Confirmed	412.93	**268.04**	0.66	434.90	572.81	0.48	712.43	**228.02**	0.58
	Recovered	201.62	**768.73**	0.85	**605.98**	794.33	0.59	717.59	**759.54**	0.59
	Death	482.04	302.25	0.79	528.53	393.48	**0.98**	536.12	653.89	0.65
DTR	Confirmed	186.93	274.68	0.36	**212.56**	849.33	0.54	564.03	483.45	0.59
	Recovered	5212.51	3013.25	0.35	3453.33	934.61	0.75	4932.40	3435.33	0.38
	Death	499.07	322.80	0.29	532.65	938.32	0.49	812.34	399.32	0.26
AR	Confirmed	5388.36	4554.12	0.36	5694.22	4495.20	0.27	5590.33	4559.42	0.32
	Recovered	4969.83	4071.04	0.45	4395.30	4390.22	0.35	5303.55	4932.34	0.44
	Death	427.62	590.67	0.47	538.59	734.66	0.37	683.27	739.33	0.67
ARIMA	Confirmed	1081.04	740.33	0.12	1201.34	690.23	0.30	1806.44	450.45	0.43
	Recovered	6120.88	5109.87	0.35	6310.43	5409.38	0.45	6385.29	5590.45	0.47
	Death	2011.25	396.68	0.36	643.60	369.44	0.26	795.95	638.28	0.11

TABLE G.4: Model Evaluation Metrics for Colombia

		S-REAL			S-SIMUL			S-SIREAL		
Model	**Cases**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**
LSTM	Confirmed	**46.49**	63.68	**0.95**	**54.01**	66.44	**0.92**	96.30	49.80	**0.88**
	Recovered	58.21	**44.59**	**0.87**	128.12	**17.45**	0.56	169.36	89.72	0.47
	Death	70.23	60.20	0.65	**13.44**	930.32	**0.83**	**27.45**	76.35	**0.64**
GRU	Confirmed	91.57	68.72	0.35	121.45	428.02	0.45	191.34	67.25	0.49
	Recovered	**59.23**	46.16	0.65	239.32	43.65	0.52	39.33	**32.01**	**0.74**
	Death	42.19	31.69	0.21	349.32	**32.54**	0.34	232.34	38.97	0.45
MLP	Confirmed	84.78	**37.02**	0.83	242.75	**39.44**	0.63	**85.90**	**33.67**	0.95
	Recovered	74.90	65.72	0.72	124.91	59.79	0.34	232.50	625.42	0.52
	Death	46.11	38.54	0.81	49.31	35.66	0.88	243.03	48.64	0.61
XGBR	Confirmed	9594.01	8396.98	0.90	1093.45	6943.48	0.45	5930.02	8495.99	0.63
	Recovered	1550.09	9454.07	0.65	2342.90	4596.33	0.36	2965.34	6956.23	0.45
	Death	248.34	226.35	0.54	325.43	219.45	0.66	251.56	233.345	0.67
RFR	Confirmed	443.49	413.82	0.90	481.72	435.44	0.85	450.45	436.22	0.85
	Recovered	86.57	948.78	0.75	124.93	1024.33	**0.85**	186.72	1044.38	0.49
	Death	59.44	44.24	0.94	102.45	234.24	0.88	142.86	124.42	0.82
SVR	Confirmed	1553.47	1112.24	0.78	1903.33	14520.41	0.68	1670.66	1630.34	0.57
	Recovered	2351.06	1596.18	0.75	3013.64	1689.21	0.26	2411.26	1609.32	0.69
	Death	**26.11**	**20.44**	0.54	67.32	32.55	0.74	46.14	**32.56**	0.64
LR	Confirmed	366.25	726.19	0.75	492.15	492.35	0.45	349.13	794.39	0.49
	Recovered	**32.16**	125.79	0.64	**59.23**	59.45	0.62	**33.49**	60.39	0.54
	Death	553.03	119.95	0.70	560.33	140.92	0.65	930.32	159.35	0.67
DTR	Confirmed	2938.42	2104.25	0.90	2494.12	2233.52	0.94	2349.56	2493.52	0.85
	Recovered	3583.22	2516.94	0.75	3693.32	2303.33	0.67	3496.32	2635.54	0.78
	Death	63.52	45.30	0.94	89.53	144.83	0.47	133.27	49.06	0.59
AR	Confirmed	6814.51	5866.10	0.35	2854.34	4950.33	0.67	6956.85	5954.21	0.73
	Recovered	5337.99	3785.63	0.55	6855.44	3842.35	0.55	5495.60	2405.32	0.61
	Death	492.71	338.48	0.35	528.63	442.44	0.54	795.44	305.33	0.66
ARIMA	Confirmed	7991.81	6139.97	0.36	3682.91	4385.37	0.38	3953.96	9053.54	0.39
	Recovered	5080.28	410.78	0.56	534.18	424.68	0.66	3920.33	4910.33	0.46
	Death	510.23	409.62	0.51	595.13	490.45	0.53	752.59	549.58	0.61

TABLE G.5: Model Evaluation Metrics for Fiji

		S-REAL			S-SIMUL			S-SIREAL		
Model	**Cases**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**
LSTM	Confirmed	**0.12**	**0.04**	0.48	**0.11**	**0.03**	0.38	0.14	**0.04**	0.32
	Recovered	0.14	0.04	0.44	**0.12**	0.21	0.27	**0.12**	**0.3**	0.29
	Death	**0.01**	**0.01**	0.10	**0.01**	**0.02**	0.10	**0.01**	**0.03**	**0.64**
GRU	Confirmed	0.14	0.09	**0.52**	0.16	0.5	0.8	0.12	1.15	0.19
	Recovered	0.14	0.04	**0.67**	2.34	5.64	0.15	5.26	3.61	**0.45**
	Death	42.19	31.69	0.21	5.95	3.26	0.31	6.17	5.99	0.08
MLP	Confirmed	0.09	0.04	0.21	14.41	1.64	0.7	**0.03**	0.05	0.04
	Recovered	**0.04**	**0.02**	0.23	23.69	21.32	**0.61**	2.09	4.41	0.30
	Death	0.11	0.11	0.01	3.12	3.71	0.10	6.70	8.90	0.16
XGBR	Confirmed	1.12	0.65	0.15	3.16	4.33	0.55	4.78	2.42	**0.85**
	Recovered	0.87	0.56	0.10	0.26	**0.14**	0.23	2.37	10.81	0.24
	Death	0.39	0.32	0.90	0.23	0.74	0.40	0.57	0.84	0.32
RFR	Confirmed	1.12	0.43	0.19	0.15	0.35	**0.75**	0.32	0.63	0.20
	Recovered	0.85	0.31	0.10	0.72	0.65	0.20	10.01	0.36	0.20
	Death	0.01	0.01	0.90	0.15	0.10	0.25	2.10	0.70	0.81
SVR	Confirmed	1.14	0.41	0.50	1.32	0.44	0.06	3.14	0.52	0.44
	Recovered	0.85	0.30	0.10	0.56	0.45	0.32	0.33	0.41	0.15
	Death	0.10	0.19	0.90	0.20	0.12	0.14	5.33	10.92	0.28
LR	Confirmed	1.13	0.44	0.50	6.19	3.37	0.03	2.18	0.46	0.40
	Recovered	0.83	0.31	0.10	0.88	0.46	0.21	6.81	43.51	0.30
	Death	0.06	0.06	0.50	4.25	10.26	**0.54**	0.65	0.25	0.23
DTR	Confirmed	1.13	0.44	0.55	3.15	5.40	0.65	2.10	0.44	0.51
	Recovered	0.85	0.30	0.10	3.84	0.39	0.27	14.65	12.32	0.53
	Death	0.06	0.06	0.90	0.65	12.69	0.34	13.89	25.12	0.60
AR	Confirmed	1.13	0.46	0.27	12.63	5.5	0.40	10.33	23.35	0.25
	Recovered	0.85	0.32	0.01	29.02	3.41	0.26	12.43	18.31	0.19
	Death	0.01	0.01	0.10	0.4	0.9	0.35	0.08	0.35	0.34
ARIMA	Confirmed	3.17	3.09	0.03	5.18	3.32	0.38	13.54	19.55	0.31
	Recovered	1.52	1.47	0.11	61.32	10.17	0.20	13.22	18.22	0.45
	Death	4.17	3.63	0.13	3.15	3.16	0.50	3.15	3.83	0.14

Appendix H

Disease Prevalence Days (S-SIMUL and S-SIREAL)

In this section, we present the disease prevalence days for the forecast models considered in this study relating to the simulated datasets together with the hybrid of simulated and real-world datasets. The interpretation of the table value is consistent as described in Section 4.6.4.4 of Chapter 4.

Tables H.1 and H.2 represents the disease prevalence days of the models relating to the simulated datasets, whereas, Tables H.3 and H.4 represents the disease prevalence days of the models relating to the hybrid of simulated and real world datsets.

TABLE H.1: Surrogate Disease Prevalence Day Forecast (S-SIMUL)

Country	**Case**	**LSTM**	**GRU**	**MLP**	**XGBR**	**SVR**
South (Africa)	C	65±4 – 150±6 230±5 – 340±4	110±5 – 114±5 260±4 – 330±5	92±4 – 1146±4 245±6 – 330±5	96±11 – 135±8 250±5 – 300±3	90±6 – 112±3 222±3 – 312±4
	R	220±5 – 130±3 224±4 – 320±4	95±2 – 135±5 280±7 – 280±8	80±5 – 135±4 160±2 – 315±3	65±5 – 150±10 250±5 – 315±3	110±10 – 150±5 250±2 – 340±2
	D	103±4 – 120±6	110±5 – 120±3	60±5 – 120±5	140±3 – 125±5	124±3 – 142±4
Italy	C	215±8 315±3 – 320±6	120±3 – 320±2 325±2	205±4 312±1 – 370±3	215±3 310±5	215±8 250±5 – 390±4
	R	240±3 – 310±3 325±4	223±6 – 315±2	575±2 – 315±3 365±3	220±3 – 320±3	230±3 – 240±3 315±7
	D	37±5 – 335±4 363±3	48±2 – 310±4 363±3	55±8 – 236±8 325±3	40±3 272±3 – 310±3 375±10	38±5 220±4 – 349±2 379±2
India	C	123±3 – 285±4 321±4 – 365±5	123±3 – 289±5 315±3 – 320±2	120±5 – 215±5 362±2 – 410±2	220±5 – 225±4 335±3 – 319±5	132±5 – 220±7 345±3
	R	215±3 – 310±3 355±4	212±8 – 322±3 312±3	244±5 – 32±6 330±4 – 395±4	125±4 – 245±4	220±4 – 320±3 335±6
	D	7±3 – 220±3 220±3	15±9 – 330±3 325±3	12±6 – 218±5 340±3	18±5 – 230±8 344±9	22±4 – 280±7 385±2
Mexico	C	253±5 – 325±2 312±4 – 345±3	245±5 – 320±4 335±5	215±4 – 320±4	100±5 – 220±2 255±5 – 330±4	273±6 – 345±3
	R	135±4 – 312±4	138±5 – 340±2	135±4 – 357±6	132±10 – 325±7	127±2 – 320±5
	D	177±3	180±9 325±5	90±6 – 230±6 220±4 – 315±6	85±4 – 230±4 345±10	55±4 – 210±3 223±5 – 352±3
Colombia	C	124±3 – 130±4 225±3 – 315±7	19±8 – 130±3 230±4 – 310±9	135±6 – 185±5 232±2 – 349±4	125±4 – 120±5 255±6 – 325±4	120±7 – 132±2 265±5 – 310±6
	R	228±2 314±4	229±4 290±5	163±4 – 210±6 230±4 – 245±3	235±5	145±4 – 230±4
	D	115±6 – 135±2 247±2 – 312±2 315±4	115±9 – 156±2 288±4 – 325±2 370±2	80±5 – 145±6 225±3 – 318±9 392±2	45±2 – 165±2 232±2 – 319±2 335±5	125±2 – 132±2 338±4 – 335±4 346±7
Fiji	C	–	–	–	–	–
	R	–	–	–	–	–
	D	–	–	–	–	–

TABLE H.2: Surrogate Disease Prevalence Day Forecast (S-SIMUL continued...)

Country	Case	RFR	DTR	LR	AR	ARIMA
South Africa	C	87±5 – 130±3 245±5 – 324±2	29±4 – 134±3 245±4 – 313±3	80±7 – 135±3 249±2 – 330±9	30±20 115±5 – 226±2	130±10 215±5
	R	86±3 – 115±3 223±3 – 312±3	127±8 – 145±2 27±3 – 313±5	98±4 – 135±5 245±3 – 322±4	134±8 – 140±5 115±4 – 135±3	215±8 235±9
	D	105±2 – 123±6	112±3 – 189±4	129±5 – 123±2	43±6	188±5
Italy	C	228±8 334±3	187±4 359±3	245±5 – 232±6 330±2	173±3	134±4
	R	264±9 – 376±5	213±2 – 313±4 359±5	244±6 – 329±4 321±4	340±2 356±3	388±4
	D	34±2 – 39±3 382±4	21±3 367±5	14±3 – 29±3 353±3	18±3	42±3
India	C	124±4 – 247±3 334±3	145±6 – 235±5 364±3	134±3 – 277±4 367±2	132±4 128±5 – 340±8	315±3 153±6
	R	223±6 – 312±4 353±4 – 317±3	29±4 – 334±3 342±4 – 386±3	246±3 – 250±5 342±4	14±2 – 376±7 273±4 – 337±8	384±3
	D	122±5 – 241±5 340±2 – 386±3	128±2 – 225±6 347±3	134±2 – 218±3 366±7	139±0	176±5
Mexico	C	92±3 – 165±4 250±3 – 350±8	221±4 250±6 – 350±3	253±4 – 376±4	147±3	134±7
	R	108±4 – 284±4	129±3 – 317±6	143±4 – 356±3	140±2	167±8
	D	67±7 – 188±3 249±6 – 370±3	63±4 – 172±8 321±8 – 340±9	92±3 – 177±3 260±4 – 332±3	58±12	123±5
Colombia	C	112±3 – 154±4 237±3 – 312±3 375±9	128±4 – 145±8 242±2 – 313±4 363±3	125±4 – 166±8 273±2 – 310±4 364±2	128±3	176±2
	R	132±2 – 213±2 265±3 – 312±8	223±7 – 213±2	280±2 – 299±2	135±5	165±9
	D	112±4 – 162±5 255±3 – 312±5 394±8	128±6 – 142±3 321±4 – 334±2 381±5	151±8 – 149±4 275±4 – 349±4 392±4	137±6	284±3
Fiji	C	–	–	–	–	–
	R	–	–	–	–	–
	D	–	–	–	–	–

TABLE H.3: Surrogate Disease Prevalence Day Forecast (S-SIREAL)

Country	Case	LSTM	GRU	MLP	XGBR	SVR
South (Africa)	C	83±4 – 149±6 268±4 – 324±3	92±3 – 160±4 284±4 – 324±2	64±3 – 142±5 265±4 – 320±5	78±3 – 1640±4 246±4 – 323±2	78±5 – 115±4 238±4 – 342±3
	R	280±5 – 143±2 243±3 – 325±6	103±2 – 135±4 253±3 – 323±6	120±5 – 152±8 271±4 – 360±3	15±5 – 136±4 268±3 – 332±7	115±6 – 165±3 284±2 – 312±3
	D	113±2 – 145±5	135±2 – 142±5	126±3 – 145±2	125±2 – 136±5	113±4 – 132±2
Italy	C	212±5 343±3 – 396±2	223±5 – 386±3 381±4	214±3 372±4 – 384±2	226±4 369±3	219±5 334±2 – 392±4
	R	267±3 – 352±5 381±3	256±5 – 368±3	242±4 – 314±2 372±4	242±6 – 383±2	255±4 – 293±2 353±4
	D	42±3 – 304±3 383±2	41±3 – 385±3 391±8	43±5 – 283±5 371±3	41±8 244±1 – 332±3 373±3	38±3 221±3 – 324±2 382±2
India	C	118±5 – 231±2 362±2 – 382±3	134±3 – 252±2 334±4 – 376±3	115±2 – 235±3 345±2 – 383±4	220±3 – 232±3 350±5 – 365±3	126±5 – 237±7 363±4
	R	236±2 – 335±3 369±4	232±2 – 322±3 383±4	240±5 – 29±5 352±3 – 367±2	116±6 – 236±3	234±3 – 288±2 354±2
	D	7±9 – 212±2 355±4	16±4 – 214±5 362±5	14±6 – 233±4 384±2	15±5 – 244±3 363±4	18±4 – 256±4 367±4
Mexico	C	234±4 – 335±7 353±3 – 387±3	262±5 – 291±6 369±3	271±4 – 304±5	86±3 – 234±3 245±3 – 332±9	253±3 – 311±6
	R	132±4 – 312±5	145±3 – 363±3	159±2 – 362±3	132±9 – 339±6	135±2 – 371±1
	D	185±3	189±3 302±3	52±3 – 212±4 222±7 – 352±3	67±2 – 215±2 385±6	45±4 – 210±9 252±8 – 353±9
Colombia	C	123±2 – 150±4 265±2 – 34±4	19±4 – 174±2 225±3 – 334±2	123±3 – 216±45 310±8 – 302±5	131±8 – 172±3 243±5 – 312±9	148±3 – 155±4 263±2 – 267±3
	R	268±2 304±9	234±2 308±3	143±1 – 190±6 229±1 – 234±9	254±2	172±4 – 252±2
	D	125±2 – 134±9 232±2 – 332±2 365±3	121±3 – 172±9 275±4 – 323±2 345±3	125±9 – 201±3 310±4 – 332±8 369±4	109±3 – 174±5 202±3 – 336±6 365±3	128±2 – 132±1 308±2 – 329±3 378±5
Fiji	C R D	– – –	– – –	– – –	– – –	– – –

TABLE H.4: Surrogate Disease Prevalence Day Forecast (S-SIREAL continued...)

Country	**Case**	**RFR**	**DTR**	**LR**	**AR**	**ARIMA**
South Africa	C	76±3 – 135±4 239±4 – 313±1	109±2 – 165±4 267±2 – 343±6	95±3 – 172±8 273±4 – 330±3	38±13 89±3 – 204±3	157±7 195±3
	R	111±1 – 162±4 240±2 – 312±4	94±3 – 152±6 278±4 – 313±8	103±5 – 138±4 271±3 – 326±2	124±6 – 175±3 123±3 – 164±8	232±7 259±2
	D	115±4 – 146±4	123±4 – 172±4	122±4 – 163±3	58±9	171±3
Italy	C	207±4 383±4	212±5 372±4	212±6 – 234±3 372±4	175±3	160±4
	R	266±7 – 373±5	238±5 – 332±5 374±6	242±3 – 287±9 323±8	305±4 356±4	372±4
	D	23±8 – 31±3 386±4	21±3 374±4	26±4 – 34±4 392±3	18±4	31±3
	C	112±4 – 246±3 346±3	122±2 – 252±8 365±3	123±8 – 262±4 368±3	128±5 112±6 – 362±6	352±3 153±5
	R	223±8 – 274±7 354±8 – 373±4	28±4 – 313±5 341±4 – 384±5	238±4 – 253±2 383±3	267±4 – 374±3 276±4 – 332±8	375±3
	D	134±5 – 234±5 353±4 – 372±6	122±3 – 249±3 361±4	132±2 – 209±3 362±6	149±12	142±9
Mexico	C	102±5 – 177±4 262±2 – 342±5	204±5 236±9 – 339±5	234±1 – 334±3	147±6	154±3
	R	106±3 – 264±6	124±2 – 345±9	133±8 – 373±3	118±3	193±10
	D	63±6 – 215±3 264±3 – 373±3	45±5 – 173±4 288±2 – 380±4	96±4 – 173±6 252±6 – 338±3	56±12	93±4
Colombia	C	124±3 – 158±3 252±3 – 314±1 373±10	129±6 – 172±2 261±3 – 322±5 384±2	123±8 – 184±3 274±2 – 288±4 372±3	128±4	161±9
	R	132±2 – 253±2 265±5 – 312±7	205±4 – 238±6	235±2 – 243±6	148±3	187±16
	D	130±2 – 148±8 272±4 – 342±3 386±3	112±3 – 173±4 312±7 – 352±6 383±2	134±5 – 148±2 292±6 – 352±8 375±3	132±2	284±6
Fiji	C	–	–	–	–	–
	R	–	–	–	–	–
	D	–	–	–	–	–

Appendix I

Datasets Correlation Exploration

In this section and as described in Section 4.3.1 we present the correlation between the confirmed, recovered, and death cases for the SIIMCF Countries.

Italy (Europe): Figure I.1 shows that the CRD datasets explored for Italy in this study are strongly correlated.

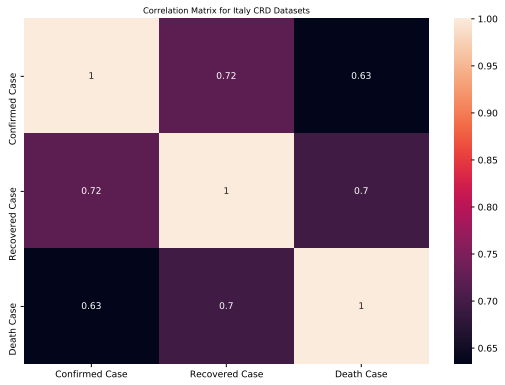


FIGURE I.1: Correlation Matrix for Italy CRD Datasets

India (Asia): As evidenced in Figure I.1, the CRD datasets explored for India in this study are strongly correlated.

Mexico (North America): The CRD datasets explored for Mexico are highly correlated as shown in Figure I.1.

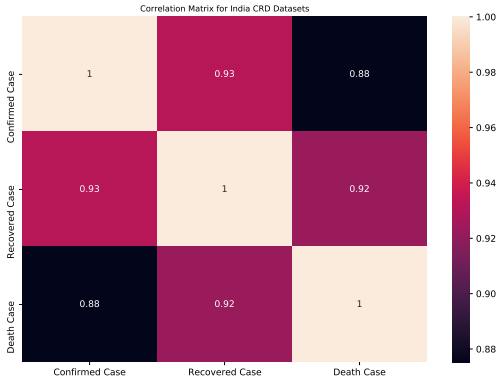
Colombia (South America): We show in Figure I.1 that the CRD datasets explored for Colombia in our study are highly correlated.

Fiji (Oceania): Figure I.1 shows that the CRD datasets explored for Fiji are weakly correlated. This shows that a variable has a weak statistical influence on the other correlated variable in the dataset.

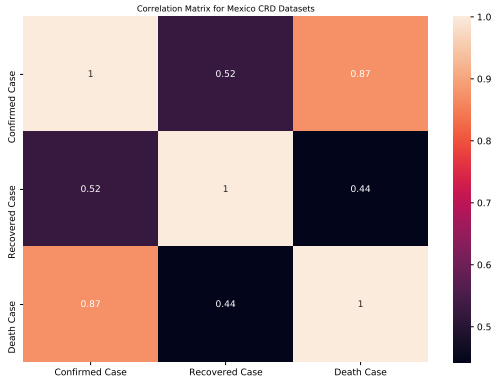
The above explanations are also consistent with the simulated datasets as evidenced in Appendix J

TABLE I.1: Dataset Correlation Plot for the SIIMCF Countries

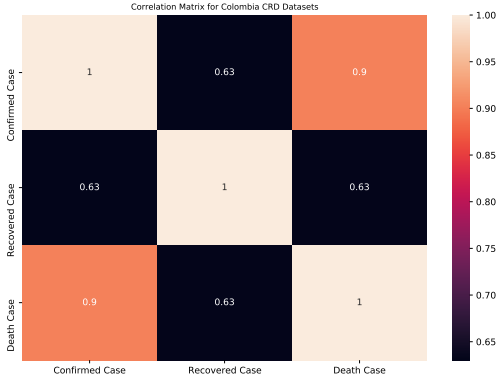
India



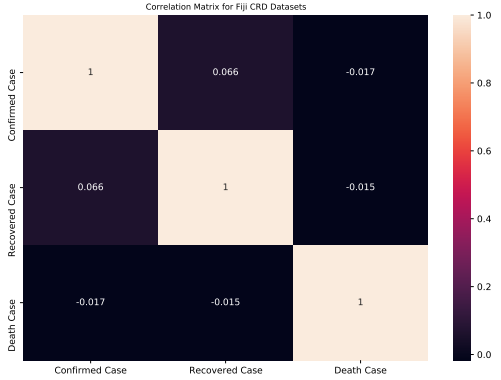
Mexico



Colombia



Fiji

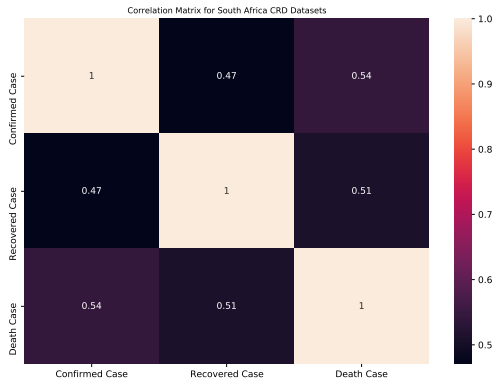


Appendix J

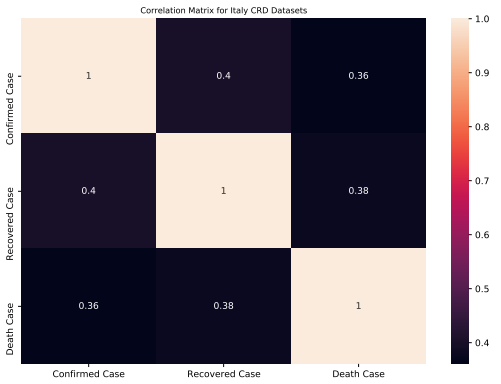
Datasets Correlation Exploration for Simulated Datasets

TABLE J.1: Dataset Correlations Plot for the SIIMCF Countries

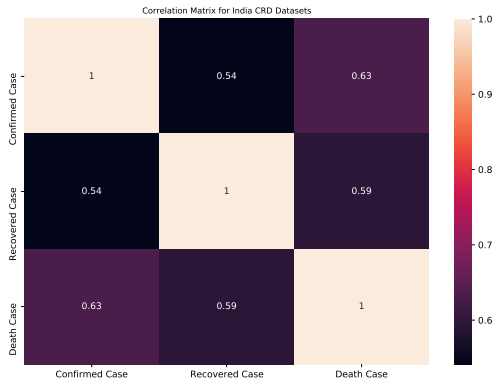
South Africa



Italy



India



Mexico

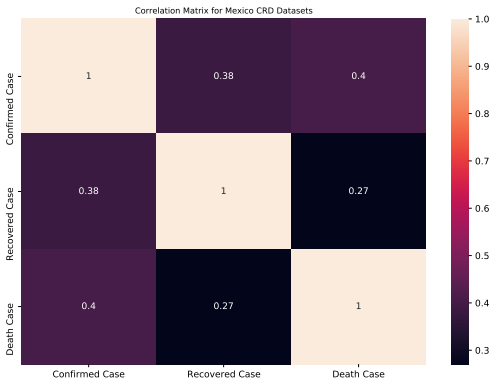
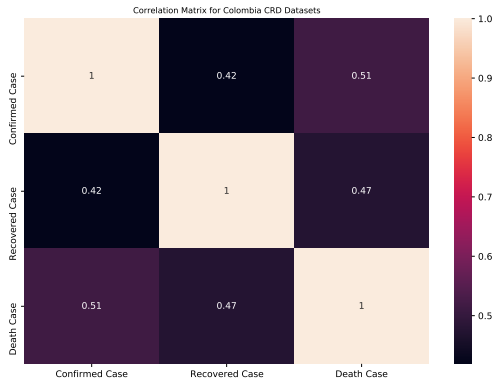
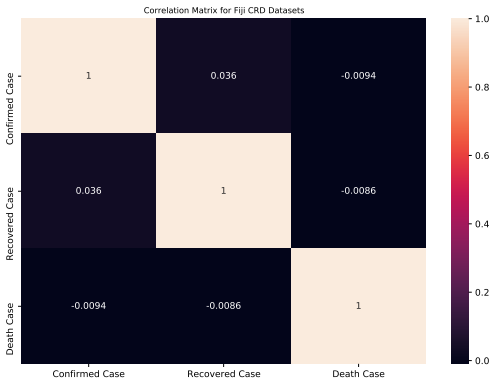


TABLE J.2: Dataset Correlations Plot for the SIIMCF Countries

Colombia	**Fiji**
	

Appendix K

Outlier Exploration of Datasets

The data exploration to check for outliers in the CRD datasets of the countries explored in our study are presented below.

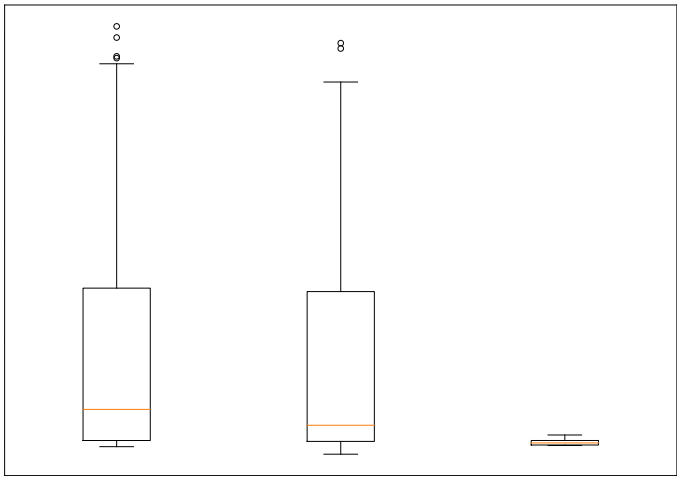
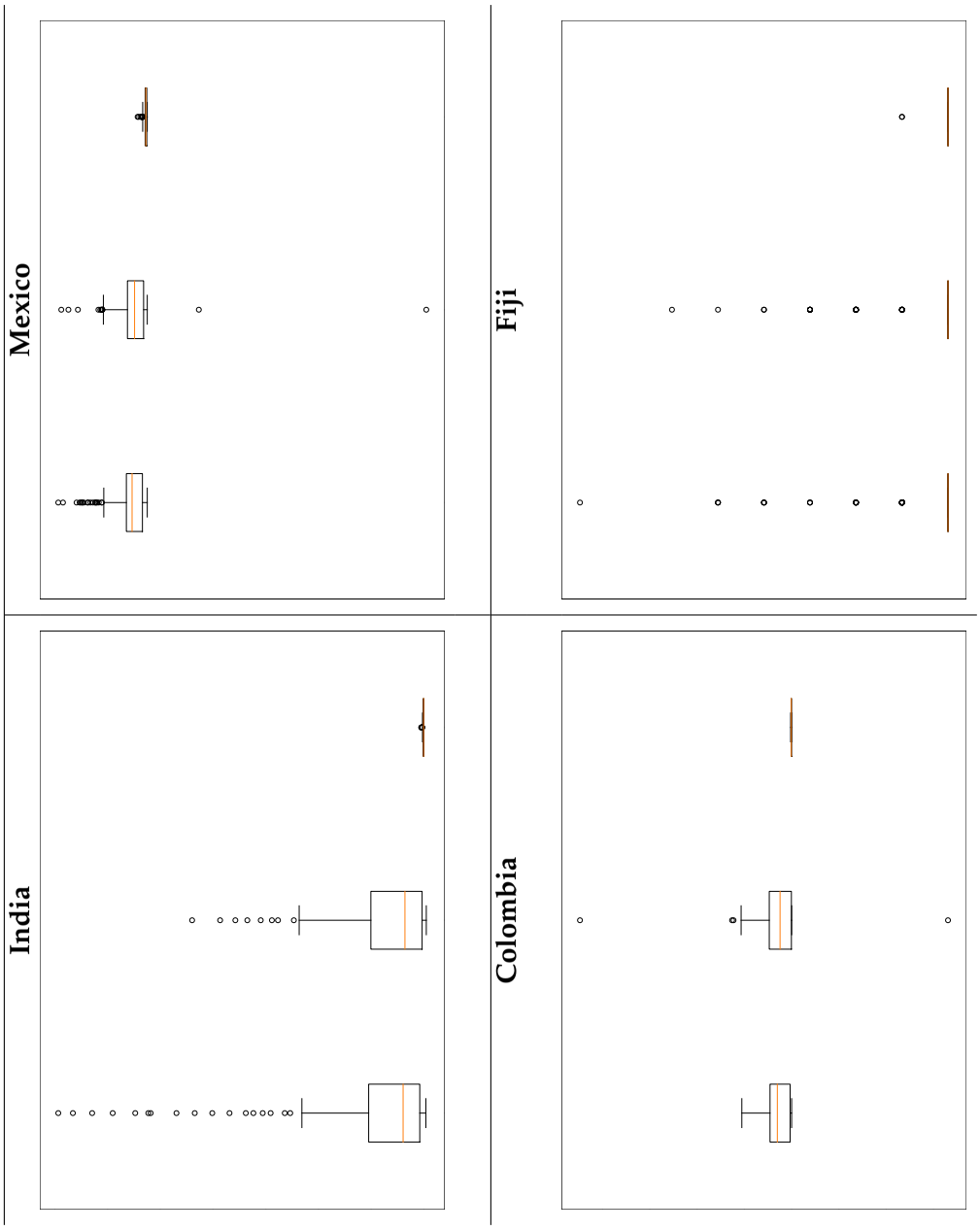


FIGURE K.1: Outlier Detection of Italy CRD Dataset

Observation - Our outlier detection study reveals that the CRD datasets for the SIIMCF countries contains outlier data. We rationalized the datasets and carried out the procedures outlined in Section 4.3.4.


Mexico
Fiji
India
Colombia

Appendix L

Visual Estimation of Datasets (A)

L.1 Real-World Datasets

This section aims to present the visuals for the ACF and PACF representing of the real world CRD datasets for all the countries investigated in this study. The interpretation of the visuals remain the same as described in Section 4.3.2 of Chapter 4.

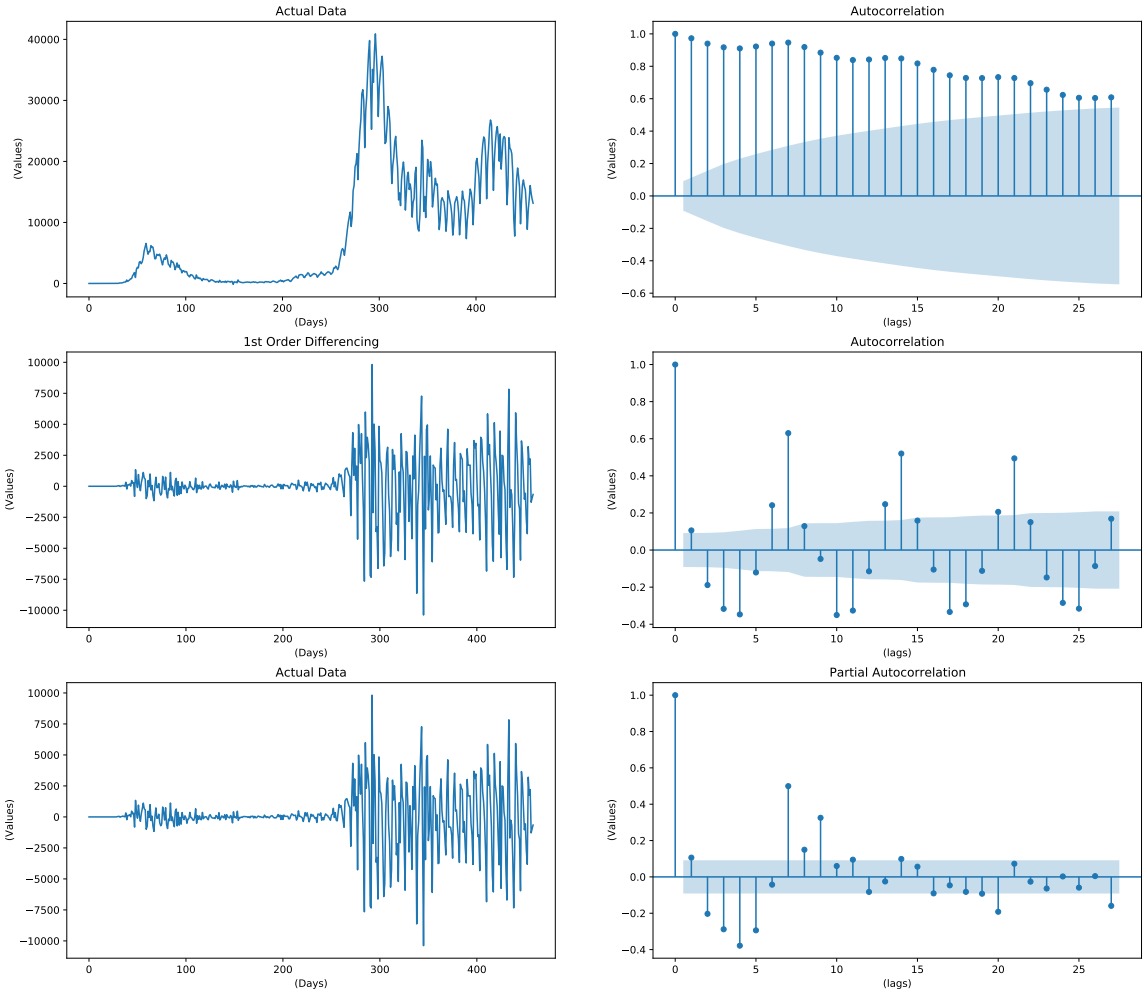


FIGURE L.1: ACF-PACF Plot of Confirmed Case in Italy

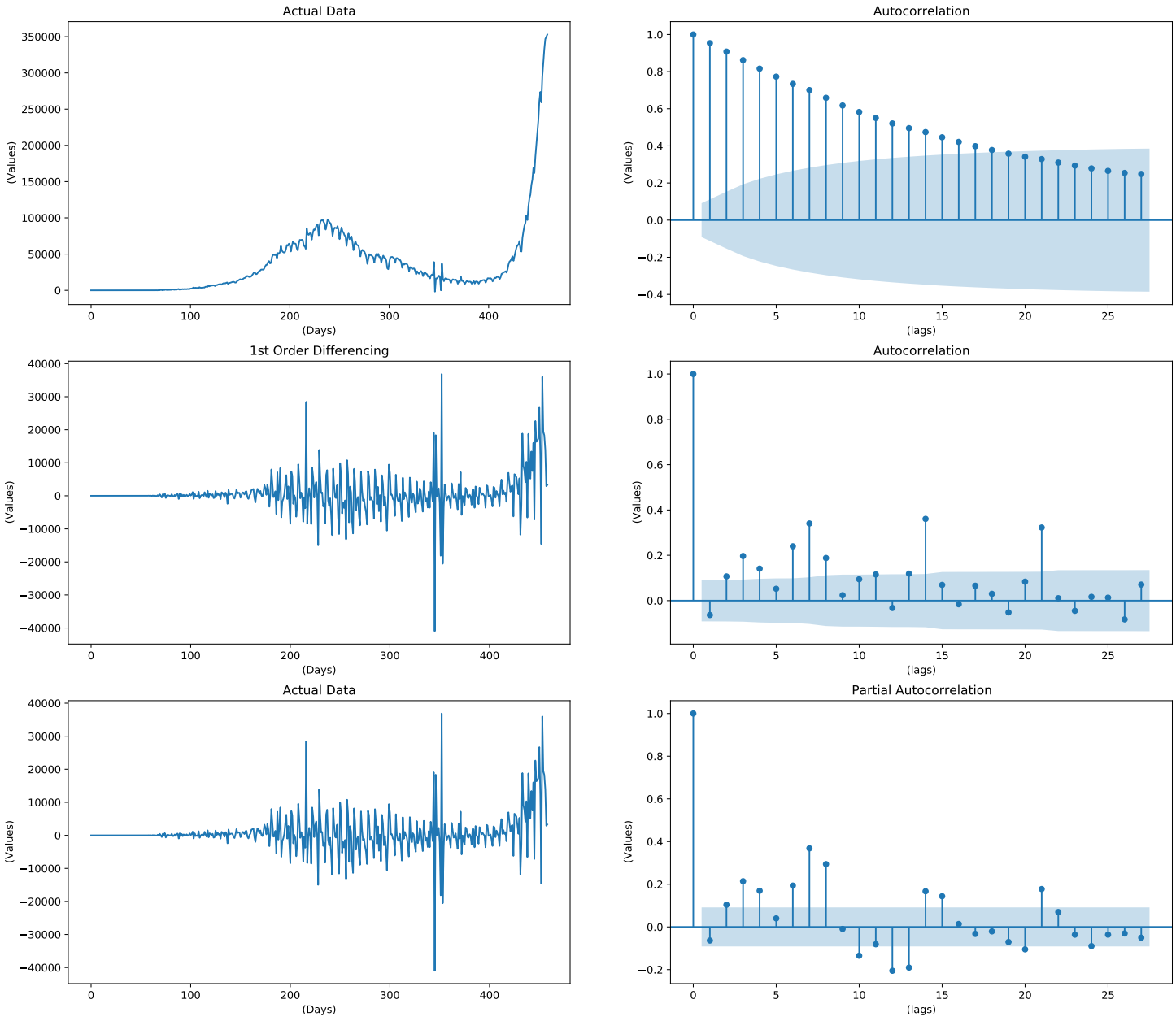


FIGURE L.2: ACF-PACF Plot of Confirmed Case in India

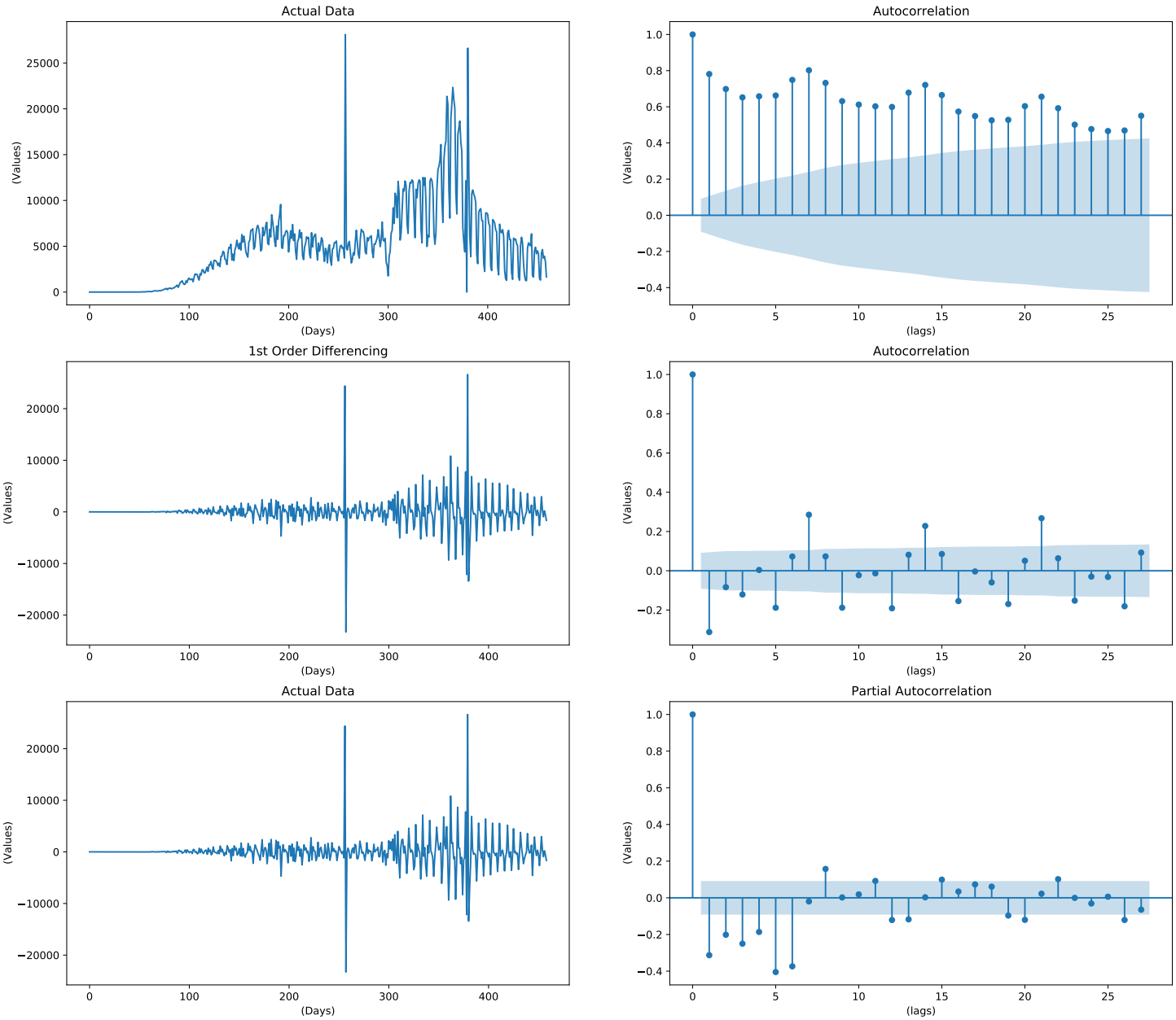


FIGURE L.3: ACF-PACF Plot of Confirmed Case in Mexico

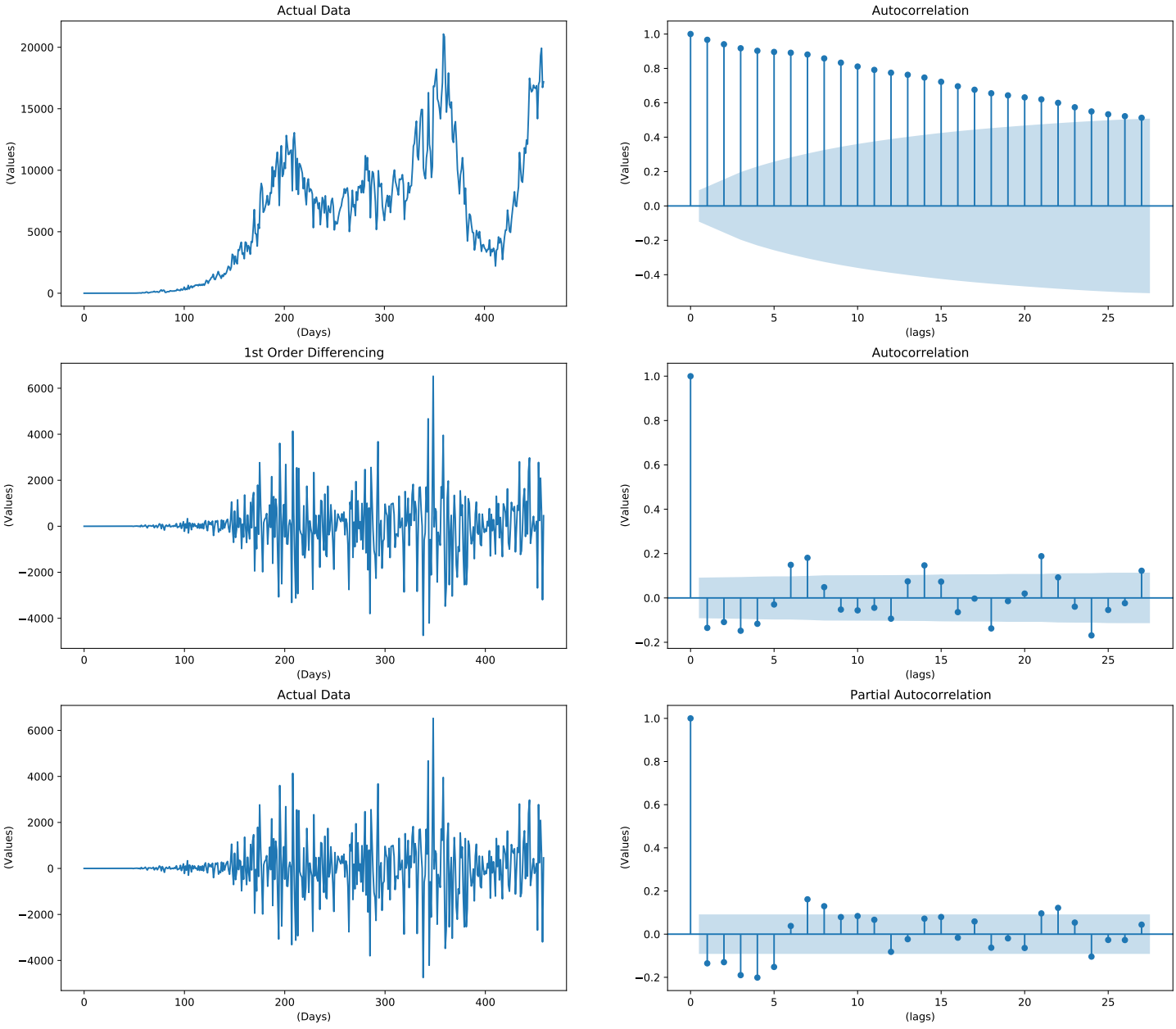


FIGURE L.4: ACF-PACF Plot of Confirmed Case in Colombia

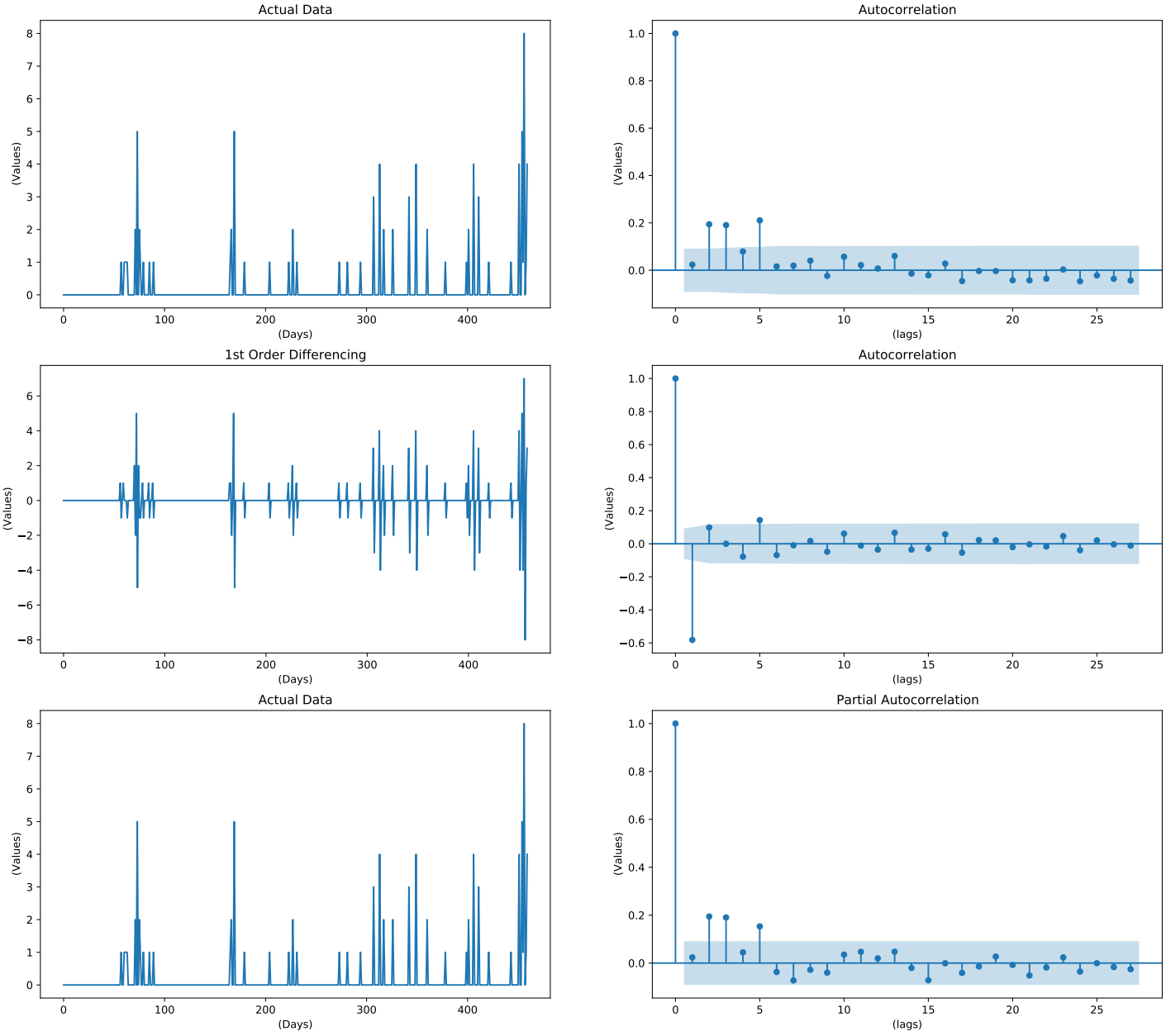


FIGURE L.5: ACF-PACF Plot of Confirmed Case in Fiji

Appendix M

Visual Estimation of Datasets (B)

M.1 Simulated Datasets

This section aims to present the visuals for the ACF and PACF representing of the simulated CRD datasets for all the countries investigated in this study. The interpretation of the visuals remain the same as described in Section 4.3.2 of Chapter 4.

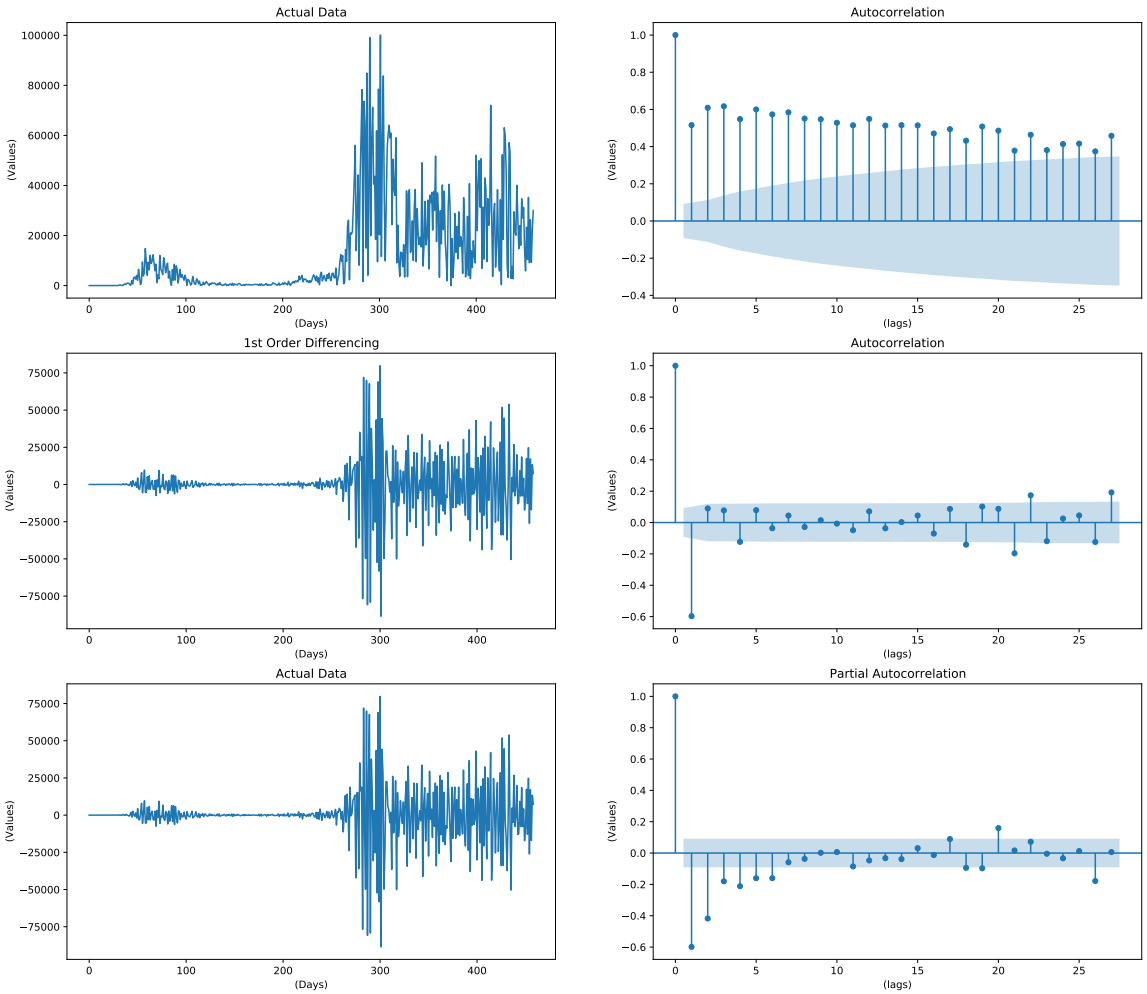


FIGURE M.1: ACF-PACF Plot of Confirmed Case in Italy

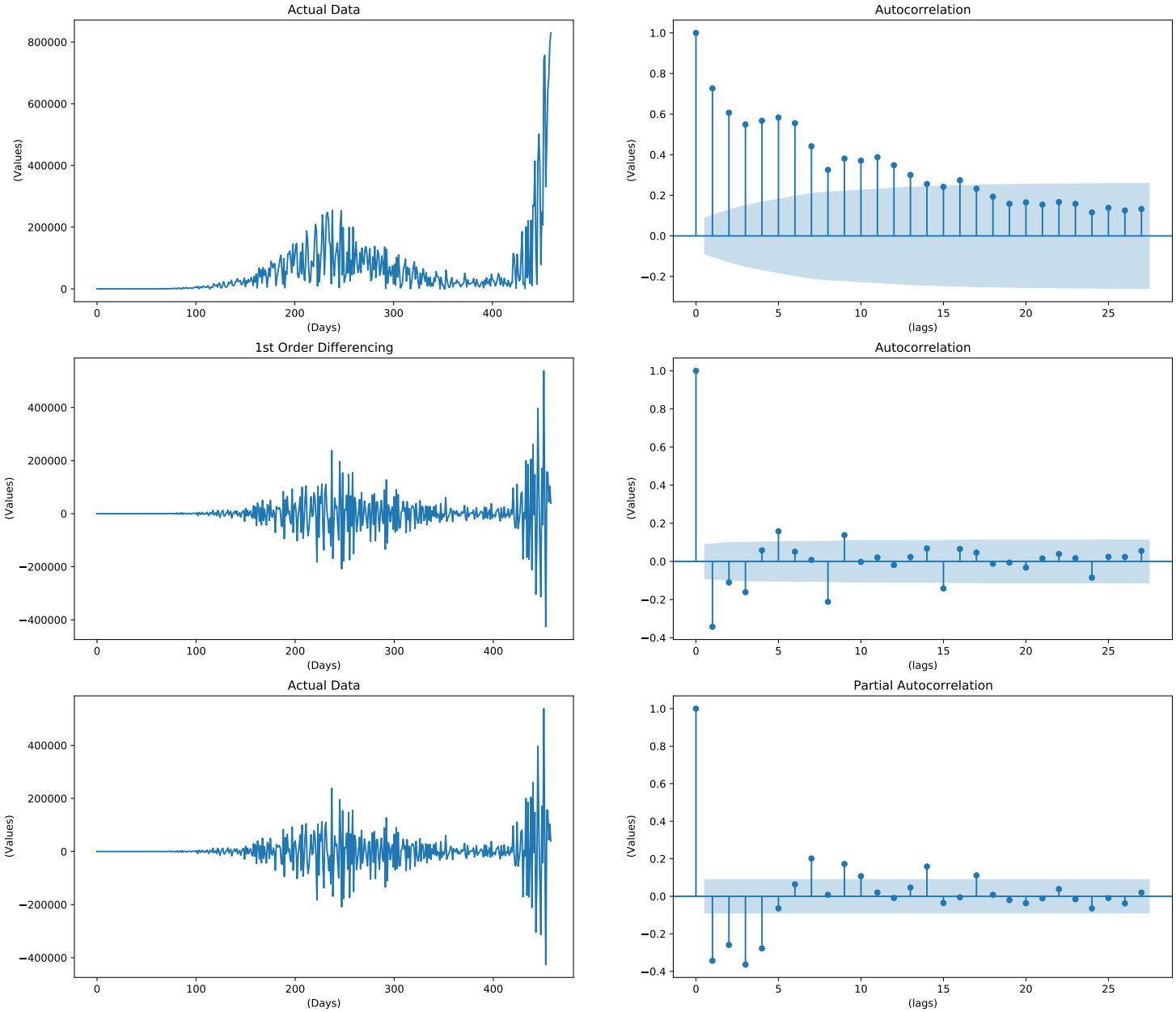


FIGURE M.2: ACF-PACF Plot of Confirmed Case in India

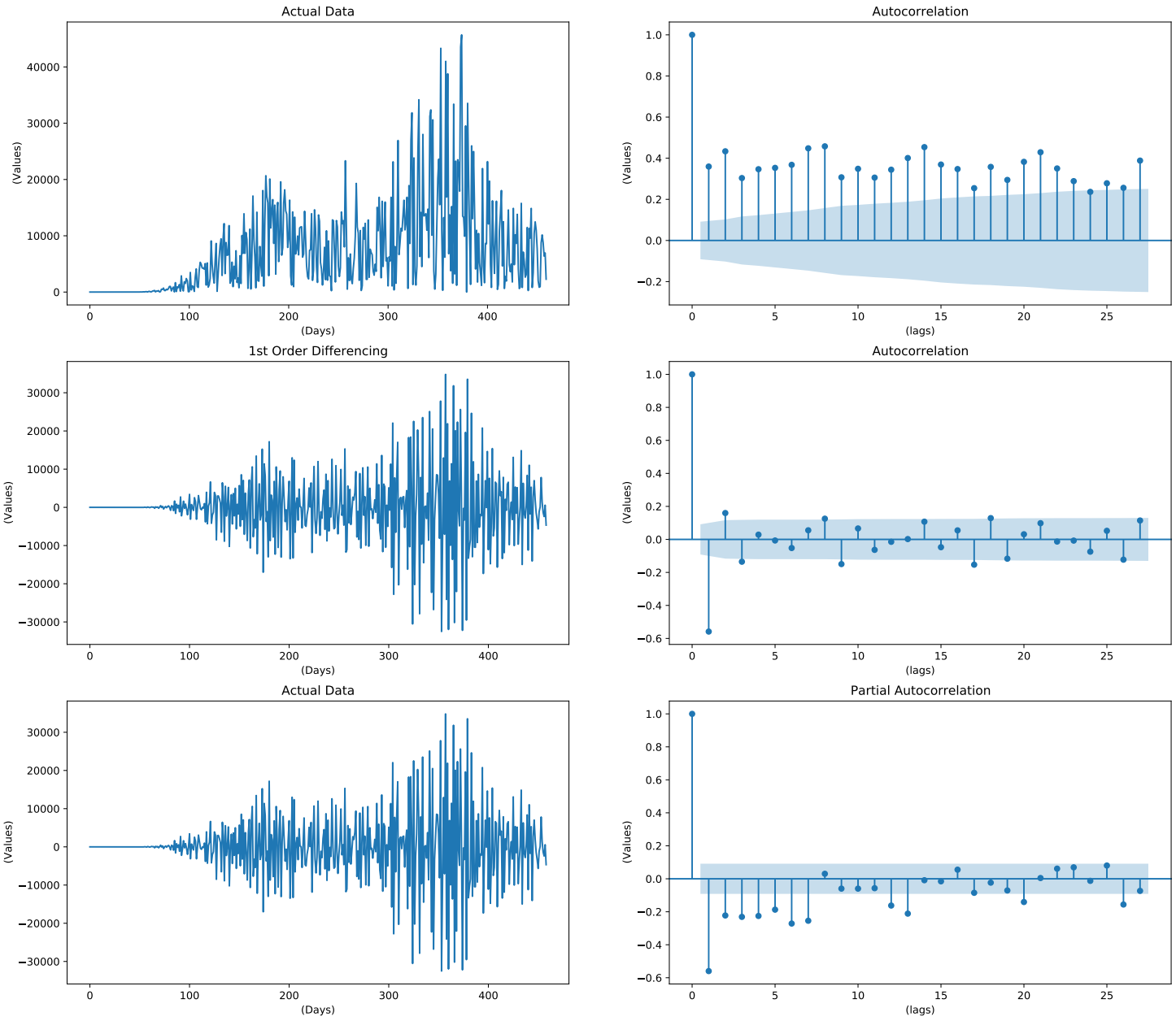


FIGURE M.3: ACF-PACF Plot of Confirmed Case in Mexico

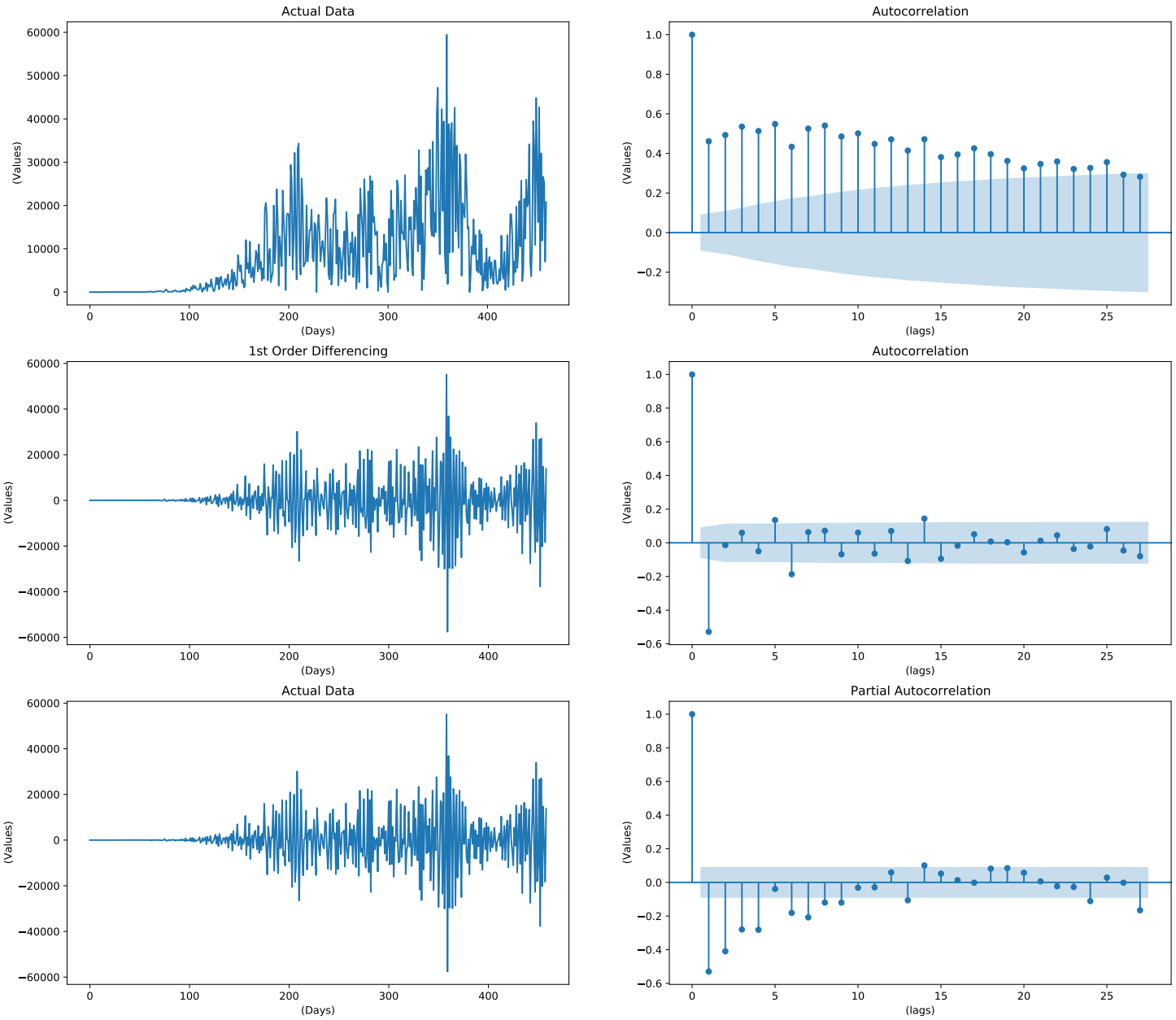


FIGURE M.4: ACF-PACF Plot of Confirmed Case in Colombia

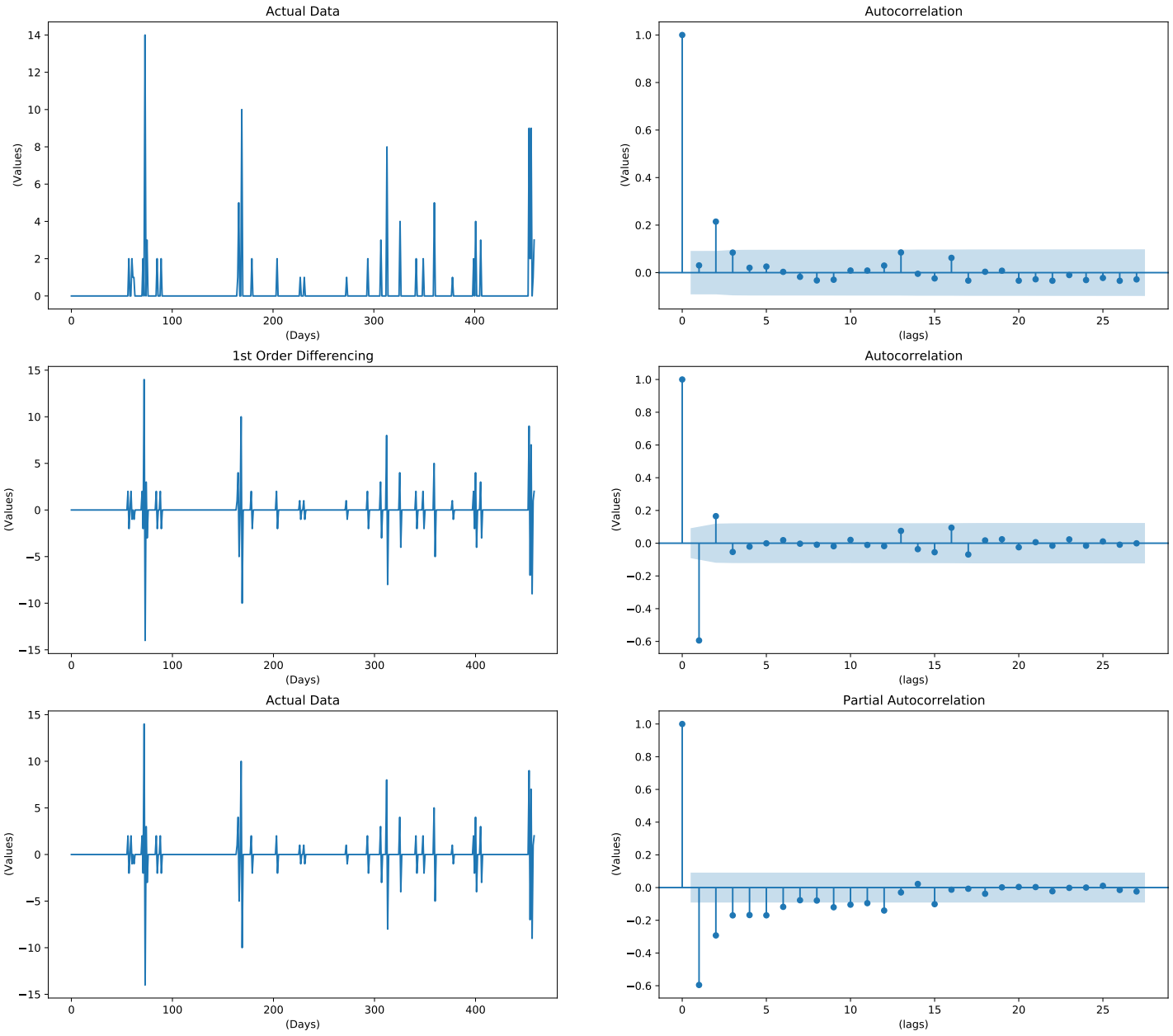


FIGURE M.5: ACF-PACF Plot of Confirmed Case in Fiji

Bibliography

[1] *Who: Statement regarding cluster of pneumonia cases in wuhan, china; 2020*. [Online]. Available: https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china.

[2] *John hopkins university: Covid-19 dataset*, https://github.com/CSSEGISandData/COVID-19, Feb. 2020.

[3] *Web md*, Apr. 2020. [Online]. Available: https://www.webmd.com/cold-and-flu/what-are-epidemics-pandemics-outbreaks.

[4] B. B. Corporation, *Bbc news covid-19: Ftse 100, dow, sp 500 in worst day since 1987*, Mar. 2020. [Online]. Available: http://https//www.bbc.co.uk/news/business-51829852.

[5] S. A. NICD, *National institute for communicable diseases in south africa*, Feb. 2020. [Online]. Available: https://www.nicd.ac.za/.

[6] *Africa centres for disease control and prevention*, Mar. 2020. [Online]. Available: https://africacdc.org/.

[7] S. A. Ministry of Health, *Department of health in south africa*, Mar. 2020. [Online]. Available: https://sacoronavirus.co.za/.

[8] W. H. Organization, *World health organization (who) for south africa region*, Mar. 2020. [Online]. Available: https://www.afro.who.int/.

[9] E. Gary, M. Weissman, A. C. Droesch, and C. Chivers, "Locally informed simulation to predict hospital capacity needs during the covid-19 pandemic," *MedRxiv - Annals of Internal Medicine*, vol. 173, no. 1, pp. 21–28, July 2020.

[10] MOHFW, *India ministry of health: Description of covid-19 in india*. [Online]. Available: https://www.mohfw.gov.in.

[11] OECG, *Organisation for economic co-operation and development*. [Online]. Available: http://www.oecd.org/.

[12] Z. Ceylan, "Estimation of covid-19 prevalence in italy, spain, and france," *Sci Total Environ*, vol. 729, p. 138 817, July 2020.

[13] A. Ahmar and E. del Val, "Sutte arima: Short-term forecasting method, a case: Covid-19 and stock market in spain," *Sci Total Environ*, vol. 729, p. 138 883, July 2020.

[14] S. R. Conway, "An agent-based model for analyzing control policies and the dynamic service-time performance of a capacity-constrained air traffic management facility," *25th Congress of International Council of the Aeronautical Sciences*, August 2006.

[15] V. A. Folcik, C. Gary, and C. G. Orosz, "The basic immune simulator: An agent-based model to study the interactions between innate and adaptive immunity," *Theoretical Biology and Medical Modelling*, vol. 4, no. 39, June 2007.

[16] N. Metawa, K. Hassan, and M. Elhoseny, "Genetic algorithm based model for optimizing bank lending decisions," *Expert Systems with Applications*, vol. 80, pp. 215–301, September 2017, ISSN: 957-4174.

[17] A. Wright, K. D. Jong, M. Vose, and L. Schmitt, *Foundations of Genetic Algorithms*. Newyork, United States of America: Springer Press, January 2005, pp. 39–56.

[18] D. M. Aleman, T. G. Wibisono, and B. Schwartz, "A non-homogeneous agent based simulation approach to modeling the spread of disease in a pandemic outbreak," *Interfaces Journal*, vol. 41, no. 3, pp. 215–301, May 2011.

[19] S. Bisgaard and M. Kulahci, *Time Series Analysis and Forecasting by Example*. John Willey and Sons. Inc., 2011.

[20] B. Douglas and L. Bernard, "A framework for rationing ventilators and critical care beds during the covid-19 pandemic," *Jama*, vol. 323, no. 18, pp. 1773–1774, May 2020.

[21] A. Cleo, L. Russo, A. Tsakris, and C. Siettos, "Data-based analysis, modelling and forecasting of the covid-19 outbreak," *Plus One Journal*, March 2020.

[22] P. Kumar, R. Chimmula, and L. Zhang, "Time series forecasting of covid-19 transmission in canada using lstm networks," vol. 135, p. 109 864, June 2020, ISSN: 960-0779.

[23] Scopus, *Database for researchers*. [Online]. Available: http://scopus.com.

[24] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, and M. Ciccozzi, "Application of the arima model on the covid-2019 epidemic dataset," *Proc. IEEE Congr. Evol. Comput. (CEC), Edingburg, UK*, p. 105 340, July 2020.

[25] M. Ahmed and T. Alkhamis, "Simulation optimization for an emergency department healthcare unit in kuwait," *European Journal of Operation Research*, vol. 198, no. 3, pp. 936–942, November 2009.

[26] J. Y. Yeh and W. S. Lin, "Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department," *Expert Systems with Application*, vol. 32, no. 4, pp. 1073–1083, May 2007.

[27] Y.Xu, J. Weaver, D. Healy, and J.Lu, "Wavelet transform domain filters: A spatially selective noise filtration technique," *IEEE Trans Image Process*, vol. 3, no. 6, pp. 747–758, November 1994.

[28] A. Bertozzi, E. Franco, G. Mohler, M. B. Short, and D. Sledge, "The challenges of modeling and forecasting the spread of covid-19," *National Academy of Science of the United States of America*, vol. 117, no. 29, July 2020.

[29] M. V. D. Schaar, "Adjutorium: Using covid-19 patient data to train machine learning models for healthcare," *Journal of Health Informatics: University of Cambridge*, April 2020.

[30] S. Emek, V. Evren, and S. Bora, "An agent-based approach in homeostatic control systems: Thermoregulation," *9th International Conference: Computational Intelligence and Communication Networks (CICN)*, pp. 113–116, July 2017.

[31] P. Rylan and van Zyl Terence L, "Surrogate assisted methods for the parameterisation of agent-based models," in *7th International Conference on Soft Computing Machine Intelligence (ISCMI)*, November 2020, pp. 78–82.

[32] F. Schweitzer, *Brownian Agents and Active Particles. On the Emergence of Complex Behavior in the Natural and Social Sciences*. Berlin: Springer publisher, 2003.

[33] A. Griffin, F. S. Arthur, and Charles, "An agent-based model of prehistoric settlement patterns and political consolidation in the lake titicaca basin of peru and bolivia - structure and dynamics," *eJournal of Anthropological and Related Sciences*, vol. 2, no. 2, Sep. 2007.

[34] A. Troisi, V. Wong, and M. Ratner, "An agent-based approach for modeling molecular self-organization," *Theoretical Biology and Medical Modelling*, vol. 102, no. 02, pp. 255–260, Jan. 2005.

[35] K. Carley, D. Fridsma, E.Casman, A.Yahja, N.Altman, L. Chen, B.Kaminsky, and D.Nave, "Biowar: Scalable agent-based model of bio-attacks," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 36, no. 2, pp. 252–265, Jul. 2006.

[36] E. Mast, G. V. Kuik, and G. V. Bussel, "Agent-based modelling for scenario development of offshore wind energy," in *Proceedings of the 2007 European Wind Energy Conference Exhibition in Milan*, T Chaviaropoulos, Ed., EWEA, pp. 1–4.

[37] K. H. van Dam, Z. Lukszo, L. Ferreira, and A. Sirikijpanichkul, "Planning the location of intermodal freight hubs: An agent based approach," *Proceedings of the 2007 IEEE International Conference on Networking, Sensing and Control*, pp. 187–192, 2007.

[38] A. C. Charania, J. R. Olds, and D. DePasquale, "Sub-orbital space tourism market: Predictions of the future marketplace using agent-based modeling," *Space Works Engineering, Inc., Atlanta, GA*, Oct. 2006.

[39] K. Mock and J. W. Testa, "An agent-based model of predator-prey relationships between transient killer whales and other marine mammal," *University of Alaska Anchorage, Anchorage*, May 2007.

[40] N. Malleson, "Using simulation to predict prospective burglary rates in leeds and vancouver," *7th National Crime Mapping Conference*, May 2009.

[41] J. Holland, "Genetic algorithms and adaptation in adaptive control of ill-defined systems," *Springer Journal, Boston, MA.*, vol. 16, pp. 215–301, May 1964.

[42] R. M. Fujimoto, "Parallel discrete event simulation," *Communications of the ACM*, vol. 33, no. 10, pp. 30–53, 1990.

[43] M. Gunal and J. Pidd, "Discrete event simulation for performance modelling in health care: A review of the literature," *Journal of Discrete Event Simulation*, vol. 107, pp. 42–51, 2017.

[44] M. Gunal and J. Pidd, *Discrete event simulation for performance modelling in health care: A review of the literature*, 2017.

[45] A. E. Hassanien and R. M. Rizk-Allah, "Covid-19 forecasting based on an improved interior search algorithm and multi-layer feed forward neural network," *Neural and Evolutionary Computing*, April 2020.

[46] R. Balling, "The maximin fitness function; multi-objective city and regional planning," *International Conference on Evolutionary Multi-Criterion Optimization (EMO 2003)*, pp. 1–15, April 2003.

[47] R. Sharapov and A. Lapshin, "Convergence of genetic algorithms," *Mathematical Theory of Pattern Recognition*, no. 16, 392–397, July 2006.

[48] S. Sivanandam, *Principles of soft computing*. New Delhi: Wiley Publishers, 2013, ISBN: 978-1-118-54680-2.

[49] E. Alba and M. Tomassini, "Parallelism and evolutionary algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 5, 443–462, November 2002.

[50] M. Dubreuil, C. Gagne, and M. Parizeau, "Analysis of a master-slave architecture for distributed evolutionary computations," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 1, 229–235, March 2006.

[51] J. H. Jones, *Evaluation of reproduction number in epidemiology simulation*, May 2007. DOI: https://web.stanford.edu/~jhj1/teachingdocs/Jones-on-R0.pdf.

[52] R. Ross, A. McKendrick, W. Hamer, and W. Kermack, "Proceedings of the royal society of london. series a, containing papers of a mathematical and physical character," *Medical mathematics journal*, vol. 115, no. 772, pp. 700–721, 1927.

[53] K. M. Sutton, "Discretizing the si epidemic model," *Rose-Hulman Undergraduate Mathematics Journal*, vol. 15, no. 12, January 2014.

[54] O. Diekmann, H.Heesterbeek, and T. Britton, "Mathematical tools for understanding infectious disease dynamics," *Princeton Series in Theoretical and Computational Biology. Princeton, NJ: Princeton University Press*, pp. 568–624, 2013.

[55] D.Greenhalgh, "Hopf bifurcation in epidemic models with a latent period and non-permanent immunity," *Mathematical and Computer Modelling*, vol. 25, no. 2, pp. 85–107, May 1998.

[56] H. Hethcote and S. Levin, "Periodicity in epidemiological models," *Applied Mathematical Ecology and Bio-mathematics journal*, vol. 18, pp. 192–211, 1989.

[57] B. Edwin and J. Worcester, *The law of mass action in epidemiology*, December 1944. [Online]. Available: https://www.pnas.org/content/31/1/24.

[58] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Communications of the ACM*, vol. 349, no. 6245, pp. 255–260, July 2015.

[59] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT press, 2016.

[60] M. Aiken, "A neural network to predict civilian unemployment rates," *Journal of International Information Management*, vol. 5, no. 1, 1996.

[61] H. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (indrnn): Building a longer and deeper rnn," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5457–5466, Jun. 2018.

[62] A. Nair, Vinod, Hinton, and E. Geoffrey, "Rectified linear units improve restricted boltzmann machines," *27th International Conference on International Conference on Machine Learning*, 807–814, April 2020.

[63] R. Gove and J. Faytong, "Machine learning and event-based software testing: Classifiers for identifying infeasible gui event sequences," *Advances in Computers*, p. 109, 2012.

[64] L. Cao, K. Chua, W. Chong, H. Lee, and Q. Gu, "A comparison of pca, kpca and ica for dimensionality reduction in support vector machine," *Neuro Computing*, pp. 321–336, 2003.

[65] Z. Li, S. Yang, and J. Wu, "The prediction of the spread of covid-19 using regression models," November 2020, pp. 25–28.

[66] B. Suvarna, L. Nandipati, and N. Bhat, "Support vector regression for predicting covid-19 cases," *European Journal of Molecular Clinical Medicine*, vol. 7, no. 3, pp. 4882–4893, 2020.

[67] N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-Shishiny, "An empirical comparison of machine learning models for time series forecasting," *Econometric Reviews*, vol. 29, no. 5, p. 594, 2010.

[68] L. Breiman, "Random forest machine learning technique," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[69] A. P. Engelbrecht, *Computational Intelligence: An Introduction*. 2007.

[70] N. Khong, Rao, and M. Mariappan, "Blood pressure measurements using non-contact video imaging techniques," *2nd IEEE International Conference on Automatic Control and Intelligent Systems*, vol. 2017, pp. 35–40, October 2017.

[71] P. Krause, D. P. Boyle, and F. Base, "Comparison of different efficiency criteria for hydrological model assessment," *Advances in Geosciences*, vol. 5, pp. 89–97, 2005.

[72] O. Balci, "Verification, validation and testing of computer models)," *Simulation Handbook: Principles, Methodology, Advances, Applications, and Practice, John Wiley & Sons, New York NY*, pp. 335–393, 1998.

[73] A. Myttenaerea, B. Golden, B. Grand, and F. Rossi,

[74] H. Nishiura, N. Linton, and A. Akhmetzhanova,

[75] Patro, S. Gopal, Sahu, and K. Kumar, "Normalization: A preprocessing stage," *IARJSET Journal*, Mar. 2015.

[76] S. Bisgaard and M. Kulahci, *Time Series Analysis and Forecasting by Example*. John Willey and Sons. Inc., 2011.

[77] R. J. Hyndman and G. Athanasopoulos, *Forecasting principles and practice*. Maryland, USA: OTexts, 2018.

[78] R. A. Yaffee and M. M. Gee, *An Introduction to Time Series Analysis and Forecasting*. Prints Media, 2010.

[79] E. J. W. E.E. Holmes M. D. Scheuerell, *Applied time series analysis for fisheries and environmental sciences*. [Online]. Available: https://nwfsc-timeseries.github.io/atsa-labs/sec-boxjenkins-aug-dickey-fuller.html.

[80] M. Cristina, M.-L. Elisa, M. Joaquín, and D. Petar, "A particle filtering scheme for processing time series corrupted by outliers," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4611–4627, 2012.

[81] *Overview of covid-19 in south africa*, Mar. 2020. [Online]. Available: https://www.covid19sa.org.

[82] *Provincial breakdown of covid-19 in south africa*, Mar. 2020. [Online]. Available: https://www.covid19sa.org/provincial-breakdown.

[83] S. Grima, M. Kizilkaya, R. Rupeika-Apoga, I. Romanova, R. D. Gonzi, and M. Jakovljevic, "A country pandemic risk exposure measurement model," *Risk Manag Health Policy*, vol. 13, 2067—2077, July 2020.

[84] R. Breban, J. Riou, and A. Fontanet, "Inter-human transmissibility of middle east respiratory syndrome coronavirus: Estimation of pandemic risk," *The Lancet Journal*, vol. 382, pp. 694–699, 9893 August 2013.

[85] *John hopkins university: Covid-19 dashboard*, https://coronavirus.jhu.edu/map.html, Mar. 2020.

[86] J. S. Arora, *Introduction to Optimum Design (Fourth Edition) - Nature-Inspired Search Methods*. The University of Iowa College of Engineering at Iowa City, Iowa: Academic Press, 2007, pp. 739–769, ISBN: 978-0-12-800806-5.