Kwame Nkrumah University of Science and Technology
College of Engineering
Department of Telecommunication Engineering

Machine Learning for Time Series Analysis of Epidemics: Case of COVID-19

A thesis submitted in partial fulfilment for a Bachelor of Science Degree in Telecommunication Engineering

By

Njini Nathan Fofeyin
Acquah Oliver
Oteng-Oppey Emmanuella
Supervisor: Prof. John Jerry Kponyo

September 2022

Declaration

We hereby state that this submission is our original work for the Bsc. Telecommunication Engineering degree and that, to the best of our knowledge, it does not contain any previously published work by another person or work that has been accepted for the award of any other degree from the university, unless appropriate citation has been made in the text.

Njini Nathan Fofeyin (3574018)	.....
(Student)	Signature
Acquah Oliver (3565318)	.....
(Student)	Signature
Oteng-Oppey Emmanuella (3572118)	.....
(Student)	Signature

Supervised by:

Prof. Ing. Jerry John Kponyo	.....
(Supervisor)	Signature

Dedication

This work is dedicated to Prof. Jerry John Kponyo, our supervisor, who has served as a mentor and has helped us in every way throughout the course of this project. It would not have been possible to complete this work without his continuous support and direction.

Acknowledgment

We are grateful to God Almighty for providing us with the fortitude, knowledge, wisdom, and strength necessary to take on this task and see it through to completion. We are appreciative of our families for helping us have a comfortable four years living on campus. Even during the most trying times, we were able to study comfortably thanks to their sacrifices.

We sincerely appreciate everything our lecturers have taught us about commitment. They have given us all the skills necessary to complete this project through their unwavering dedication to teaching and their unwavering commitment to excellence.

We would also like to express our gratitude to the Department of Telecommunication Engineering at KNUST for providing us with a top-notch engineering education and preparing us for the future. Last but not least, we appreciate our friends for helping to make our time at KNUST special.

Contents

1 Introduction **14**

1.1 Background 14

1.2 Problem Statement 16

1.3 Research Objectives 17

1.3.1 General Objectives 17

1.3.2 Specific Objectives 17

1.4 Significance of Studies 17

1.5 Organization of the report 18

2 Literature Review **19**

2.1 Introduction 19

2.2 Synopsis of Compartmental models 20

2.3 Regressive and Curve Fitting Models 24

2.3.1 Applications of Regressive techniques to Epidemics 24

2.3.2 Regression Analysis 26

2.3.3 Exponential Smoothing 28

2.3.4 Autoregressive Integrated Moving Average 29

2.4 Network Models 32

2.5 Machine Learning Models 32

2.5.1 Introduction to ML/DL and Applications to Epidemics 32

2.5.2 Review of Supervised Learning for Epidemics 35

2.5.3 Unsupervised Learning for Modeling COVID-19 Spread 38

2.5.4 Forecasting with Facebook Prophet 40
2.5.5 Deep Learning Models for Forecasting 41
2.5.6 Hybrid Models: Neural Prophet 43

3 Research Methodology **44**
3.1 Implementation Frameworks and Tools 44
3.2 Datasets and Preprocessing 44
3.3 Python Implementation of Facebook Prophet 45
3.4 LSTM Neural Network Architecture 46
3.4.1 Keras Implementation of LSTM 46
3.5 Implementation of Neural Prophet 48
3.6 Performance Metrics for Prediction and Forecasting 49

4 Results and Discussion **50**
4.1 Prophet Fitting and Forecasts 50
4.2 Time Sequence LSTM Forecasting 51
4.3 Neural Prophet Prediction and Forecasting 52
4.4 Performance of Models for a 90-day Forecast 53
4.5 Regional Level Fitting and Forecasting 54
4.5.1 Approach 1 54
4.5.2 Approach 2 55
4.6 Use Case: Data-Driven Decision Making 57

5 Conclusion and Recommendations **58**
5.1 Conclusion 58
5.2 Recommendations 59
5.3 Future Work and Open Problems 59

List of Figures

1.1 Coronavirus Structure, [Source: CDC] 14

1.2 Chart of Covid-19 distribution in Ghana as of June 2022 16

2.1 Demonstration of Compartmental Models 22

2.2 Inefficiency of SIR in dynamic spread. Source: Moein et al. 24

2.3 Underfitted Linear Regression on Ashanti Regional Data 27

2.4 Order Selection and Polynomial Fitting 28

2.5 Simple Exponential Smoothing Fit for Ghana's National COVID-19 data 29

2.6 Autocorrelation and Partial Autocorrelation Plots 31

2.7 Net-logo approach to simulate spread in a community 32

2.8 Applications of AI/ML in Healthcare 34

2.9 General Workflow Diagram for Supervised Learning 37

2.10 Use of agglomerative and divisive functions on a dataset of five items, a, b, c, d, e. (b) Using partition clustering on a dataset with fourteen clusters 39

2.11 How a Multi-layer/Deep Neural Network Works! 42

2.12 GRU vs RNN vs LSTM Unit Design 42

3.1 National Data Preprocessing 45

3.2 Regional Level Data Import and Preprocessing 45

3.3 Single LSTM Unit 47

3.4 LSTM Architecture, implemented for $t = 90, 150, 300,$ 47

3.5 Minimized Loss for 360 epochs of training 48

4.1 National COVID-19 Daily Forecasting with Prophet 51

4.2 National Cumulative COVID-19 Forecasting with Prophet Models 51
4.3 Using LSTM for COVID-19, National Level Forecasting 52
4.4 Estimation of Future COVID-19 Cases for Western Region (Neural Prophet) . . 53
4.5 National Cumulative COVID-19 Forecast using Neural Prophet 53
4.6 Comparison of model performance for a 90-day forecast 54
4.7 Approach 1 to Estimating COVID-19 spread in Regions 55
4.8 Approach 2: Forecasting Ghana's Regional Daily Cases Independently 56
4.9 Approach 2: Forecasting Ghana's Regional Cumulative Cases Independently . . 56
4.10 A Simple COVID-19 Dashboard for Ghana and its Regions 57

List of Tables

3.1 Training of Neural prophet and loss minimization for 212 epochs 49

4.1 Comparison of model performance for a 90-day forecast 53

List of Abbreviations

KNUST	Kwame Nkrumah University of Science and Technology
RNA	Ribonucleic Acid
WHO	World Health Organization
SARS-COVID 2	Severe Acute Sespiratory Syndrome Coronavirus 2.
MEMS	Mathematical Epidemic Models
AR	Autoregression
MA	Moving Average
ARIMA	Autoregressive Moving average
SIR	Susceptible-Infected-Recovered
MAPE	Minimum Absolute Percentage Error
RMSE	Root Mean Squared Error
SVM	Support Vector Machines
SVR	Support Vector Regression
GPR	Gaussian Process Regression
EEMD	Ensemble empirical mode decomposition
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
LSTM	Long-Short Term Memory
GRU	Gated Recurrent Unit
GAN-GRU	Generative Adversarial Network GRU
CAN	Susceptible-Infected-Recovered
VAN	Variational Auto-Encoder
MLP	Multi-layer perceptron
AR-Net	Auto-regressive Neural Network

List of Algorithms

1 ARIMA Model for Forecasting 31

Abstract

The outbreak of COVID-19 in the world has resulted in major changes to society including lockdown, volatile economy, overwhelming healthcare systems, loss of jobs and layoffs, and other socio-economic changes. The world was thrown into a state of pandemonium since the emergence of the novel coronavirus (covid-19) in December 2019. Over 500 million cases of covid-19 have been recorded with about 6 million deaths. This case is no different in Ghana, with the country plummeting into economic crisis and setbacks in the education and healthcare system, work and lifestyle as a result of this pandemic. About 165,000 cases have been recorded in Ghana as of July 2022 with 1453 death cases. Forecasting future repercussions and estimating the future number of cases is essential for handling the virus and limiting its effects on society. Scientists have attempted to forecast the future trend of COVID-19 in different countries and regions. Unfortunately, very few such attempts have been made in Ghana and only a handful (2 papers) are publicly available. There is a need to model and forecast the pandemic's spread in Ghana.

In the world, techniques such as compartmental, statistical and machine learning models and deep learning models have been used for modelling the pandemic and forecasting its spread. Research has shown that machine learning-based models and statistical models are effective in providing clear-cut forecasts that are more useful and accurate. As such, our research focuses on designing 5 statistical and machine learning models for fitting current COVID-19 data in Ghana and using the model to forecast the future trend of expected cases. With such clear–cut information regarding the disease, authorities are well-informed to make appropriate decisions concerning resource allocation and population control to minimize the spread. In this study, we tested the performance of some common statistical and machine learning models to predict

epidemics. The project also provides baseline models that will be made publicly available for future researchers to quickly use when modelling and forecasting the trend of other epidemics.

The authors have discussed and implemented several machine learning and statistical models including polynomial regression, long short-term memory neural networks, Facebook prophet and neural prophet (hybrid of neural networks and prophet model). The results from these models are analyzed based on demographic information and ground truths about the population. Based on the analysis, one of these models is then recommended for use based on performance and convenience metrics. An experiment on transfer learning is performed to understand the transferability of the models on other COVID-19 datasets that share similar characteristics to the original training dataset.

Chapter 1

Introduction

1.1 Background

Coronavirus is a family member of a large Ribonucleic acid (RNA) [65]. The outermost part of the virus contains glycoprotein spikes (see figure 1.1), which allows the virus to become trapped in living cells. Once inside the living cell, the virus replicates the RNA and reproduces. The SARS- COV-2 (severe acute respiratory syndrome coronavirus 2) variant is called novel coronavirus. This is called novel because this contagious virus has never spread to humans before and it recently evolved to spread rapidly from person-to-person [34, 65].

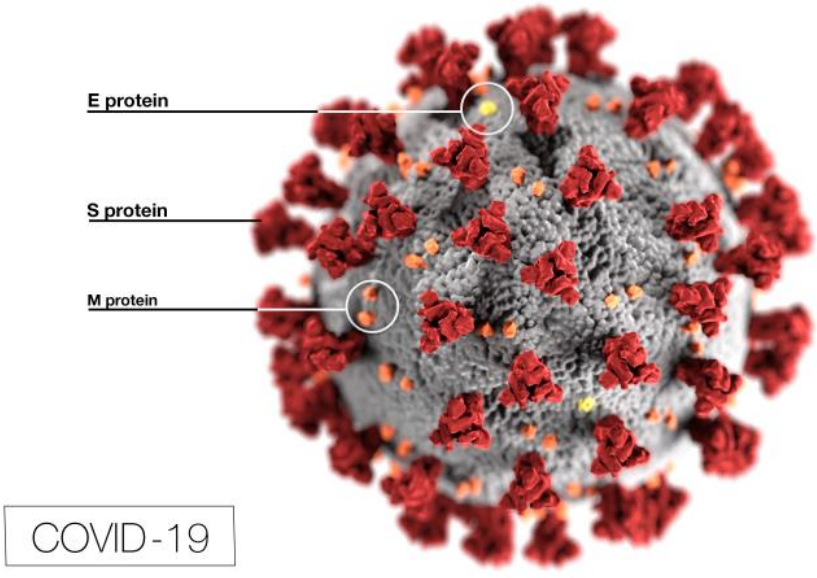


Figure 1.1: Coronavirus Structure, [Source: CDC]

In a short period, the virus had spread to seven continents and more than 6.5 million people have died as of now as a result of the spread. According to the World Health Organization, the period of incubation for the said virus ranges from 2 to 14 days in the human body. The reason why coronavirus is lethal is that it shows different symptoms in different people. For some people, it shows slight symptoms such as loss of smell and taste and headaches whiles others suffer from acute shortness of breath, physical weakness and recurrent coughs. In the face of the spurring growth of the epidemic and the incessant emergence of mutant strains or so-called variants of the virus, it is very important to implement emerging technologies for precise prediction on the number of infected cases to enhance planning towards future economic development of all countries in the world in which Ghana is inclusive.

Machine learning has been widely used in disease recognition [39], classification and analysis of influencing factors. It has been applied in the field of Covid-19 disease prediction, classification, drug efficacy and has been seen to show great advantages. In view of the prediction, [70] analyses of the spread and development trend of the epidemic, the prediction models built by scholars mainly focus on the dynamic model and statistical models. Compartmental mathematical models such as SIR [43, 4] models were used to understand the Covid-19 epidemic as well as to fit the number of cumulative confirmed cases for revealing the patterns and rate of spread of the virus in different countries.

However, all the mathematical models based on the first wave [11] of the pandemic were not able to provide an accurate forecast of the second wave of the virus due to the substantial parametric changes from the first to the second wave [7]. One of the general issues in using compartmental mathematical models relates to the gradual loss of accuracy resulting from the time-invariant formulation of hyper-parameters which prevents the models from following the evolution over time of the epidemiological phenomenon under investigation. An alternative approach based on machine learning which uses only the initial set and avoids recalculation of hyper-parameters has been implemented. The expected loss of accuracy in the models is corrected based on the use of models that are specifically tailored for time series problems [73]. Hybrids of deep learning and statistical techniques, so-called hybrid models [78, 76] are also used.

Epidemics have huge impacts on society socially and economically. The recent Novel Coronavirus has led to most countries declaring states of emergency and resulted in drastic changes to livelihood and the economy. This has set-back society in several domains. There is the need to understand epidemic spread and make future predictions related to the spread for better control and preparedness

1.2 Problem Statement

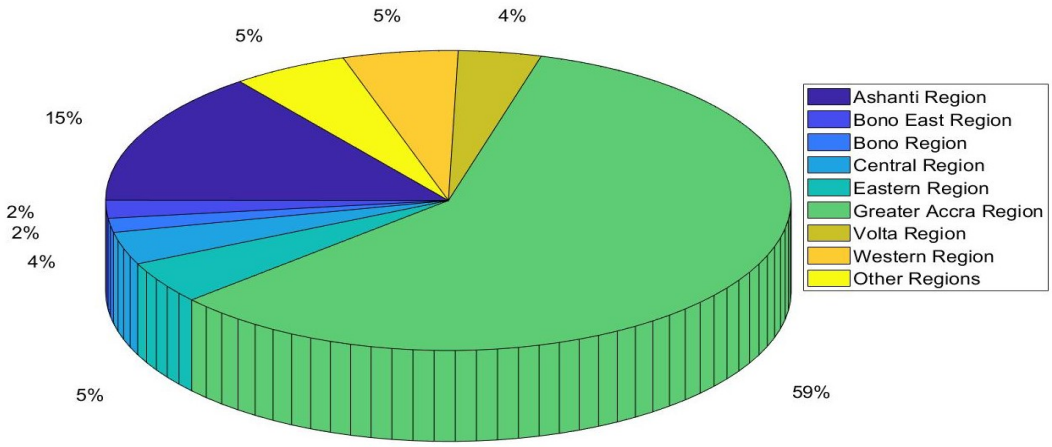


Figure 1.2: Chart of Covid-19 distribution in Ghana as of June 2022

As the corona virus continues to spread in the country, the question that we are trying to answer is "how do we predict the future trends and cases for each for the nation and each region?", and "what is needed to make these predictions?". We need a strong model or models that predict how the virus could spread across different districts, regions and at the national level. The Goal is to build a model that predicts and forecasts the spread in a month or two or years. Our goal is forecast the spread for both the long-term and short-term, using Covid-19 pandemic as the case study in Ghana.

1.3 Research Objectives

1.3.1 General Objectives

The end goal of this project is to predict and forecast the extent of Covid-19 spread by reviewing machine learning techniques, defining and implementing the suitable models for modeling the outbreak, since the traditional way, i.e., the compartmental modeling method is not sufficient to predict and model the dynamics of the viral spread. Dynamic aspects of the virus such as cold seasons, new variants and seasons of high mobility of individuals need to be modeled and the associated long-term predictions and forecasts made. Our focus is to use statistical and deep learning models which are capable of learning these complex non-linearities and relation and making useful forecasts.

1.3.2 Specific Objectives

For the above–mentioned objectives to be successfully met, the following specific objectives must be achieved.

1. Complete a deep study of mathematical models and network science techniques currently employed in modeling epidemic outbreak and outline their insufficiencies.

2. Obtain reliable time-series datasets on Covid-19 for training the models.

3. Apply suitable machine learning and deep learning and mathematical models to the data and improve on model accuracy.

4. Evaluate the performance of the models and chose a suitable general model

5. Document project and validate results.

1.4 Significance of Studies

This work focuses on developing machine learning models which has been shown to provide excellent results in similar problems and other fields of study. Our proposed models bring

about an increase in precision prediction of the covid-19 virus. As a result, this project has the potential to bring about the following benefits.

- Allocation of health resources to suite the healthcare, economic and societal needs;
- Better Covid-19 management decisions for spread control; and
- Set a framework for analyzing future epidemics in Ghana based on time series data.

1.5 Organization of the report

Chapter 2 starts by providing the reader with an overview of the epidemiology and a historical perspective to the modeling of epidemics. It then provides an in-depth survey of the field of Machine learning and its significance to epidemics. The chapter gives the reader the basic foundation needed in machine learning to understand the rest of this document. A comprehensive review on machine learning and epidemics spread, recent advances in the area of machine learning to model epidemics is also detailed.

Chapter 3 is the research methodology. It provides insights to our design approach and the various steps and algorithms used.

The fourth chapter discusses the implementation results as well as the performance of the proposed models. Our solution is evaluated in comparison to other models and the performance of proposed models are compared, one to another. This section illustrates how the specific objectives set out for this project are met.

The report concludes by summarizing the findings from the research and its applicability to society. It also proceeds to provide recommendations and suggestions for future work.

Chapter 2

Literature Review

2.1 Introduction

An infectious disease that spreads rapidly over a vast area and impacts lots of individuals at once is called an epidemic. Epidemics occur when a disease spreads from patient to patient more quickly than medical professionals can stop it. Sometimes a disease is referred to as a pandemic if it spreads globally. These illnesses need to be treated right away if you want to prevent them from spreading widely and endangering everyone on earth. The provision of timely vaccination and medical care is crucial. Epidemiology is simply the study of epidemics.

With Covid-19, the spread typically happens within 3 days to two weeks. Infectious disease epidemics are typically brought on by a combination of factors, such as changes in the ecology of the host population (such as elevated stress levels or increased vector species density), genetic changes in the pathogen reservoir, or the introduction of an emerging pathogen to a host population.

An epidemic often starts when the transmission threshold is crossed and host immunity to either an established pathogen or a newly developing novel pathogen is abruptly decreased below that found in the endemic equilibrium. In problems involving sufficient amounts of data, machine learning methods have become increasingly popular. Healthcare is an applicable domain for these techniques and has been demonstrated by the prediction of diseases using health data such as the West Nile Virus [4]. Once an epidemic starts, infectious cases may be recorded daily and a cumulated. This results in a sequential dataset recorded daily or within

other time frames. As a result, we end up with a time series dataset which can be used for modeling and forecasting the spread (future number of cases) for the epidemic.

This part begins by giving a summary of traditional epidemic (compartmental) models. Network based simulation is also discussed and their limitations as applied to forecasting and modeling epidemic spread are discussed. We then proceed to discuss data fitting and certain historical time series techniques such as ARIMA and exponential smoothing. The limitations of these are also explained. Consequently, we then review and adopt machine learning techniques that are suitable for time series-based and sequence-like datasets to overcome the limitations in traditional models. This chapter details all these techniques and how they have been used by other researchers in modeling epidemic spread.

The proceeding section discusses Mathematical Epidemic Models (MEMs). MEMs include constraints including identifiability in data fitting, case specificity, and robustness [17], have been used extensively to predict future patterns in disease transmission, acquire insight into disease dynamics, and design control tactics to reduce outbreaks [17, 10].

Later, a discussion of how machine learning models outperform compartmental models and other historical models in terms of disease prediction serves as the conclusion for this chapter.

2.2 Synopsis of Compartmental models

Historically, a number of mathematical, so-called compartmental models have been put out to forecast the spread of a disease in a population. Though they offer lesser insights into the dynamics of epidemics than typical statistical methods, these compartmental models based on dynamic differential equations [53]. With the emergence of machine learning techniques, these compartmental models have received comparatively less attention [76].

Deterministic compartmental models are the most practical for modeling infectious diseases. The equations model may model the number of susceptible individuals (S), infections (I), recovered and removed (death) cases. The equations' parameters may include one or more of the following: time (t), transmission rate(α), infection rate (β), recovery or removal rate or death rate (γ). These rates are estimated based on the early stage of the epidemic spread. The equations' parameters can be changed to better model environmental factors, virulence

and societal constraints [11]. Depending on the number of compartments that are used in the model, the models' names are frequently abbreviated as SIR, SEIR, SEIS and SIDARTHE [4, 32] since they are based on flow patterns across compartments like susceptible (S), exposed (E), infected (I), and recovered (R) etc. The traditional susceptible, exposed, infectious, recovered (SEIR) model is the most widely applied theory among all of these models.

SIR

To understand compartmental modeling of epidemics, we will start from the simplest of the models. The goal of this section is to give the reader a basic understanding of the SIR model and highlight the key improvements of the other versions of the model such as SEIR and SEIS models. Finally, the use cases and limitation of these compartmental models is explained.

Infectious diseases that spread from person-to-person and where recovery from an infection gives long-lasting resistance can be forecast using this SIR and similar models. For the simple SIR model, there are three chambers in it;

- "S" stands for the number of susceptible people. A susceptible person gets infected with the disease and moves into the infectious compartment when they come into touch with an infectious person.

- "I" is for the number of infectious people. These are those who have contracted the disease and are able to spread it to other people.

- "R" stands for the total number of people who were removed, recovered, or died. These are people who were infected, got better, and either entered the removed compartment or passed away.

For infectious diseases like measles, mumps and rubella that are spread from person to person and where recovery gives long-lasting resistance, this model is reasonably auguring [38].

S, I, and R, are variables that reflect the population of each compartment at a specific time [38]. Figure 2.1 illustrates that the number of susceptible, infected, and recovered individuals may change over time (even if the total population, N size does not change).

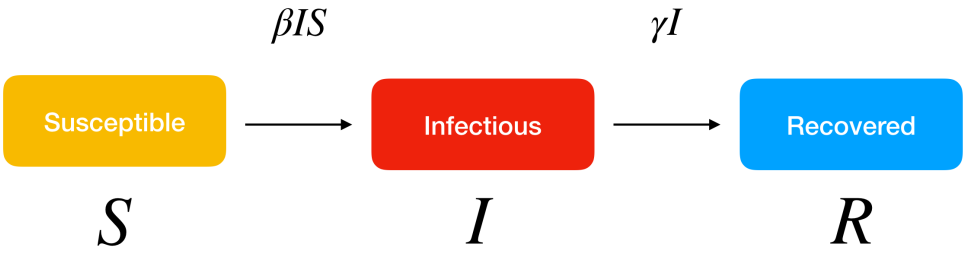


Figure 2.1: Demonstration of Compartmental Models

Consequently, the $S(t)$, $I(t)$ and $R(t)$ are all functions of time(t) and the system can be mathematically modeled by differential equations [21] as defined below. Please note that with the simple SIR model, only a subset of the parameters, including infection rate (β) and recovery rate, γ are used. For more sophisticated models, the transmission rate are other parameters may be factored in.

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

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

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

To perform a one-step forecast from the above models, we can rewrite and perform substitutions on the expressions above to obtain $S(t+1)$, $I(t+1)$ and $R(t+1)$ as a function of $S(t)$, $I(t)$, $R(t)$. This is the forecast that we seek to obtain. To obtain this, we will approximate

$$dS(t) = \Delta S(t) = S(t+1) - S(t) \tag{2.4}$$

We also perform a similar assumption for $dI(t)$ and $d[R(t)]$. Consequently, we obtain:

$$S(t+1) = S(t) - \beta \{S(t) \times I(t)\} dt \tag{2.5}$$

$$I(t+1) = I(t) + \{\beta S(t) \times I(t) - \gamma I(t)\}\, dt \tag{2.6}$$

$$R(t+1) = R(t) + \gamma I(t+1) dt \tag{2.7}$$

To implement this model, we clearly need to define the initial values $S(0)$, $I(0)$, $R(0)$ and the parameters $\beta < 1$ and $\gamma < 1$ and dt (in days), according to the initial data collected and estimated from the spread. It is worth mentioning that the Disease Free Equilibrium is achieved when $(N, S, I, R) = (N, 0, 0, 0)$ [18, 36] which is the goal of epidemic spread control.

The model is dynamic in that the values in each compartment may change over time, as suggested by the variable function of t. However, the model is unable to model unexpected changes in the disease spread such as outbreak due to more fatal variants, changing the trend of the curve.

Differential equations may be developed to fit the situation of a particular disease based on different compartmental models in a particular community to forecast potential outbreaks and control them.

Similar to the original SIR model as explained above, models such as SEIS and SIDARTHE have been designed to offer improvements for disease scenarios with more compartments. Adopting either of these models based on the particular scenario gives useful information for only the early stage of the disease due to dynamic changes in disease spread. Consequently, all compartmental models face this fundamental limitation.

The limitation of Compartmental Models

It maintains that in the event of irregularly changing contact rate, increase in virulence of the agent and increase in recovery rate which could be due to a vaccine; the above system and its parameters remain unchanged although the real system is changing. These classical compartmental models are therefore insufficient for modeling a disease such as Covid-19 which is highly dynamic. Figure 2.2 further shows the inefficiency of SIR models for Covid-19 as modeled, based on initial parameter estimates in Wuhan China which is the source of the virus.

The model is unable to fit the initial dynamic spread of the infectious cases [44].

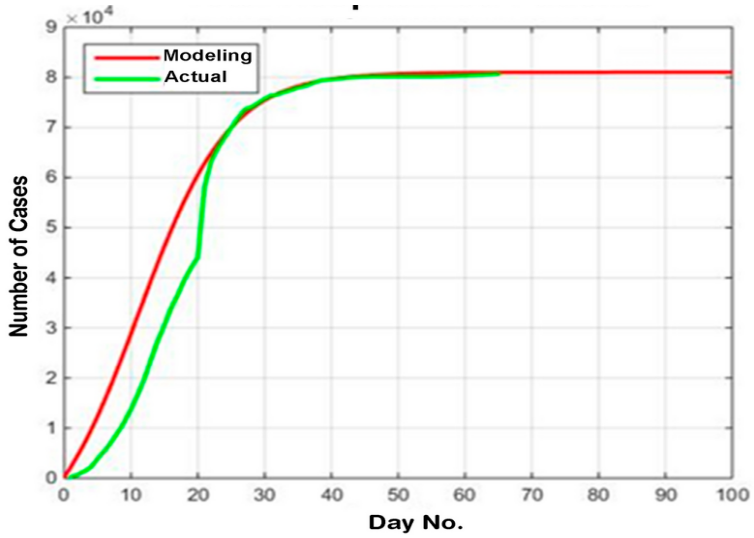


Figure 2.2: Inefficiency of SIR in dynamic spread. Source: Moein et al.

Another demonstration of the insufficiency of compartmental models is seen in [18, 36]. An endemic illness with a brief infectious period, like measles in the UK before a vaccine was introduced in 1968, makes the significance of this dynamic aspect particularly clear. Due to the fluctuation in the number of susceptible (S(t)) over time, many illnesses frequently go through cycles of outbreaks. As more people become sick and enter the infectious and removed compartments during an epidemic, the number of susceptible people rapidly declines. The disease cannot recur until the population of those susceptible has increased, such as by the birth of additional offspring into the susceptible compartment. Researchers have also tested the performance of SIR on Covid-19 and it is clear that without unique improvements to model the dynamic changes in parameters such as transmission rate, the model is unable to give a useful forecast of the spread.

2.3 Regressive and Curve Fitting Models

2.3.1 Applications of Regressive techniques to Epidemics

For infectious diseases, time series forecasting is a well-known difficulty. The COVID-19 pandemic time series has already been predicted by a large number of academics. The majority of studies [15, 40, 29] used statistical models, traditional machine learning, and deep

learning (see section 2.5 for details). Utilizing data from January 22, 2020 to April 13, 2020, statistical models, including the autoregressive integrated moving average (ARIMA), single exponential, double exponential, moving average, and S-curve models were used to predict daily new COVID-19 cases in India [40]. With a minimum mean absolute percentage error (MAPE) of 4.1, experimental results of the study demonstrated that ARIMA (2,2,2) outperformed other techniques. With little training data, exponential smoothing models were used to predict a variety of trends and seasonal patterns [51].

Researchers have showed a strong interest in studying India's dynamic expansion. Swaraj et al. created a model for predicting the COVID-19 epidemic in India that used ARIMA and a nonlinear autoregressive neural network (NAR). The hybrid model significantly reduces evaluation metrics (RMSE: 16.23 percent, MAE: 37.89 percent, and MAPE: 39.53 percent) as compared to the single ARIMA model [64]. Wadhwa et al., [71] also investigated the impact of lockdown policy on disease transmission by projecting the number of active cases across India. They created a graphical depiction of the COVID-19 cases three months ahead using the Linear Regression (LR) model [71]. In another work, Khan et al., [27] used three (3) machine learning models (Decision Tree (DT), SVM, and Gaussian Process Regression (GPR)) to examine the time point at which the number of instances in India stops increasing, allowing them to analyze policy restrictions. According to their findings, the GPR model surpasses the other models with a 95% accuracy.

In another study, Silva, Francisquini, and Nascimento examined the number of infections in the top 27 impacted Brazilian cities using a single ARIMA and a hybrid model that combines the Ensemble Empirical Mode Decomposition (EEMD) approach with the ARIMA. Their findings suggest that the ensemble model outperformed the single model by 26.73 percent [62].

Having obtained an overview of the applications of regressive techniques to epidemics, we focus on discussing the mathematical and technical details of the common and most outstanding of these models in the next sections.

2.3.2 Regression Analysis

In regression, the goal is to obtain a simple fit to the data. Using a regression model for forecasting will require extrapolation to be performed on the obtained model to the new, out of sample data [33]. Several regression techniques exist including logistic, linear, polynomial, binomial and multinomial regression. We have explained a few of these that are relevant to our research in the proceeding sections.

Logistic Regression

The likelihood of the target value is estimated using the logistic regression method. The target value is discrete, which means that the data is coded in binary or higher number of levels form. In the binary case, 1 denotes success and 0 denotes failure [73].

$$p(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 t)}} \tag{2.8}$$

This probabilistic model as seen in equation 2.8 above can be used to determine if the number of COVID-19 cases may hit a certain target. $\beta_0 = 0$ in our case represents the y-intercept while β_1 represents the rate parameter. $p(t)$ represents the probability of COVID-19 cases reaching a specified number, after a specified time t, in days. This concept is echoed in other advanced models such as the prophet model with a logistic growth trend.

Linear Regression

In Linear Regression, the goal is to fit a linear curve to the sequential data. Extrapolation of the linear model curve can then be made to produce a forecast. Equation 2.9 and figure 2.3 below demonstrate this technique for the Ashanti Regional data in Ghana; which has only a single time (in days) independent variable. We can observe underfitting due to the low order of the model and its inability to fit the seasonal or dynamic aspects of the data. We will attempt to make improvements to this by using higher order models.

$$\overrightarrow{y} = \beta_0 + \beta_1 \overrightarrow{t} \tag{2.9}$$

It is important to note that $\overrightarrow{y}$ and $\overrightarrow{t}$ are vectors of cumulative cases and days

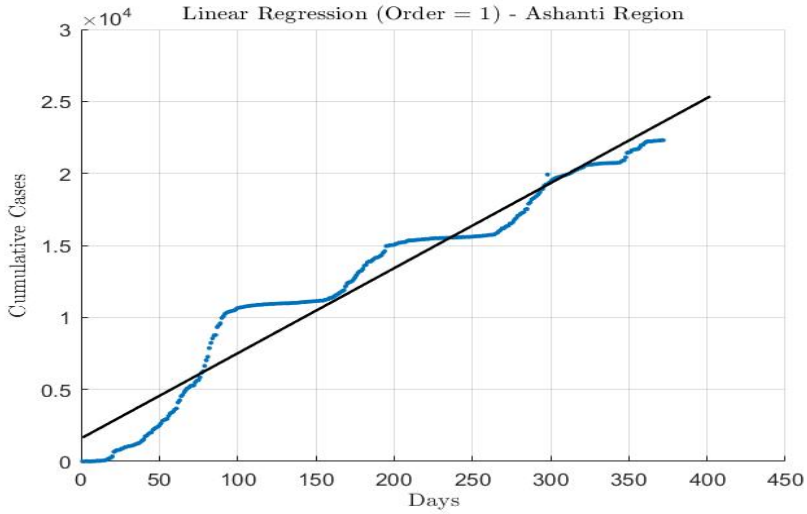


Figure 2.3: Underfitted Linear Regression on Ashanti Regional Data

The size, n of the vector $\overrightarrow{\beta} = [\beta_0, \beta_1, \cdots, \beta_n]$ depends on the number of independent input variables, defined by $n = m + 1$, where m is the number of independent input vectors. The general solution for β can be obtained by solving the least squares problem stated in equation 2.10 below.

A is the matrix of independent input parameters with size $n \times m$.

$$\overrightarrow{\beta} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\overrightarrow{y} \tag{2.10}$$

It is worth mentioning that Linear regression may be used for both daily records and cumulative records of COVID-19 cases. However, it is more useful for cumulative cases as there is better linearity and correlation instead of random records.

Polynomial Regression

Polynomial regression for a single variable fit is defined as:

$$\overrightarrow{y} = \beta_0 + \beta_1\overrightarrow{t} + \beta_2(\overrightarrow{t})^2 + \cdots + \beta_3(\overrightarrow{t})^n \tag{2.11}$$

The $\overrightarrow{\beta}$ entries are also obtained in a least squares method as seen in equation 2.10 with entries of **A** being exponents of the preceeding column.

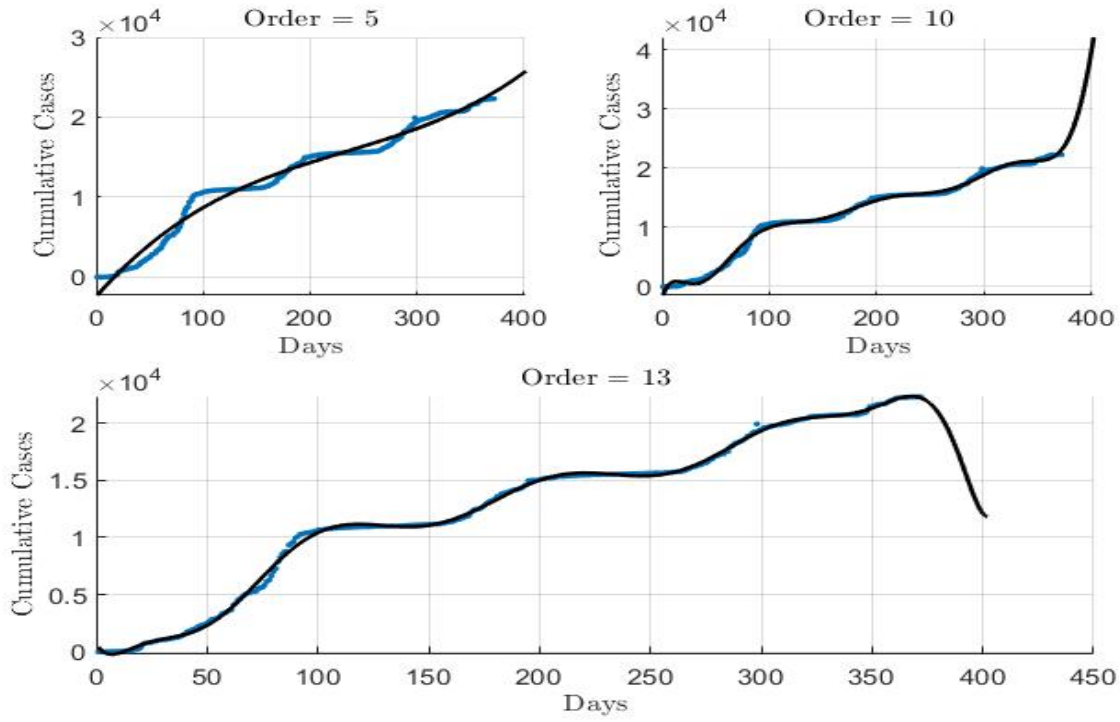


Figure 2.4: Order Selection and Polynomial Fitting

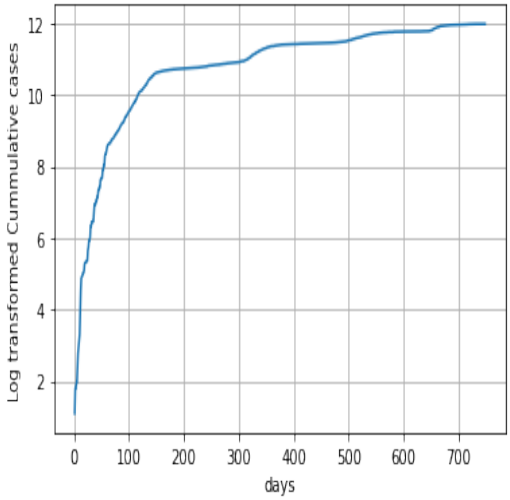
In the single variable polynomial regression we implemented in MATLAB, the tendency of overfitting is observed. The model order must therefore be selected carefully in such a way that the extrapolation makes sense and the fitting is valid. Nikhil et al. in [47] implement a degree 2 regression model which turns out to be useful but displays the limitations of the regression models. Consequently, the limitation of polynomial fitting is the trade-off between order selection and the fitting of data dynamics. We can see this from figure 2.4 that an unexpected extrapolation (forecast) occurs at Order = 13.

2.3.3 Exponential Smoothing

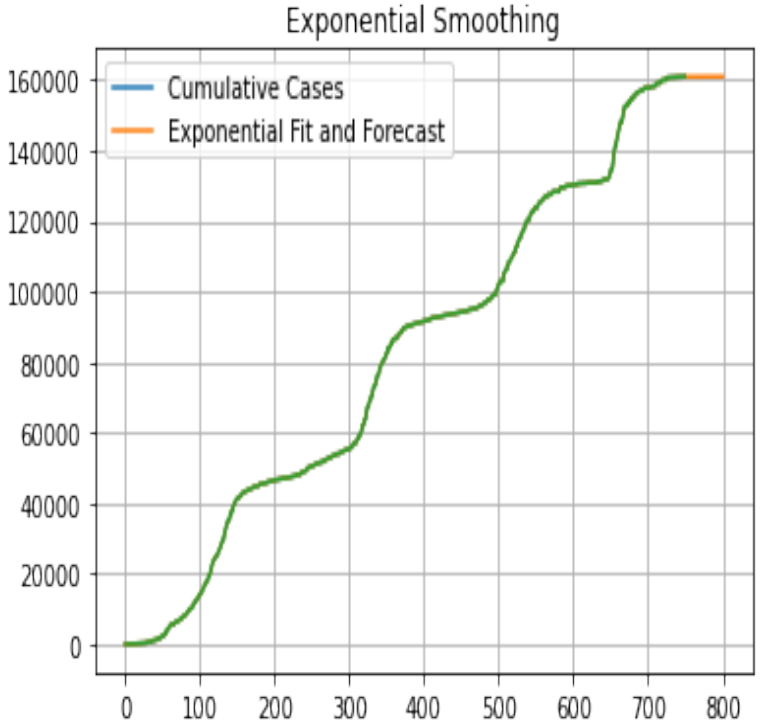
Exponential Smoothing refers to a suite of techniques for fitting exponential curves on data that follow an similar pattern. While we can clearly see growth behaviour in the COVID-19 cumulative trend cases, we can identify that there is an element of seasonality to it. Therefore, we need to transform the data to a scale with a trend that is more exponential. We perform this task by implementing a logarithmic transformation using the `np.log()` function in python. The exponential smoothing is performed using `statmodels.tsa.holtwinters.SimpleExpSmoothing()`

object in Python.

After fitting the new data to the exponential model and rescaling the prediction to its original scale, we obtain the fits seen in 2.5b. We can clearly identify from the plots that exponential smoothing does not provide any suitable forecast although it gives an almost perfect fit on the training set.



(a) Log Transformation of Cumulative Cases



(b) Exponential Fit and Forecast of Cumulative Cases

Figure 2.5: Simple Exponential Smoothing Fit for Ghana's National COVID-19 data

2.3.4 Autoregressive Integrated Moving Average

ARIMA is a time-series forecasting method that combines the autoregressive (AR) and moving average (MA) statistical methods with differencing for modeling of time series data [60, 57]. Equations 2.21 and 2.13 below demonstrate the Box-Jenkins equation, showing the AR and MA components of the model. ARIMA apparently shows good performance and competitive RMSE for a worldwide and other COVID-19 time series [65, 35]. In addition, SutteARIMA averages the forecasting results of the -Sutte Indicator and ARIMA [3], and Holt-Winters, which can capture three important aspects of time-series data: the average, trend, and seasonality [59], have been used to predict the pandemic's development.

An ARIMA model is usually denoted as `ARIMA(p,d,q)`, where:

- p denotes the order of the AR. term and the number of lagged regressors used as a predictors;

- d denotes the differencing factor needed to make the time series data stationary; if $d = 0$, then the data is assumed to be stationary and the model defaults to an ARMA model; and

- q is the order of the M.A. term and the number of lagged predicted errors that are used in the ARIMA model [69, 47].

ARIMA improves on ARMA by using the difference factor (d) which makes a non-stationary series to be stationary. For a set of lagged data points, the MA term (window size) q, reveals the relationships between observations and the residual error terms (q). ARIMA can cope with non-stationary time-series data by using an integrated step to transform the non-stationary series to a stationary series via differencing, whereas ARMA only applies to stationary processes.

$$\hat{y}(t) = c + \phi_1 \hat{y}(t-1) + \phi_2 \hat{y}(t-2) + \cdots + \phi_p \hat{y}(t-p) + \epsilon \tag{2.12}$$

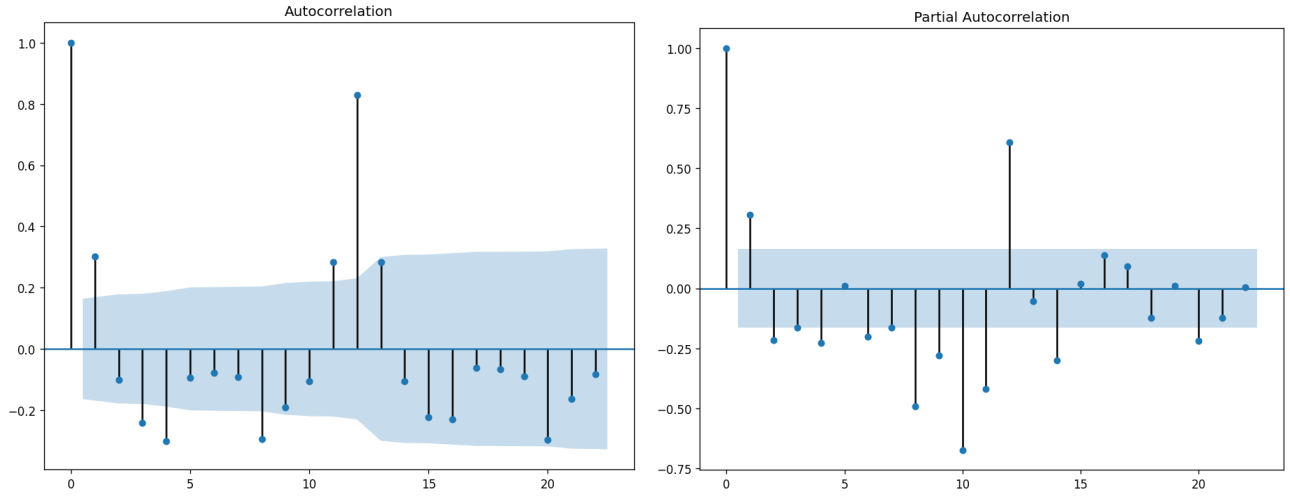
$$\Rightarrow \hat{y}(t) = c + \underbrace{\sum_{i=1}^{p} \phi_i \hat{y}_{t-i}}_{\text{AR}} + \underbrace{\sum_{i=1}^{q} \theta_i \epsilon_{t-i}}_{\text{MA}} + \epsilon_t \tag{2.13}$$

All parameters of the ARIMA model are determined by performing statistical tests for stationarity such as the Dickey-Fuller test; and by using statistical plots such as autocorrelation (ACF) to determine q and Partial autocorrelation (PCF) to determine p. An algorithmic overview of the ARIMA model can be seen in algorithm 1 below.

p and q are determined from the PACF and ACF plots by counting the major number of lags whose values are beyond the confidence bound, see figure 2.6b and 2.6a. A trade-off between model complexity in selecting the number of lags and the p and q values must be made.

Algorithm 1 ARIMA Model for Forecasting

p-value = $ADF(series)$
if p-value > 0.05 **then** ▷ Time Series is Stationary
    Choose d
    $series = function(series, d)$ ▷ Differencing based on window size d
end if
Visualize $y(t)$
Choose number of lags l ▷ Based on visualization or stationarity test
for $n \leftarrow 0$ to l **do**
    $p = count(PACF >> threshold)$
    $q = count(ACF >> threshold)$
end for
$model \leftarrow ARIMA(p, d, q)$
$model \leftarrow fit(model, data)$
$future \leftarrow forecast(model, future)$ ▷ Future cases
if p-value > 0.05 **then**
    future = cumsum(future) ▷ Inverse Difference via Cumulative Summation
end if



(a) ACF Plot, $q \approx 2$ (b) PACF Plot, $p \approx 5$

Figure 2.6: Autocorrelation and Partial Autocorrelation Plots

Shortcomings of ARIMA

Clearly, ARIMA is only suitable for stationary time series and for datasets with strong seasonality, yet stationary. The major challenge with using ARIMA for our datasets is perhaps the lack of strong seasonality and the wide range of transformations required will result in lack of accuracy in the inverse transformation of the prediction and forecasts. Implementations by other authors also show ARIMA is unreliable for long-term, time series forecasting as the trend behaviour of the forecast becomes inconsistent with the trend of the original data.

2.4 Network Models

Communities are essentially social networks with densely connected nodes. The nodes represent individuals in the community while the links represent interactions between individuals [43]. We can model a the outbreak of a disease in a community by thinking of links to represent a person-to-person interaction that may results in a probabilistic transfer of the infection. Previous work by Kponyo et al. in [30] demonstrate this for a simulated community, bench-marking it against the Ebola virus spread. The software tool used is Net-logo. See figure 2.7

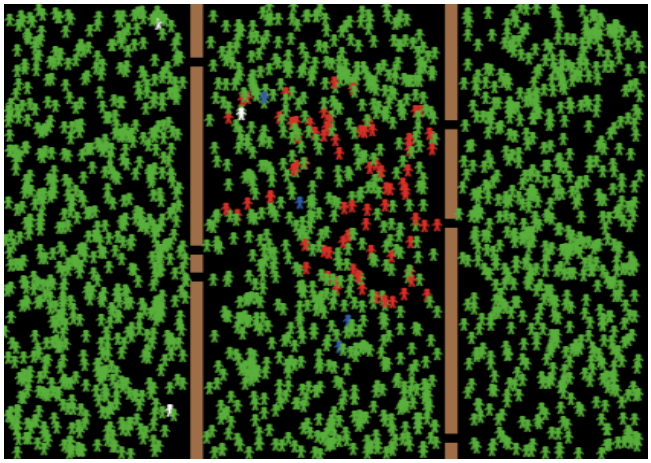


Figure 2.7: Net-logo approach to simulate spread in a community

Kuzdeuov et al., [32] designed a network simulator specifically for epidemic spread simulations based on randomness. The issue with network modeling of epidemics is the lack of reality in the modeling, since the extremely large graphs with trillions of relationships that are all meaningful, essentially creates a multi-graph that is computationally expensive to simulate.

2.5 Machine Learning Models

2.5.1 Introduction to ML/DL and Applications to Epidemics

Machine Learning refers to a computer programming where computers make decisions based on data instead of explicit instructions. It is a technique for data analysis that involves creating and fitting models and enables computers to "learn".

Several investigations [22, 31] revealed various laboratory results at the start of the COVID-19 pandemic. The majority of the instances are minor, and patients' clinical results vary widely

[23, 5]. As a result, identifying risk groups based only on gender and age may be problematic. In addition to these, it is critical to anticipate which individuals are more likely to develop serious disease and face a higher risk of mortality. These are critical considerations when clinical resources and equipment (hospital beds, medical masks, respirators, hospital capacity, etc.) are limited and health care workers are forced to make judgements about patients with no prior experience to guide them. Because of these constraints, such judgments must be made by an artificial intelligence (AI) enabled system. In healthcare systems, AI is actively used to provide clinical decision support [66, 13, 54]. Machine learning classifiers are useful for interpreting medical data such epilepsy [42], nerve and muscle illnesses [75, 24]. Deep learning methods are also useful for predicting clinical outcomes in cancer [46], viral illnesses and biological investigations [58]. Such methods are effective and can be used to predict COVID-19 infection in a clinical setting.

Recurrent neural network (RNN) versions, one type of deep learning model, has been widely utilized to forecast infectious disease outbreaks. In India, a RNN variation known as long short-term memory (LSTM) was used to forecast COVID-19 cases [68]. The experiment employed COVID-19 data from January 30, 2020, to April 4, 2020, with 80% of the data used for training and the remaining 20% for forecasting. For the days from April 5 to April 9, the model generated an error percentage ranging from 1.6 to 6.4 for all confirmed cases. To predict COVID-19 transmission in Canada, Italy, and the USA, an LSTM-based network was used [9]. For training (January 20, 2020 to December 3, 2020) and forecasting of confirmed, negative, released, and deceased COVID-19 cases, several RNN variations, including LSTM and GRU, were used [14]. The evaluation metrics employed were accuracy and RMSE.

Machine learning models have been used successfully used to understand other aspects of the pandemic including the design antibodies [37], using medical image datasets, particularly chest X-rays [67], modeling and understanding mutations [45, 72], detecting whether a patient is infected with SARS-CoV-2, and forecasting pandemic trends.

We focus on anticipating pandemic patterns for Ghana i.e. the national level and the regional level. In this scenario, we may compare our findings to those of earlier studies. This section is a list of previous forecasting studies that used machine learning which will be relevant in our

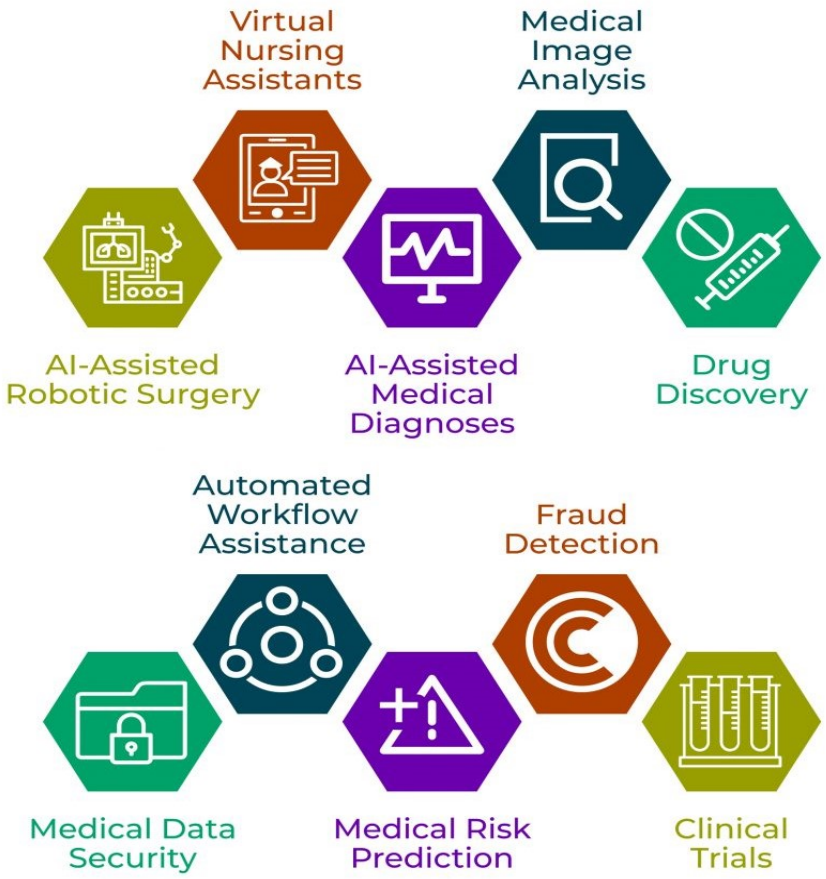


Figure 2.8: Applications of AI/ML in Healthcare

studies as well. Brazil, being one of the most severely impacted nations by the epidemic, has been extensively examined by experts. To examine the cumulative confirmed cases in Brazil, Ribeiro et al. employed autoregressive integrated moving average (ARIMA), cubist regression, random forest, ridge forest, SVR, and stacking-ensemble learning, respectively [55].

A study used training on restricted data sets of 30 and 40 days to forecast the pace of spread in Brazil using the Gated Recurrent Unit (GRU) [19]. They discovered that the best accuracy of 85 percent was reached on a 30-day time-step utilizing validation data from 4/7/2020 to 6/13/2020. However, as the predicting period lengthens, the accuracy falls precipitously (to a high of 68%), showing that the model performs rather badly in long-term forecasting.

Russia, in addition to Brazil and India, has been extensively researched. Wang et al., [72] created an LSTM model to estimate pandemic trends 150 days in advance using daily new confirmed cases in Russia, Peru, and Iran [72]. In another study, the Bayesian model was used to assess the impacts of lockdowns on COVID-19 transmission in the top five nations from March

1 to June 29, 2020. (India, Brazil, Russia, USA and UK). It has been established that if the lockdowns are relaxed, the outbreak pace would dramatically rise in Brazil, India, and Russia [16]. In [12], Dairi et al. compared the prediction performance of machine learning methods (LR and SVR) and deep learning methods (the hybrid LSTM-CNN, the GAN-GRU hybrid, CAN, CNN, and LSTM).Deep learning methods beat traditional machine learning tools in terms of predicting performance, with LSTM-CNN demonstrating the highest accurate prediction with a MAPE of 3.718 percent.

More deep learning algorithms, notably RNN, LSTM, BiLSTM, GRU, and VAE, were investigated in order to forecast COVID-19 instances in various nations (Italy, France, Spain, China, Australia and the USA) by Zeroual et al. in [77]. The study concentrated on real-time cumulative daily cases (from January 22, 2020 to March 11, 2020), and it created forecasts for 10 days in the future, updating them every ten days. Another investigation employed models including ARIMA (see section 2.3.4), cubist regression (CUBIST), random forest (RF), support vector machine (SVR), and stacking-ensemble learning for short-term forecasting: one, three, and six days ahead [51]. This analysis focused on daily cumulative cases of COVID-19 in 10 Brazilian states. The analysis revealed that in the evaluation, the SVR had the best mean absolute error (MAE) and symmetric MAPE. Utilizing data, cumulative confirmed and recovered cases from around the world were subjected to autoregressive time-series models based on the two-piece scale mixture normal distribution (which does not rely on the assumption of the symmetric distribution of error components).

2.5.2 Review of Supervised Learning for Epidemics

The identification of COVID-19 has resulted in the use of supervised learning techniques such as regression, classification, and feature extraction [50].As opposed to unsupervised learning, supervised learning uses labelled data. For supervised learning, the general implementation workflow can be seen in figure 2.9 below. This starts from the dataset through modeling with the necessary techniques, updating parameters and making the predictions. Several COVID-19 related tasks can be modeled using supervised learning. Some of these tasks include:

1. Prediction and forecasting of the coronavirus transmission over areas,

2. Analysis of the expansion rate and types of treatment across different countries,

3. Correlation between the effect of weather conditions and the coronavirus,

4. Analysis of the virus's transmission rate (see Yadav, Perumal, and Srinivas,[74]).

Some studies have concentrated on forecasting by using logistic models, neural network-based prediction and a hybrid ensemble of non-linear autoregressive neural network with type 2 fuzzy and firefly algorithm[26, 56]. Other researchers have also analyzed the geographical relationship in COVID-19 dissemination using ML[41], ML-based evaluation of COVID-19-related health opinions and online content, and others have used ML techniques for examining the consequences of the COVID-19 pandemic on young students' activities, mental health, and learning styles (Khattar, Jain, and Quadri, 2020 [28]).

After analyzing the current COVID-19 vaccine candidates, Ong et al., 2020 [48] suggested a model based on ML and reverse vaccinology (RV) for the prediction of potential vaccine candidate proteins. The primary function of RV is to investigate the pathogen genomes' bioinformatics. As a result, vaccination candidates with potential are found. The UniProt and NCBI databases served as the source of the dataset used in this investigation (see Bairoch et al., 2008 [6]). It is made up of all the proteins and SARS-CoV2 sequences that have been taken from recognized human coronavirus strains. To predict the biological properties of the proteome in this study, the authors used Vaxign and Vaxign-ML [20, 49]. They improved the Vaxign-ML model based on reverse vaccinology, utilizing XGBoost, Machine Learning, Support Vector Machines, K Nearest Neighbour, and Logistic Regression (LR) techniques were used to predict the protein levels of all SARS-CoV2 proteins.

According to time series data compiled from the Hungary statistical reports of infected cases and death rates, Pintér et al., 2020 [52] investigated the possibility of using a hybridization model of network-based fuzzy inference system and multi-layered perceptron-imperialist competitive algorithm to forecast the outbreak of COVID-19. The suggested prediction model's effectiveness was assessed using three metrics: mean absolute percentage error (MAPE), root mean square error (RMSE), and determination coefficient. The suggested prediction model performed well in calculating the overall mortality and forecasting the COVID-19 epidemic.

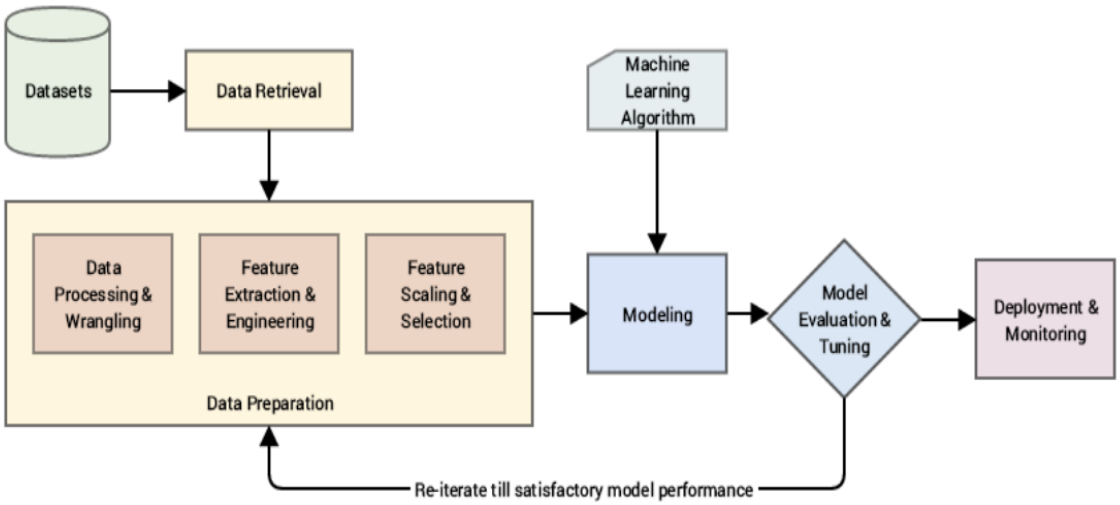


Figure 2.9: General Workflow Diagram for Supervised Learning

An online survey was developed by Fayyoumi, Idwan, and AboShindi,2020 [15] for normal and COVID-19 cases in Jordan. The existence of signs and symptoms in both groups was determined using data from the questionnaire. The researchers compiled a COVID-19 dataset comprising various patients' indications and symptoms. The researchers then used this dataset as input to a variety of machine learning (ML) models (SVM, multi-layer perceptron [MLP]), as well as statistical methods (i.e., LR), to predict future COVID-19 patients. The classification accuracy revealed that MLP performed at its peak (91.62 percent). SVM performed the best in terms of precision (91.67%).

Kavadi et al., [25] suggested a different strategy for stopping the COVID-19 outbreak in India. The researchers made use of the Indian COVID-19 database. The suggested method combines partial derivative regression and nonlinear machine learning (NML), two well regarded methodologies. The dataset was normalized using PDL, and prediction was performed using NML. According to experimental findings, this technique is better than earlier efforts in terms of classification accuracy and forecast time.

The COVID-19 trend was predicted by Zheng et al. and Wang et al. in [78, 72] using the logistic modeling, Facebook Prophet and LSTM models. The most recent COVID-19 epidemical dataset pertaining to time series data at the country level is the one used in this investigation. This study includes a large number of nations, including Brazil, Russia, India, Peru, and Indonesia. In actuality, logistic modeling was used to determine the pandemic trend's cap

value (refer section 2.3.1). The Facebook Prophet model's learning model was then developed using the results in order to anticipate the COVID-19 epidemic trend. The proposed strategy will allow decision-makers in a single country to move quickly during a COVID-19 outbreak, according to the experiment's findings.

We can derive from this review that traditional machine learning and deep learning-based techniques have been used extensively for diverse epidemic related tasks, including prediction and forecasting of the number of cases

2.5.3 Unsupervised Learning for Modeling COVID-19 Spread

Unsupervised learning uses unlabeled data to perform tasks such as clustering, association and dimensionality reduction. The goal is to group data into comparable clusters without using any extra information. In general, there are two types of clustering techniques: hierarchical clustering and partition clustering. Hierarchical approaches may be divided into two types: divisive (top-down) and agglomerative (bottom-up). Starting with all items in one cluster, divisive clustering attempts to separate them into smaller groups until a stopping requirement is met. In contrast, agglomerative clustering considers each object as a single cluster before consolidating them into bigger clusters until a stopping requirement is met (see Abasi et al., [2]) designed a partition algorithm to split texts into distinct groups. These strategies often employ each cluster's centroid to gather comparable data [1]. This approaches' ultimate goal is to efficiently disseminate a huge quantity of data into a number of heterogeneous clusters, each containing homogenous data (Abasi et al.,2020).

Using the k-means technique, Siddiqui et al. in [61] looked at the relationship between patient temperature and the COVID-19 case status (i.e., suspected, confirmed, and death). The three phases of the process are database design, clustering, and data collection. The "coronavirus disease (COVID) situation reports" collected from the WHO are the dataset utilized in the initial analysis. The infection rates in different parts of China are included in this dataset. The dataset's seven (7) features (i.e., region, population (in 10,000s), suspected cases, confirmed cases, death, lowest temperature and highest temperature features. The rationale is that one of the key elements in determining COVID-19 case status is patient temperature. In

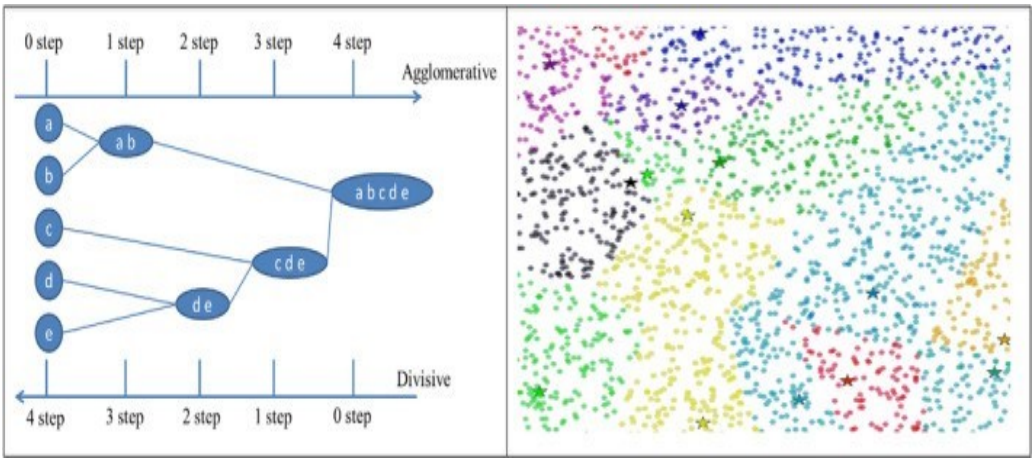


Figure 2.10: Use of agglomerative and divisive functions on a dataset of five items, a, b, c, d, e. (b) Using partition clustering on a dataset with fourteen clusters

the final stage, new patterns are discovered using clustering algorithms based on k-means. The patterns showed how the three COVID-19 states' three regions were affected by temperature (i.e., suspected, confirmed and death).

Castillo-Cara and Carrillo-Larco, [8] also employ kmeans to generate data-driven country clusters for forecasting COVID-19 impact (2020). It was influenced by health system coverage, socioeconomic status, air pollution measures, and illness prevalence estimates. The researchers compare the clusters based on the case fatality rate, the number of fatalities, the number of verified COVID-19 cases, and the sequence in which the nation detected the first case. The model was created and used to define clusters with 155 nations. The model employs three principal component analysis (PCA) parameters, and five or six clusters yield the best results by merging nations into similar groups. According to the findings, a model of five or six clusters can stratify nations based on the reported number of COVID-19 cases. This is visualized in figure 2.10 above.

Having seen the machine learning models other researchers have adopted for COVID-19 related tasks, we will narrow it down to the most useful ones for forecasting future cases. The preceeding sections discuss these models which include Facebook's prophet model, LSTMs and Neural Prophet.

2.5.4 Forecasting with Facebook Prophet

SJ and B in [63], researchers at facebook invented the prophet model for forecasting at scale. The model requires time recorded data which could be dates or datetime. The utilized concepts such as trend modeling, seasonality modeling and factoring of holiday and error effects.

Trend Analysis

The trend function in facebook prophet may either be a logistic growth function or a linear function. The linear growth function is estimated from the training data and is adaptable the function is adaptable to changes in the time series data. On the other hand, the cap or maximum value, C has to be defined for both forecasting and training for the logistic growth prophet trend model. C is adapted during training to fit the data hence it is a function of time, $C(t)$. The logistic trend function for prophet is defined in equation 2.14 below.

$$g(t) = \frac{C(t)}{1 + e^{-k(t-t_0)}} \tag{2.14}$$

Seasonality Modeling

The above-mentioned tendencies are summarized into weekly periods known as seasons. The Trend encapsulates seasonality. Seasonality is measured by calculating repeating patterns throughout a given time period. The Fourier Transform is used to find these periods. After then, the seasonal impacts are smoothed.

$$s(t) = \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi t}{P} + b_n \sin \frac{n\pi t}{P} \right) \tag{2.15}$$

Holiday Effects and Error Factor

Holidays are events that occur at a given time and drastically alter the way a time series progresses [33]. Because holidays do not occur on a regular basis, they must be included as input. Holidays have an impact on the days around the dates supplied as input [28]. All of these are added together to form the forecasting equation represented by the equation below. The error is expected to be normally distributed.

$$y = g(t) + s(t) + h(t) + \epsilon \tag{2.16}$$

2.5.5 Deep Learning Models for Forecasting

Overview of MLPs and Neural Networks

The foundations of deep learning (neural networks) are built on multi-layer perceptrons. All types of neural networks, from DNNs to CNNs LSTMs, transformers and more advanced models are built on this concept.

MLPs are based on the universal approximation theorem which states that a neural network with 1 hidden layer can approximate any continuous function for inputs within a specific range every. We therefore generalize this concept to solve difficult problems that cannot be explicitly defined mathematically. One of such problems include time sereies forecasting of COVID-19 which is our goal in this project. Figure 2.11 shows a simple MLP with one hidden layer. The so-called activation functions give importance to a specific mapped relationship between an input parameter and the output of a single node. Examples of activation functions include ReLu, softmax, sigmoid, tanh etc. The hyperbolic tanh function are generally preferred for LSTM memory cells while sigmoid is used used for gate activation (see figure 3.3.

GRUs, RNNs and LSTMs

RNNs have evolved from simple architectures to GRUs and LSTMs. Standard RNNs have a vanishing gradient issue and essentially lack memory. The GRU implements two gate operating mechanisms called Update gate and Reset gates to solve the problem faced by simple RNNs. LSTM fu

For the LSTM unit which has been adopted for forecasting tasks in a lot of already reviewed research, we can define its unit's functions as shown in equation 2.17. i, o, and f are the parameters used to update the input, output and forget parameters respectively. The sigmoid function is used for this purpose and its parameters are the output from the previous hidden state (h_{t-1} and input, x_t. The "memory" values of the cells, C_t are updated during the training using the tanh hyperbolic function. All these expression and more can be easily visualized and

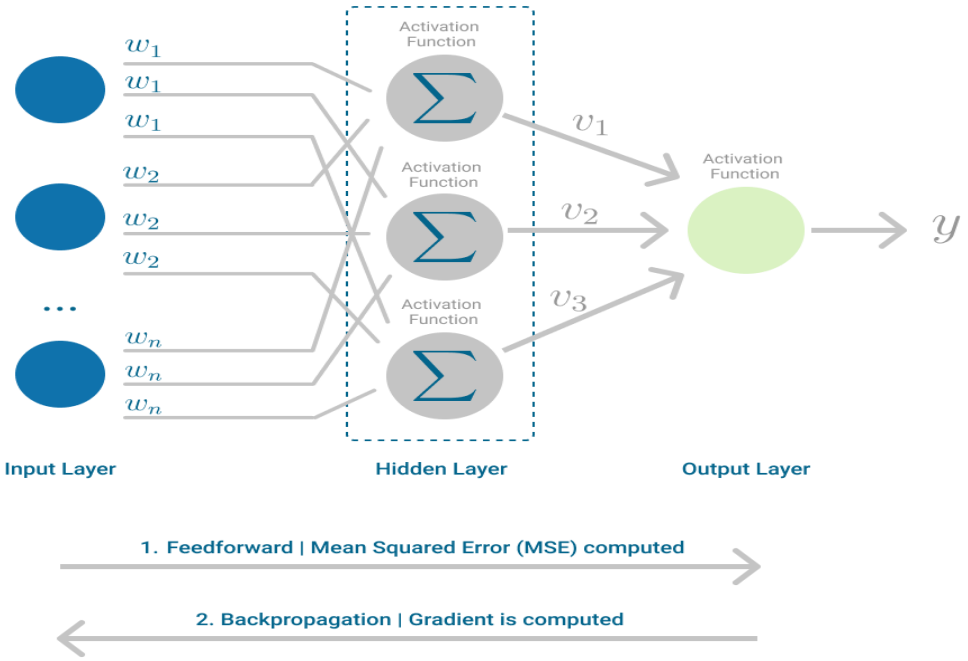


Figure 2.11: How a Multi-layer/Deep Neural Network Works!

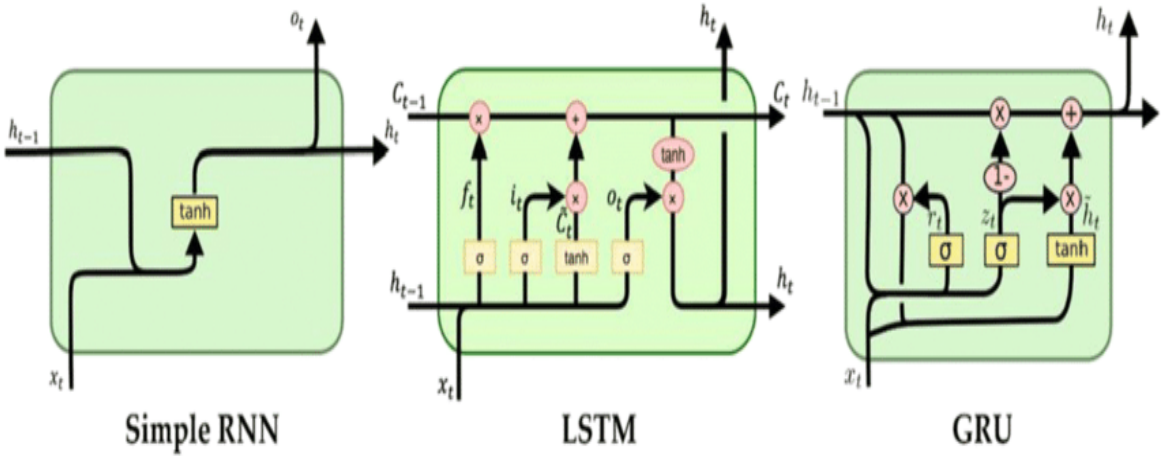


Figure 2.12: GRU vs RNN vs LSTM Unit Design

obtained from figures 2.12 and 3.3.

LSTM is capable of both short-term and long-term forecasting due to the memory improvement of the architecture from simple RNNs and GRUs. Depending on the number of time steps we need to forecast, we may defined a general LSTM neural network architecture with an equivalent number of dense LSTM layers.

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [C_{t-1}, h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \cdot [C_{t-1}, h_{t-1}, x_t] + b_i) \\
o_t &= \sigma(W_o \cdot [C_t, h_{t-1}, x_t] + b_o)
\end{aligned}
\tag{2.17}
$$

2.5.6 Hybrid Models: Neural Prophet

The Neural Prophet Model typically adopts many concepts from the Facebook Prophet model such as the linear trend and seasonality modeling with fourier series. Its goal is to bridge the explainability gap between deep learning models and explainable models like facebook prophet. Other components of the NeuralProphet model are auto-regression, special events, future regressors, and lagged regressors. Forecast regressors are external variables with known future values, while lagged regressors are those which only have observed values.

The bridge between deep learning and facebook prophet is in modeling of the regressors. Auto-regression (for forecast regressors) is handled using an implementation of AR-Net, an Auto-Regressive Feed-Forward Neural Network for time series while lagged regressors are also modelled using separate Feed-Forward Neural Networks. FFNNs have been explained in section 2.5.5 and demonstrated in figure 2.11 as a multi-layer perceptron.

The unique aspect of NeuralProphet is that both the future regressors ($\tilde{y}_{t+1}$, $\tilde{y}_{t+2}$, $\cdots$,$\tilde{y}_{t+n}$) and lagged regressors ((y_{t-n}, y_{t-n+1}, $\cdots$,y_{t-1}) are both utilized to estimate $\hat{y}_t$

Chapter 3

Research Methodology

3.1 Implementation Frameworks and Tools

In this research, Python was used for implementing and visualizing the results of the machine learning models while MATLAB was used for comparative analysis and some visualizatons. The APIs used, included numpy, pandas, scipy, prophet, NeuralProphet and tensorflow. We also used the streamlit library in python to build the dashboard.

3.2 Datasets and Preprocessing

We utilize daily and cumulative Covid-19 time series data for Ghana to fit and make forecasts with machine learning models. The data for the regional cases is collected from the Humdata website while the dataset for the national cases is obtained from Our World in Data.

The `.csv` file containing the collected data for the national cases also consists of records from other nations and states. We must therefore process this data to extract the daily and cumulative cases for Ghana alone. Figure 3.1 showcases the preprocessing stage to obtain a clean sequences of infected cases with their respective dates.

Humdata has records of the number infected cases for each of the 16 regions in Ghana. We obtain this dataset and restructure it using the pandas library. (see figure 3.2).

For the regional dataset, 70% is reserved for training the models while 30% is used for forecasting. For the National cases, we have over 840 datapoints. We utilized 750 of these data

```
import pandas as pd
import numpy as np
from prophet import Prophet
import matplotlib.pyplot as plt
df = pd.read_csv("https://raw.githubusercontent.com/njininathan/covid-forecasting/master/time_series_covid19_confirmed_global.csv",index_col =
df = df.drop(["Lat","Long","Province/State"],axis=1)
```

```
[ ] dates = list(df.columns)
    data_cumm = list(df.loc["Ghana"])
    gh_daily = list(np.diff(data_cumm)) # 3 compensates for the initial 3 recorded cases
    gh_data = pd.DataFrame(list(zip(dates, gh_daily,data_cumm)))
    gh_data.columns = ["ds","y","y_cumm"]
    gh_data = gh_data.drop(np.arange(51)).reset_index(drop=True)
    gh_data.head(3)
```

	ds	y	y_cumm
0	3/13/20	3	0
1	3/14/20	3	3
2	3/15/20	0	6

```
[ ] gh_data[["ds","y"]].plot()
```

Figure 3.1: National Data Preprocessing

```
# import and preprocess data for all regions
# load data
cases_ghana = read_csv("https://raw.githubusercontent.com/njininathan/covid-forecasting/master/cases_ghana.csv")
# Preprocessing of data
cases_ghana['cumulative_cases'] = abs(cases_ghana['cumulative_cases'])
cases_ghana['cases'] = abs(cases_ghana['cases'])
all_regions = cases_ghana.name.unique()

#all_regions['cases']
region_daily_cases_dict = dict()
region_cumm_cases_dict = dict()
for region in all_regions:
    region_daily_cases_dict[region] = cases_ghana.query('name==@region')[['date','cases']].reset_index(drop=True)
    region_daily_cases_dict[region].columns = ['ds','y']

    region_cumm_cases_dict[region] = cases_ghana.query('name==@region')[['date','cumulative_cases']].reset_index(drop=True)
    region_cumm_cases_dict[region].columns = ['ds','y']

# print example
region_daily_cases_dict["Greater Accra Region"].tail(3)
```

	ds	y
496	2022-07-03	291
497	2022-07-07	177

Figure 3.2: Regional Level Data Import and Preprocessing

points for training all the models for both cumulative and daily cases. The rest of the data points are used for forecasting.

3.3 Python Implementation of Facebook Prophet

We implement the facebook prophet model for both the logistic and linear trend models. The logistic trend model has an aspect of "cheating" to it as it requires a cap value to be set for the future forecast. The maximum future value is unknown and it can only be guessed.

In the implementation of the future forecast, we use `future_cap = 169000` to set the limit of the forecast. We also observe that the logistic growth model provides us with a prophet upper and lower bound output, `y_upper` and `y_lower` that are much more narrower due to a better confidence level set from the logistic model.

On the other hand, we would prefer using the linear trend model as it forecasts by itself without the need of making guesses about the forecast behaviour. This fundamentally obeys the rule of forecasting which entails making an inference on unseen and out of sample range data. The linear prophet model provides a much wider estimate for `y_upper` and `y_lower`.

3.4 LSTM Neural Network Architecture

We implement 3 separate LSTM neural network architectures. The model consists of 2 LSTM layers, each with 50 multiple parallel LSTM units, structurally identical but each eventually "learning to remember" some different thing. We also train the same for `num_of_forecast_steps` units. The model takes a much more significant amount of time to train for higher time steps. Smaller versions of this architectures have been implemented in literature but usually for a 1-step forecast. We do this for a 90-day, 150-day and 300-day forecast. The code is openly available on github and a saved version of the trained models is available upon request. For the daily cases, we train the same number of models. We also factor out the zeros and do thesame training.

All of the trained models show useful forecasts as shown in figures 4.3d 4.3b, 4.3a. The analysis of the forecasts in comparison with reasonable expectations and demographic information is discussed in chapter 4.

3.4.1 Keras Implementation of LSTM

The input data is preprocessed and the scaled between 0 and 1 using a `MinMaxScaler` from the scipy libarary. The input size is set to 120 data points (window size) and the forecast steps are set. The data is therefore structured to a 120×1 input array and fed into the LSTM neural network which is designed to accept this specific input size.

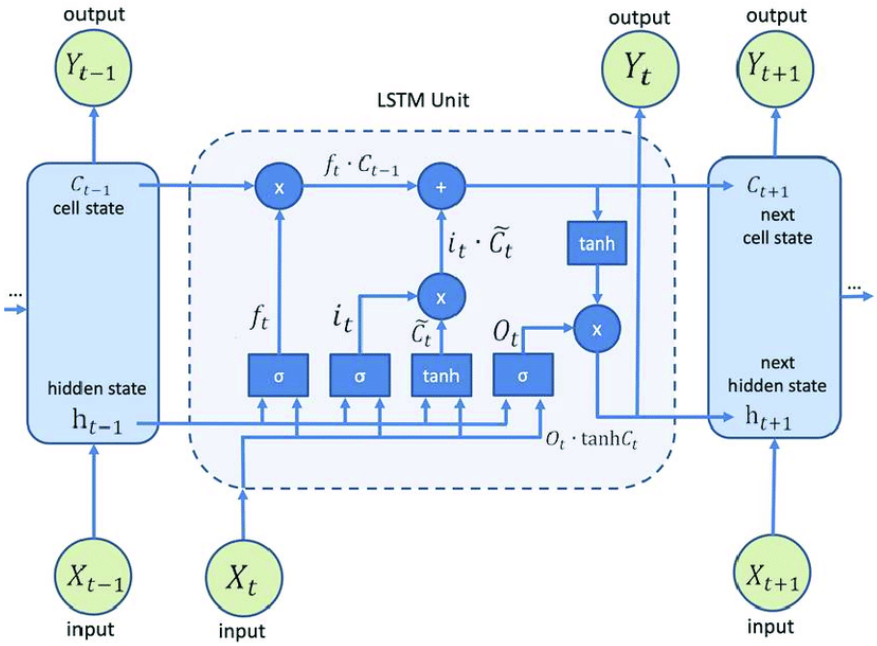


Figure 3.3: Single LSTM Unit

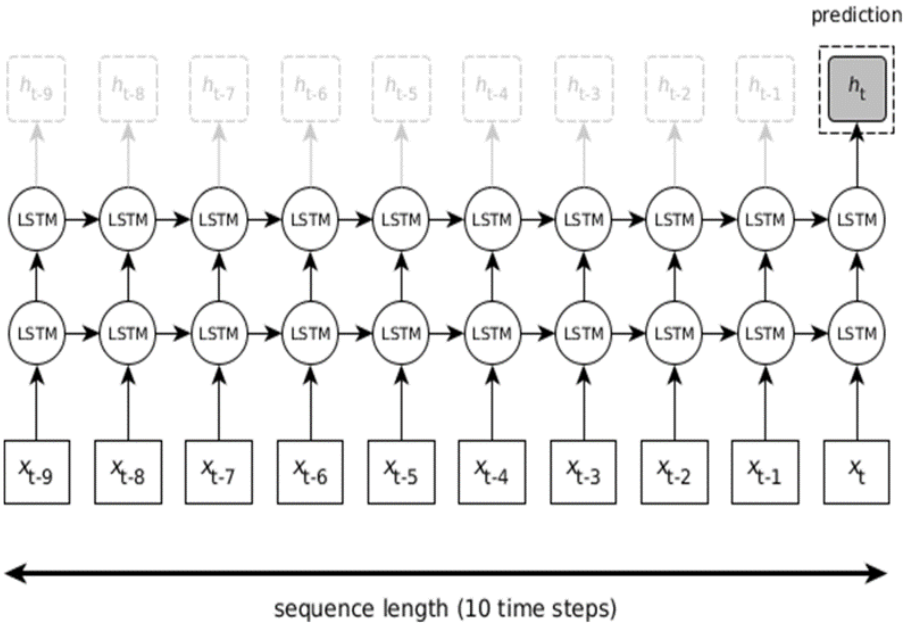


Figure 3.4: LSTM Architecture, implemented for $t = 90, 150, 300$,

In training the model, the `.Sequential()` object is used to initialize the model. The LSTM with 50 and `number_of_forecasts` dense layers, the models are trained for 360 epochs to ensure convergence of the loss function as shown in figure 3.5. The `number_of_forecasts` represents the number of time-step forecasts we select (90, 150 and 300).

With the model trained for the specified number of epochs and loss convergence observed, we can therefore rescaled the output from the LSTM forecast layers and plot them as shown in

section 4.2 and its related figures.

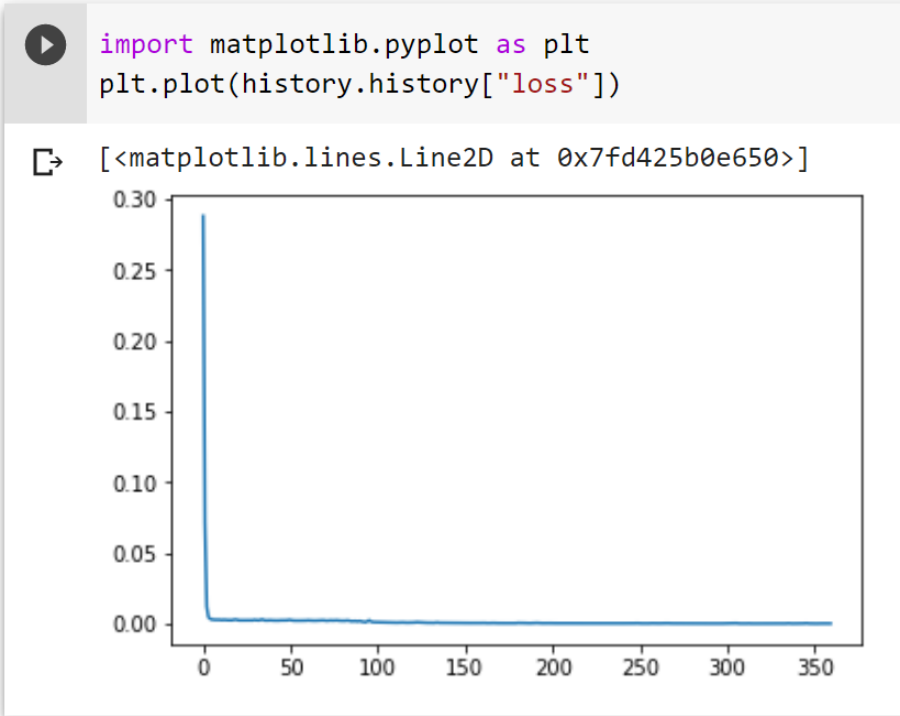


Figure 3.5: Minimized Loss for 360 epochs of training

3.5 Implementation of Neural Prophet

Implementing the Neural Prophet model, we used the NeuralProphet library which is based on Pytorch and Stan. This model is designed to be automated with the number of epochs for optimizing the lagged parameters and the AR-Net are designed to be limited once their loss functions converge. These are generally small models and our implementation showcases that most of the models are trained for epochs between 100 and 300 epochs before convergence.

We utilized the `NeuralProphet()` object and the `.fit(data)` method of the NeuralProphet API to train the model. All other parameters such as AR-Net parameters, FFNN size and number of regressors are inbuilt while others are automatically computed from the data as explained in section 2.5.6. The evaluation of the model as compared to other trained models is seen in table 4.1 in section 4.7 for all the models at the National Level.

The evaluation metrics of neural prophet include SmoothL1Loss, MAE, RMSE and RegLoss. These are shown in table 3.1 and the minimization of the metrics can be observed for 212 epochs

	SmoothL1Loss	MAE	RMSE	RegLoss
0	1.491633	309901.908208	365343.911414	0.0
1	1.360393	288236.806091	342098.865113	0.0
2	1.210593	263390.476239	314399.556612	0.0
3	1.031154	233163.243536	279377.342026	0.0
4	0.816925	196421.404924	239113.367430	0.0
...	...	...	...	...
208	0.000053	1239.861891	1595.521234	0.0
209	0.000052	1222.928644	1583.774331	0.0
210	0.000052	1222.569856	1581.974095	0.0
211	0.000052	1219.610959	1593.373291	0.0
212	0.000051	1219.074339	1581.173339	0.0

Table 3.1: Training of Neural prophet and loss minimization for 212 epochs

of training.

3.6 Performance Metrics for Prediction and Forecasting

The APIs provide the model metrics such as the loss, accuracy, smoothl1loss, RMSE and MAPE for evaluating the models. Diverse parameters are used for each of the models on the training dataset. However, our goal is to forecast the spread and compare it with the ground truths. The metrics used for evaluating the forecast accuracy are the MAPE, MSE and RMSE. We do this by evaluating the forecast on a 90-day scale for the National Cumulative Cases. We specific dataset since there is greater correlation and less randomness in the data points.

In addition, several visualizations of the forecast and predictions are made to see which models make the most sense in visual expectations. This is is extremely important aspect and relevant as the above mentioned parameters are not sufficient for providing genuinely reasonable forecasts since they are just mathematical metrics - they are still however still useful in making a model choice for a forecast.

We are also concerned about choosing models that are fast, convenient to use and that can easily scale without needed several minutes to hours of training. We realize that in section 4.4 that prophet in preferred, whatever the evaluation criterion.

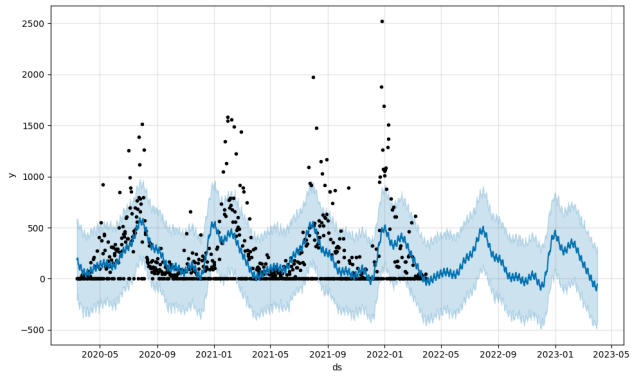
Chapter 4

Results and Discussion

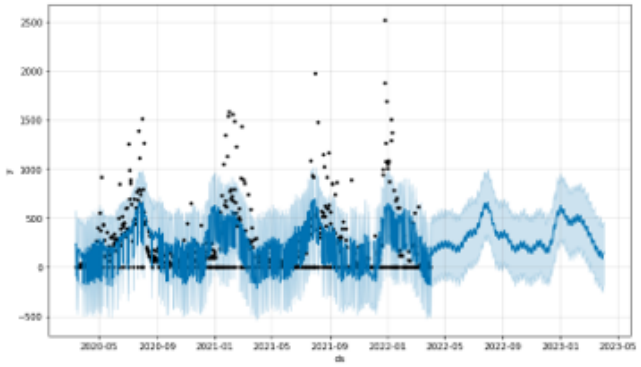
4.1 Prophet Fitting and Forecasts

Figure 4.1 shows the daily projections for the case where zeros are factored as holidays (figure 4.1b and where zeros are not factored figure 4.1a. We observe that the trend and seaonality of the forecasts is similar, however, the projections do not hit the zero point in the forecast. This is because the holidays in the future are unknown since days with zeros recorded are not predefined. In using 4.1b for analysis, days with zeros recorded in that future should be used to replace the forecasted values before decisions are made.

In making forecasts for the cumulative cases, both the logistic trend and linear trend functions are used as shown in figure 4.2. We would prefer using the linear prophet model as explained in section 3.3. The logistic growth model is only suitable for modeling the cumulative cases and not the daily cases.

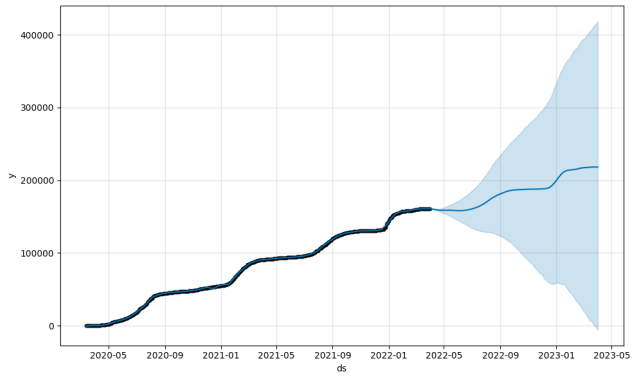


(a) Daily Cases with No Holidays (Original data)

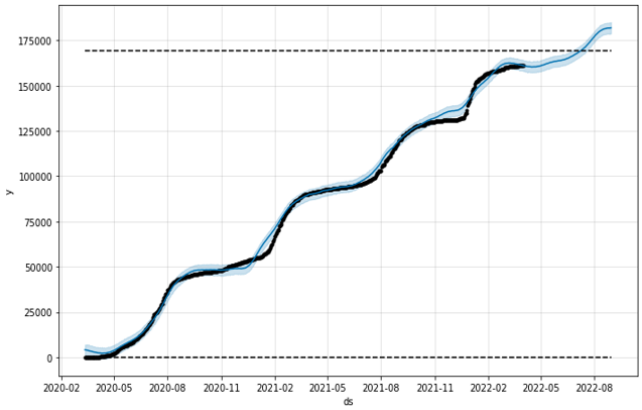


(b) Zeros factored as holidays

Figure 4.1: National COVID-19 Daily Forecasting with Prophet



(a) Forecasting with Linear Prophet



(b) Forecasting with Logistic Prophet

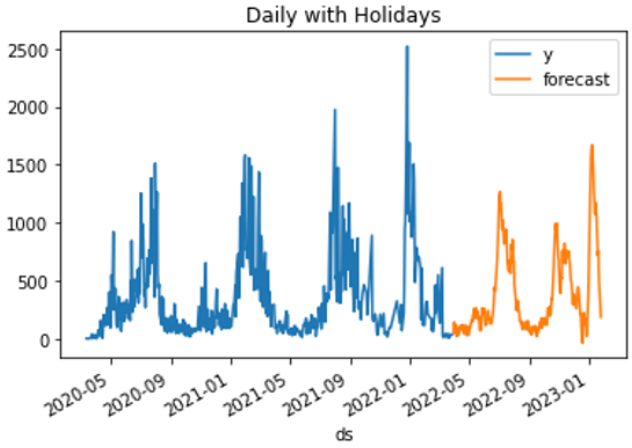
Figure 4.2: National Cumulative COVID-19 Forecasting with Prophet Models

4.2 Time Sequence LSTM Forecasting

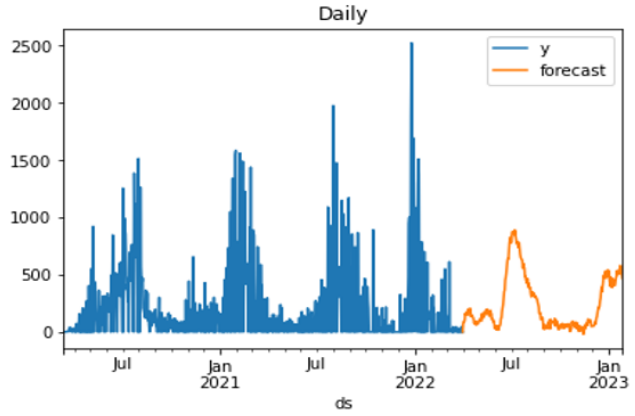
With LSTM forecasting, we can observe the fitting when zeros are removed in the preprocessing and when they are allowed in figures 4.3a and 4.3b. We can observe that LSTM provides an excellent trend forecast which perfectly fits demographic expectations in both cases. A critical observation is that the forecast when removing and keeping the zeros are significantly different in the seasonality aspects. Both cases are however still useful for having expectations are they closely align with forecasts from prophet in figure 4.1

Forecasts are also made for the national cumulative cases. The model is run a few times for a useful observation of the forecast to be obtained. It is then saved for future use to enable reproducibility. The reproducibility issues comes about since our LSTM model consists over of 35,190 trainable parameters. We include figure 4.3c to showcase a 90-day step prediction from

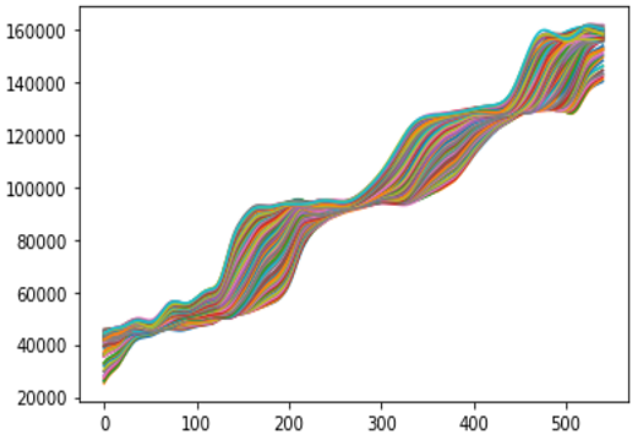
each of the dense layers, for one of the LSTM models built for cumulative forecasts. Every layer ensures to make a forecasts that is as close to the known label as possible during the training.



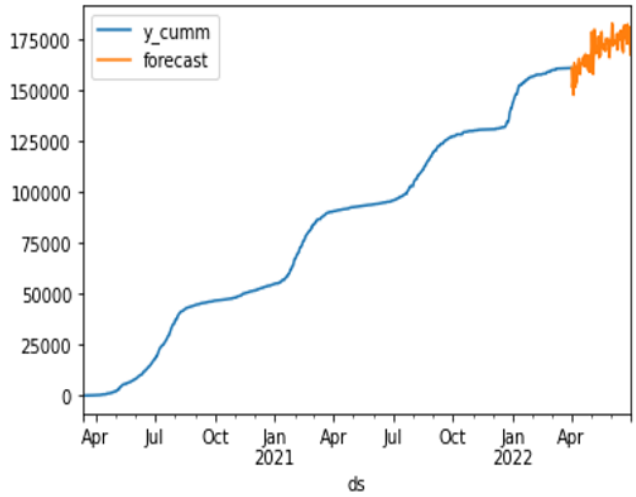
(a) LSTM Forecast with Zeros as Holidays



(b) LSTM Forecast with No Holidays



(c) LSTM Dense Layer Outputs (Predictions)



(d) LSTM Cumulative Forecast

Figure 4.3: Using LSTM for COVID-19, National Level Forecasting

4.3 Neural Prophet Prediction and Forecasting

We showcase the neural prophet forecast and predictions in figures 4.5 and 4.5a. Unfortunately, neural prophet does not perform well on daily forecasts, hence we do not display it here. We can notice that the prophet forecasts for the national cumulative cases are a lot higher compared to prophet and LSTM forecasts.

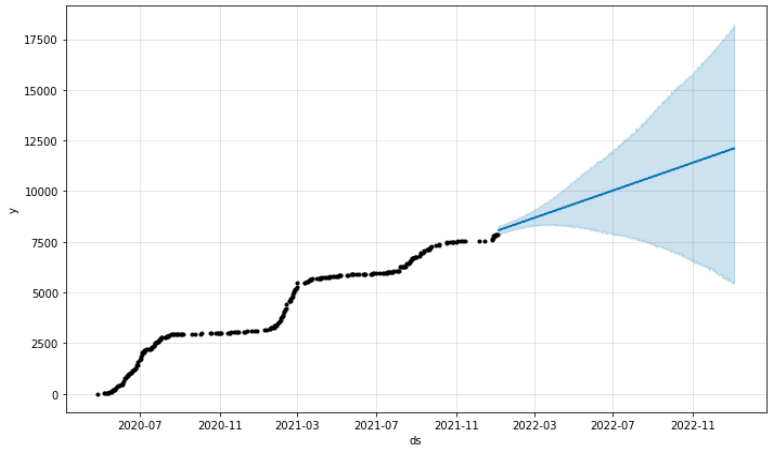
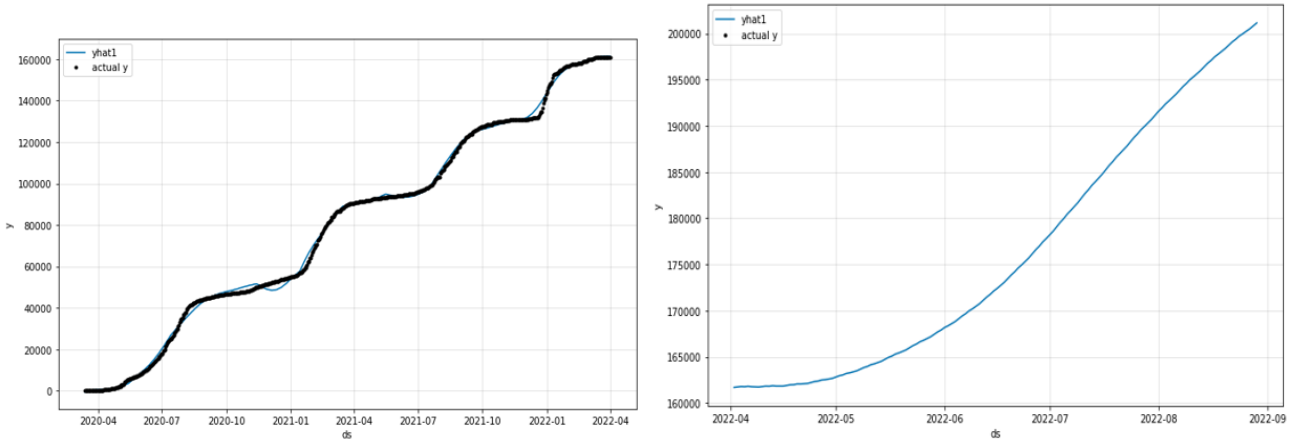


Figure 4.4: Estimation of Future COVID-19 Cases for Western Region (Neural Prophet)



(a) Neural Prophet Fitting (Prediction)

(b) Plot of Neural Prophet Forecast

Figure 4.5: National Cumulative COVID-19 Forecast using Neural Prophet

4.4 Performance of Models for a 90-day Forecast

Due to the limited amount of data and the need to split into training and test (forecast) samples, we will only perform analysis for a 90-day forecast on the National Cases. We observe that for the forecast, the prophet model performs best in terms of the measurement metrics and other expected desires from the forecast, as shown in table 4.1.

	MAPE	MSE	RMSE	Forecast Consistency (linearity)
Linear Prophet	0.0176	9.6299e+6	3.1032e+3	★★★★★
LSTM	0.0231	1.9633e+07	4.4309e+03	★
NeuralProphet	0.0287	5.9154e+03	3.4992e+07	★★★

Table 4.1: Comparison of model performance for a 90-day forecast

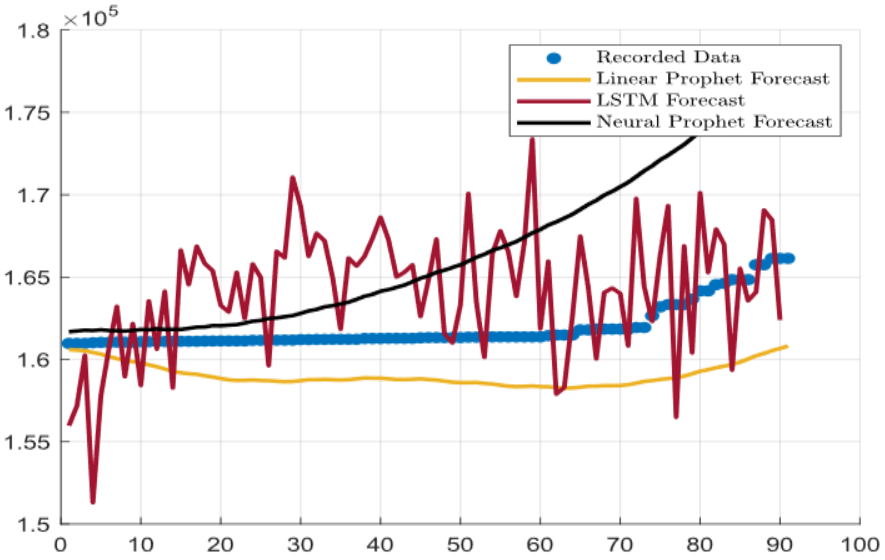


Figure 4.6: Comparison of model performance for a 90-day forecast

4.5 Regional Level Fitting and Forecasting

To estimate the expected number of future cases for each region, we may take 2 approaches based on the scenario. The second approach describes a regional head's perspective to projecting the expected number of cases while the first approach describes the state's perspective. It is important to state that while approach 1 and 2 may not yield exactly the same estimates for the regional cases, the 2 approaches are suitable and both give room for reasonable estimates and their outputs can be combined in a compromise towards disease control.

4.5.1 Approach 1

In approach 1, we consider a decision maker in charge of COVID-19 control in the entire Ghana. His/her interest is to understand the spread of the disease in the entire country and to make a forecast nation-wide. By estimating the percentage distribution of the spread across different regions, they can therefore ensure fairness and impact in the distribution of resources such as hospital staff, testing kits, protection equipment and nose masks. This is demonstrated in figure 4.7 below.

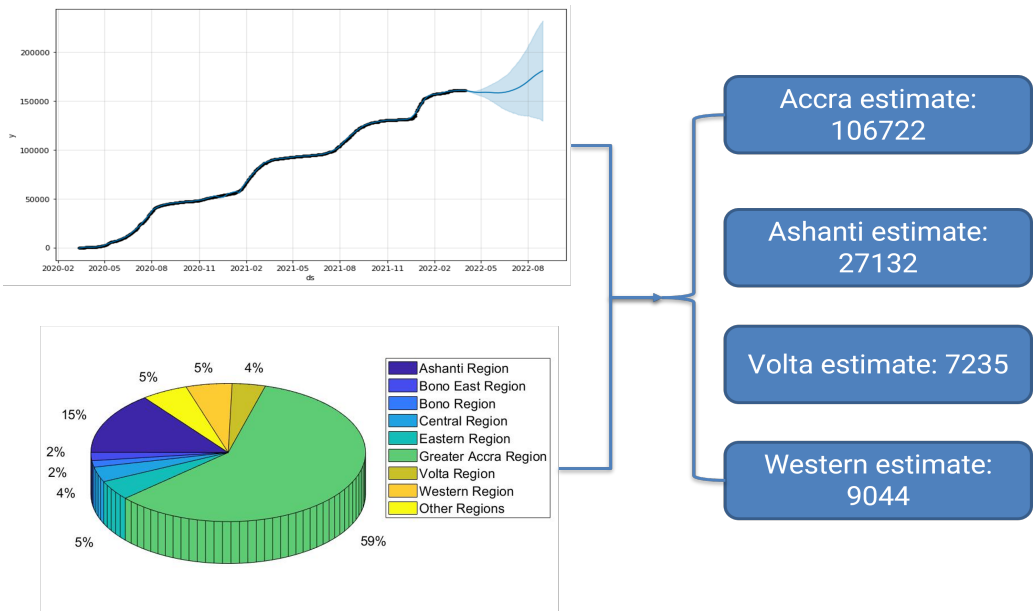
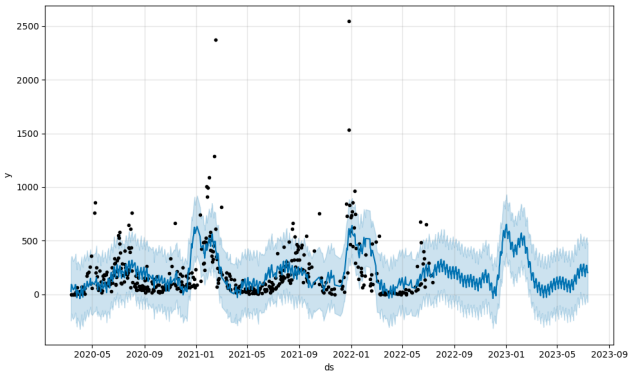


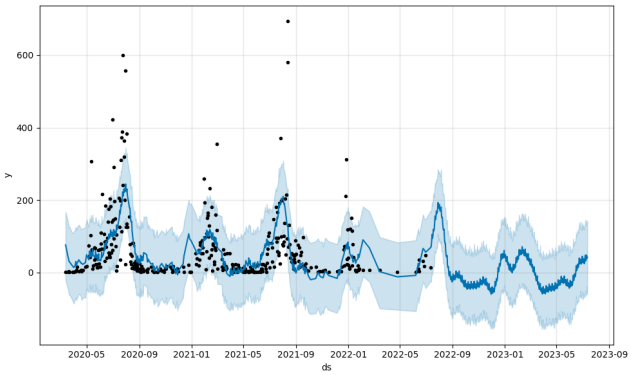
Figure 4.7: Approach 1 to Estimating COVID-19 spread in Regions

4.5.2 Approach 2

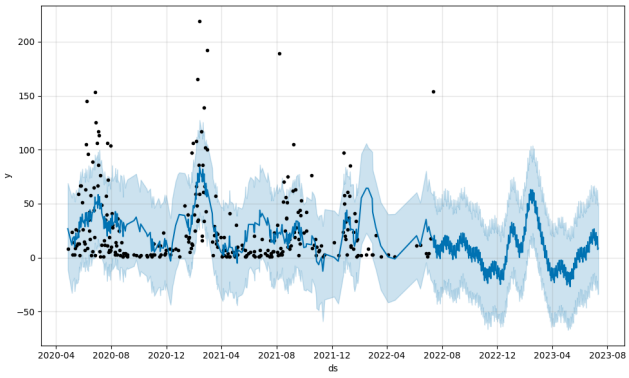
In the second approach, we propose that a regional head would make a forecast and determination of the needs of the region based on a forecast from regional data instead of a percentage distribution from the national cases. Essentially, the regional head is only concerned about the spread in his region and his decisions or requests from the central government should be based on analysis from his own region. It is therefore needful to do forecasts based on regional perspective as well. Figures 4.9 and 4.8 demonstrates the spread for 4 select region at the cumulative level and for the active cases. We can clearly observe that the forecasts meet demographic expectations with rising cases between June to August and the forecasts also increase reasonably with time. The daily forecasts also show reasonable seasonality and the number of cases generally stay low as more vaccinations are taking place. Outliers are however observed when cases go below the zero point as see in figure 4.8 which can be quickly assumed to be zeros.



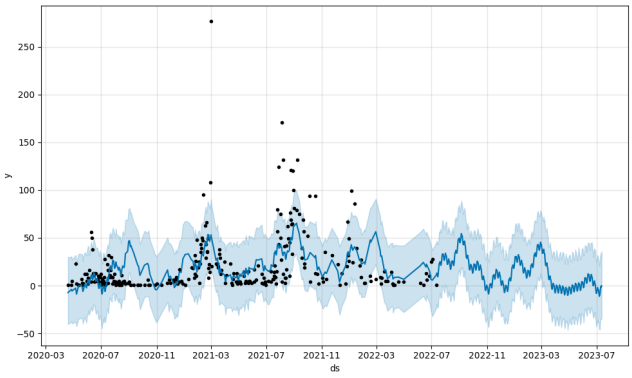
(a) Accra Regional Case Fitting Forecast



(b) Ashanti Regional Case Fitting and Forecast

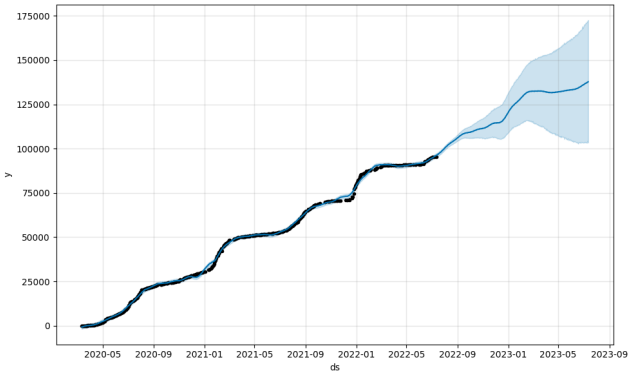


(c) Western Regional Case Fitting and Forecast

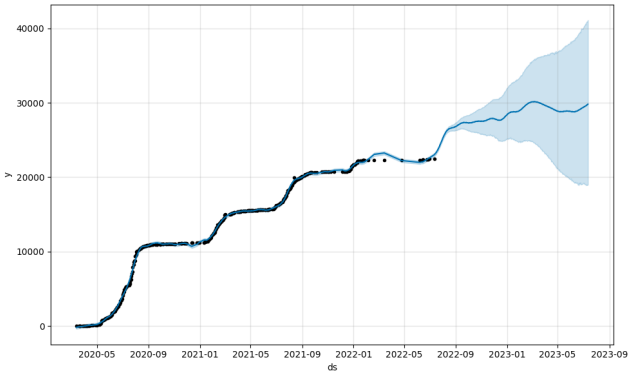


(d) Volta Regional Case Fitting and Forecast

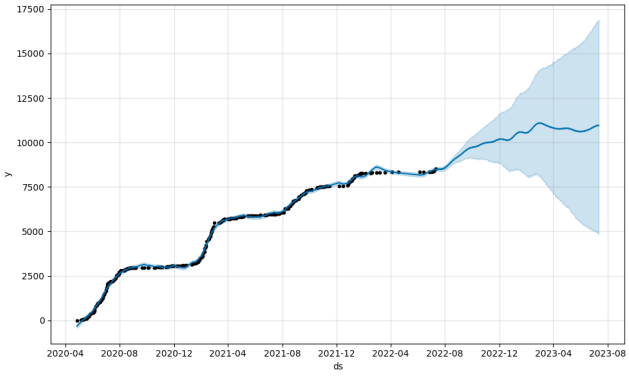
Figure 4.8: Approach 2: Forecasting Ghana's Regional Daily Cases Independently



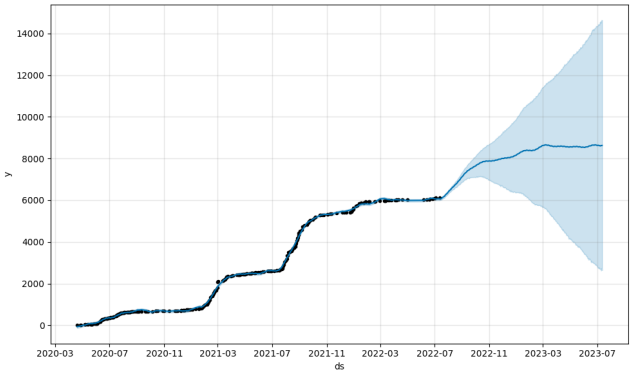
(a) Accra Regional Cumulative Case Forecast



(b) Ashanti Regional Cumulative Cases



(c) Western Regional Cumulative Case Forecast



(d) Volta Regional Cumulative Cases

Figure 4.9: Approach 2: Forecasting Ghana's Regional Cumulative Cases Independently

4.6 Use Case: Data-Driven Decision Making

We demonstrate a dashboard prototype that can be used to visualize the forecasts using the prophet model. The dashboard showcases forecasts for both national and regional forecasts. This is done for both cumulative and daily number of cases.

It is worth noting that forecasting for the cumulative cases turns out to be more useful in decision making since the goal is to estimate the total future cases. The daily cases provide us a nice visual of the seasonality aspects and situations of expected increases which could be due to a new variant, cold weather or high mobility and interaction of individuals.

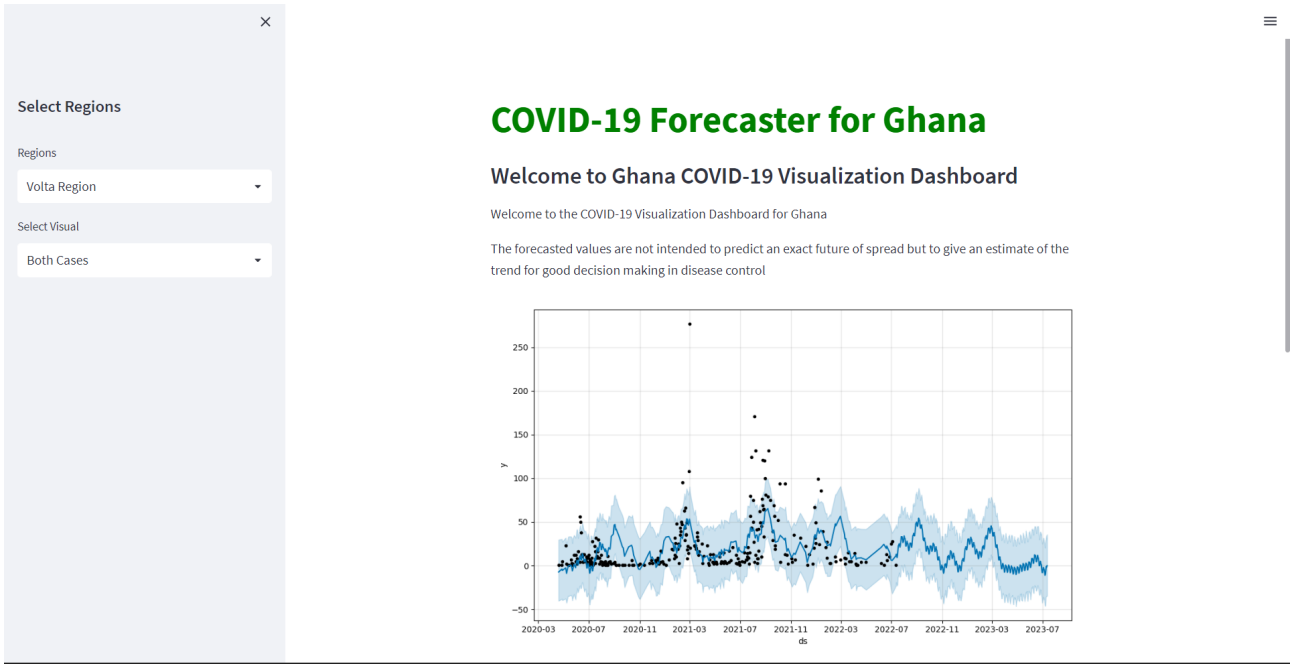


Figure 4.10: A Simple COVID-19 Dashboard for Ghana and its Regions

Chapter 5

Conclusion and Recommendations

5.1 Conclusion

In this study, machine learning models were implemented to determine the expected spread trend and number of cases in Ghana. The suggested solution performs better than conventional ones like the SIR, mathematical models, etc. Our solution outperforms the conventional methods of forecasting after the models have been trained with the data sets. The suggested solution has the power to completely alter the game, particularly in the field of health and governance. By putting the suggested solutions into practice, the overall objective of predicting the scope of the epidemics spread can be accomplished. This would result in advantages like:

1. Control and preparedness;
2. Allocation of health resources to better suite the needs;
3. Better Management decisions;
4. Development of COVID-19 clinical management training resources for health workers based on the most up to date forecast;
5. Significantly reducing the associated cost in supplies and staff redundancy.

We also observed that the same model architecture can be used to fit all sort of datasets, large and small. This is done for the regional and national, cumulative and daily active cases alike.

5.2 Recommendations

We observe that all the models forecast rises in the number of cases or the prediction of a new variant due to the seasonality factors at certain times of the year. High record of cases are observed in June to August and December to February. This may be due to changes in the climatic condition or season (cold weather) or high number of international travels. We can make several recommendations based on these observations to mitigate this situation. Some of these include:

1. Increasing the use of vaccines;
2. Increasing the use of nose marks and hand washing in cold seasons;
3. Provide facilities for respiratory hygiene and hand hygiene;
4. Cleaning the environment and disinfecting frequently touched surfaces;
5. Regular review and updating of safety protocols;
6. Facilitatation of contact tracing and isolation of individuals suspected of having COVID-19, especially at airports.

These should particularly be enforced in seasons with high number of expected cases.

5.3 Future Work and Open Problems

While this research attempts to use the best models to predict and forecast the spread of an epidemic (COVID-19) on a daily and cumulative level, there are several improvements and other open problems that can be made. While LSTMs suffer reproducibility issues, one may utilize an already trained model alongside a statistical mode such as moving average (MA) model to produce a more linear and smoothed output.

For open problems, dataset quality remains an important issue to solve. Time series data for all aspects or compartments of the disease should be recorded daily to ensure consistency in data. Another open problem is using multivariate data such as time series records of death

cases, recovered, hospitalized and infected inviduals to make a forecast of infected individuals. Multiple variables are known to contain more information, hence multivariate models can encode more accurate information and provide better forecast.

Also, creating better hybrid models by investigating a combination of network-based and machine learning-based models would be a novel area to look into.

Bibliography

[1] Ammar Abasi et al. "A novel hybrid multi-verse optimizer with K-means for text documents clustering". In: *Neural Computing and Applications* 32 (Dec. 2020). DOI: 10.1007/s00521-020-04945-0.

[2] Ammar Kamal Abasi et al. "Link-Based Multi-Verse Optimizer for Text Documents Clustering". In: *Appl. Soft Comput.* 87.C (Feb. 2020). ISSN: 1568-4946. DOI: 10.1016/j.asoc.2019.106002. URL: https://doi.org/10.1016/j.asoc.2019.106002.

[3] Ansari Ahmar and R. Rusli. "Will Covid-19 cases in the World reach 4 million? a forecasting approach using SutteARIMA". In: *JOIV : International Journal on Informatics Visualization* 4 (Sept. 2020), p. 159. DOI: 10.30630/joiv.4.3.389.

[4] Achyuth Ajith et al. "A Study on Prediction and Spreading of Epidemic Diseases". In: *2020 International Conference on Communication and Signal Processing (ICCSP)*. 2020, pp. 1265–1268. DOI: 10.1109/ICCSP48568.2020.9182147.

[5] Yan Bai et al. "Presumed Asymptomatic Carrier Transmission of COVID-19". In: *JAMA* 323 (Feb. 2020). DOI: 10.1001/jama.2020.2565.

[6] Amos Bairoch et al. "The Universal Protein Resource (UniProt)". In: *Nucleic Acids Research* 36 (Jan. 2008). DOI: 10.1093/nar/gkm895.

[7] Giuseppe C. Calafiore, Carlo Novara, and Corrado Possieri. "A Modified SIR Model for the COVID-19 Contagion in Italy". In: *2020 59th IEEE Conference on Decision and Control (CDC)*. 2020, pp. 3889–3894. DOI: 10.1109/CDC42340.2020.9304142.

[8] Manuel Castillo-Cara and M Carrillo-Larco. "Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach". In: (2020). DOI: 10.12688/wellcomeopenres.15819.3.

[9] Vinay Ch and Lei Zhang. "Time series forecasting of COVID-19 transmission in Canada using LSTM networks". In: *Chaos, Solitons Fractals* 135 (May 2020), p. 109864. DOI: 10.1016/j.chaos.2020.109864.

[10] Gerardo Chowell. "Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts". In: *Infectious Disease Modelling* 2.3 (2017), pp. 379–398. ISSN: 2468-0427. DOI: https://doi.org/10.1016/j.idm.2017.08.001.

[11] Dmytro Chumachenko et al. "Forecasting of COVID-19 Epidemic Process by Support Vector Machine Method in Ukraine and Neighboring Countries". In: *2021 IEEE 2nd KhPI Week on Advanced Technology (KhPIWeek)*. 2021, pp. 589–594. DOI: 10.1109/KhPIWeek53812.2021.9569968.

[12] Abdelkader Dairi et al. "Comparative Study of Machine Learning Methods for COVID-19 Transmission Forecasting". In: *J. of Biomedical Informatics* 118.C (June 2021). ISSN: 1532-0464. DOI: 10.1016/j.jbi.2021.103791. URL: https://doi.org/10.1016/j.jbi.2021.103791.

[13] Thomas Davenport and Ravi Kalakota. "The potential for artificial intelligence in healthcare". In: *Future Hospital Journal* 6 (June 2019), pp. 94–98. DOI: 10.7861/futurehosp.6-2-94.

[14] Shawni Dutta and Samir Kumar Bandyopadhyay. "Machine Learning Approach for Confirmation of COVID-19 Cases: Positive, Negative, Death and Release". In: *medRxiv* (2020). DOI: 10.1101/2020.03.25.20043505.

[15] Ebaa Fayyoumi, Sahar Idwan, and Heba AboShindi. "Machine Learning and Statistical Modelling for Prediction of Novel COVID-19 Patients Case Study: Jordan". In: *International Journal of Advanced Computer Science and Applications* 11 (Jan. 2020). DOI: 10.14569/IJACSA.2020.0110518.

[16] Navid Feroze. "Forecasting the patterns of COVID-19 and causal impacts of lockdown in top five affected countries using Bayesian Structural Time Series Models". In: *Chaos, Solitons Fractals* 140 (2020), p. 110196. ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2020.110196.

[17] Jennifer A. Gilbert et al. "Probabilistic uncertainty analysis of epidemiological modeling to guide public health intervention policy". In: *Epidemics* 6 (2014), pp. 37–45. ISSN: 1755-4365. DOI: https://doi.org/10.1016/j.epidem.2013.11.002.

[18] Tony L Goldberg and Nathan D Wolfe. "Infectious Diseases in Primates: Behavior, Ecology and Evolution. Oxford Series in Ecology and Evolution. By Charles L Nunn and Sonia Altizer." In: *The Quarterly Review of Biology* 82.3 (2007), pp. 289–289. DOI: 10.1086/523174.

[19] Mohamed Hawas. "Generated Time-series Prediction Data of COVID-19s Daily Infections in Brazil by Using Recurrent Neural Networks". In: *Data in Brief* 32 (Aug. 2020), p. 106175. DOI: 10.1016/j.dib.2020.106175.

[20] Yongqun He, Zuoshuang Xiang, and Harry Mobley. "Vaxign: The First Web-Based Vaccine Design Program for Reverse Vaccinology and Applications for Vaccine Development". In: *Journal of biomedicine biotechnology* 2010 (Jan. 2010), p. 297505. DOI: 10.1155/2010/297505.

[21] Herbert W. Hethcote. "The Mathematics of Infectious Diseases". In: *SIAM Review* 42.4 (2000), pp. 599–653. DOI: 10.1137/S0036144500371907.

[22] Chaolin Huang et al. "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China". In: *The Lancet* 395 (Jan. 2020). DOI: 10.1016/S0140-6736(20)30183-5.

[23] Kai-Qian Kam et al. "A Well Infant with Coronavirus Disease 2019 (COVID-19) with High Viral Load". In: *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 71 (Feb. 2020). DOI: 10.1093/cid/ciaa201.

[24] P.A Karthick, Diptasree Maitra, and Ramakrishnan Swaminathan. "Surface electromyography based muscle fatigue detection using high-resolution time-frequency methods and

machine learning algorithms". In: *Computer methods and programs in biomedicine* 154 (Feb. 2018), pp. 45–56. DOI: 10.1016/j.cmpb.2017.10.024.

[25] Mr Kavadi et al. "Partial derivative Nonlinear Global Pandemic Machine Learning prediction of COVID 19". In: *Chaos, Solitons Fractals* 139 (June 2020), p. 110056. DOI: 10.1016/j.chaos.2020.110056.

[26] Gülşen Keskin, Şenay Doğruparmak, and Kadriye Ergün. "Estimation of COVID-19 patient numbers using artificial neural networks based on air pollutant concentration levels". In: *Environmental Science and Pollution Research* (May 2022). DOI: 10.1007/s11356-022-20231-z.

[27] Farhan Mohammad Khan et al. "Projecting the criticality of COVID-19 transmission in India using GIS and machine learning methods". In: *Journal of Safety Science and Resilience* 2.2 (2021), pp. 50–62. ISSN: 2666-4496. DOI: https://doi.org/10.1016/j.jnlssr.2021.05.001.

[28] Anuradha Khattar, Priti Rai Jain, and S. M. K. Quadri. "Effects of the Disastrous Pandemic COVID 19 on Learning Styles, Activities and Mental Health of Young Indian Students - A Machine Learning Approach". In: *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. 2020, pp. 1190–1195. DOI: 10.1109/ICICCS48265.2020.9120955.

[29] Dmitriy Klyushin and Kateryna Golubeva. "Novel Nonparametric Test for Comparing Machine Learning Models for COVID-19 Outbreak Prediction". In: *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*. Vol. 1. 2021, pp. 1–4. DOI: 10.1109/CSIT52700.2021.9648801.

[30] Jerry John Kponyo et al. "An Algorithm to Determine the Extent of an Epidemic Spread: A NetLogo Modeling Approach". In: *Engineering and Applied Sciences* (2021), pp. 66–69. DOI: 10.1109/RTEICT52294.2021.9574032.

[31] Liu Kui et al. "Clinical characteristics of novel coronavirus cases in tertiary hospitals in Hubei Province". In: *Chinese medical journal* 133 (Feb. 2020). DOI: 10.1097/CM9.0000000000000744.

[32] Askat Kuzdeuov et al. “A Network-Based Stochastic Epidemic Simulator: Controlling COVID-19 With Region-Specific Policies”. In: *IEEE Journal of Biomedical and Health Informatics* 24.10 (2020), pp. 2743–2754. DOI: 10.1109/JBHI.2020.3005160.

[33] Mazharul Islam Leon et al. “Predicting COVID-19 infections and deaths in Bangladesh using Machine Learning Algorithms”. In: *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. 2021, pp. 70–75. DOI: 10.1109/ICICT4SD50815.2021.9396820.

[34] Qun Li et al. “Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia”. In: *New England Journal of Medicine* 382.13 (2020). PMID: 31995857, pp. 1199–1207. DOI: 10.1056/NEJMoa2001316. eprint: https://doi.org/10.1056/NEJMoa2001316. URL: https://doi.org/10.1056/NEJMoa2001316.

[35] Zeyuan Liu et al. “Coronavirus Epidemic (COVID-19) Prediction and Trend Analysis Based on Time Series”. In: *2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID)*. 2021, pp. 35–38. DOI: 10.1109/AIID51893.2021.9456463.

[36] Yixiao Ma et al. “COVID-19 Spreading Prediction with Enhanced SEIR Model”. In: *2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*. 2020, pp. 383–386. DOI: 10.1109/ICAICE51518.2020.00080.

[37] Rishikesh Magar, Prakarsh Yadav, and Amir Barati Farimani. “Potential neutralizing antibodies discovered for novel corona virus using machine learning”. In: *Scientific Reports* 11 (Mar. 2021). DOI: 10.1038/s41598-021-84637-4.

[38] Lorenzo Mangoni and Marco Pistilli. “Epidemic Analysis of COVID-19 in Italy by Dynamical Modelling”. In: *SSRN Electron.J.* (Apr. 2020). DOI: 10.2139/ssrn.3567770.

[39] L. William Mary and S. Albert Antony Raj. “Machine Learning Algorithms for Predicting SARS-CoV-2 (COVID-19) – A Comparative Analysis”. In: *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*. 2021, pp. 1607–1611. DOI: 10.1109/ICOSEC51865.2021.9591801.

[40] Mohammad Masum et al. “Comparative study of a mathematical epidemic model, statistical modeling, and deep learning for COVID-19 forecasting and management”. In: *Socio-Economic Planning Sciences* 80 (2022), p. 101249. ISSN: 0038-0121. DOI: https://doi.org/10.1016/j.seps.2022.101249.

[41] Patricia Melin et al. “Analysis of Spatial Spread Relationships of Coronavirus (COVID-19) Pandemic in the World using Self Organizing Maps”. In: *Chaos, Solitons & Fractals* 138 (May 2020), p. 109917. DOI: 10.1016/j.chaos.2020.109917.

[42] Negar Memarian et al. “Multimodal Data and Machine Learning for Surgery Outcome Prediction in Complicated Cases of Mesial Temporal Lobe Epilepsy”. In: *Comput. Biol. Med.* 64.C (Sept. 2015), pp. 67–78. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2015.06.008.

[43] JC Miller and IZ Kiss. “Epidemic spread in networks: Existing methods and current challenges”. In: *Math Model Nat Phenom.* 2014, 9(2):4–42. DOI: 10.1051/mmnp/20149202.

[44] Shiva Moein et al. *Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan.* 2021. DOI: 10.1038/s41598-021-84055-6.

[45] Baishali Mullick et al. “Understanding Mutation Hotspots for the SARS-CoV-2 Spike Protein Using Shannon Entropy and K-Means Clustering”. In: *Comput. Biol. Med.* 138.C (Nov. 2021). ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2021.104915. URL: https://doi.org/10.1016/j.compbiomed.2021.104915.

[46] Khushboo Munir et al. “Cancer Diagnosis Using Deep Learning: A Bibliographic Review”. In: *Cancers* 11.9 (2019). ISSN: 2072-6694. DOI: 10.3390/cancers11091235. URL: https://www.mdpi.com/2072-6694/11/9/1235.

[47] Nikhil et al. “Polynomial Based Linear Regression Model to Predict COVID-19 Cases”. In: *2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT).* 2021, pp. 66–69. DOI: 10.1109/RTEICT52294.2021.9574032.

[48] Edison Ong et al. “COVID-19 Coronavirus Vaccine Design Using Reverse Vaccinology and Machine Learning”. In: *Frontiers in Immunology* 11 (July 2020), p. 1581. DOI: 10.3389/fimmu.2020.01581.

[49] Edison Ong et al. “Vaxign-ML: Supervised Machine Learning Reverse Vaccinology Model for Improved Prediction of Bacterial Protective Antigens”. In: *Bioinformatics (Oxford, England)* 36 (Feb. 2020). DOI: 10.1093/bioinformatics/btaa119.

[50] Yaohao Peng and Mateus Nagata. “An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data”. In: *Chaos, Solitons Fractals* 139 (June 2020), p. 110055. DOI: 10.1016/j.chaos.2020.110055.

[51] Fotios Petropoulos and Spyros Makridakis. “Forecasting the novel coronavirus COVID-19”. In: *PLOS ONE* 15.3 (Mar. 2020), pp. 1–8. DOI: 10.1371/journal.pone.0231236. URL: https://doi.org/10.1371/journal.pone.0231236.

[52] Gergő Pintér et al. “COVID-19 Pandemic Prediction for Hungary: A Hybrid Machine Learning Approach”. In: *Int. J.Adv.Comput.Sci.Appl.* 11 (May 2020), pp. 122–126. DOI: 10.14569/IJACSA.2020.0110518.

[53] J. Read et al. “Novel coronavirus 2019-nCoV (COVID-19): early estimation of epidemiological parameters and epidemic size estimates”. In: *Philosophical Transactions of the Royal Society B* 376 (1829 July 2021).

[54] Sandeep Reddy, John Fox, and Maulik Purohit. “Artificial intelligence-enabled healthcare delivery”. In: *Journal of the Royal Society of Medicine* 112 (Dec. 2018), p. 014107681881551. DOI: 10.1177/0141076818815510.

[55] Matheus Henrique Dal Molin Ribeiro et al. “Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil”. In: *Chaos, Solitons Fractals* 135 (2020), p. 109853. ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2020.109853. URL: https://www.sciencedirect.com/science/article/pii/S0960077920302538.

[56] Furqan Rustam et al. “COVID-19 Future Forecasting Using Supervised Machine Learning Models”. In: *IEEE Access* 8 (2020), pp. 101489–101499. DOI: 10.1109/ACCESS.2020.2997311.

[57] Alok Kumar Sahai et al. “ARIMA modelling and forecasting of COVID-19 in top five affected countries”. In: *Diabetes and Metabolic Syndrome: Clinical Research and Reviews* 14.5 (2020), pp. 1419–1427. ISSN: 1871-4021. DOI: https://doi.org/10.1016/j.dsx.2020.07.042.

[58] Andrew Senior et al. “Improved protein structure prediction using potentials from deep learning”. In: *Nature* 577 (Jan. 2020), pp. 1–5. DOI: 10.1038/s41586-019-1923-7.

[59] Vikas Sharma and Unnati Nigam. “Modeling and Forecasting of COVID-19 Growth Curve in India”. In: *Transactions of the Indian National Academy of Engineering* 5 (Sept. 2020), pp. 1–14. DOI: 10.1007/s41403-020-00165-z.

[60] Sima Siami-Namini and Akbar Siami Namin. “Forecasting Economics and Financial Time Series: ARIMA vs. LSTM”. In: *ArXiv* abs/1803.06386 (2018).

[61] Mohammad Khubeb Siddiqui et al. “Correlation Between Temperature and COVID-19 (Suspected, Confirmed and Death) Cases based on Machine Learning Analysis”. In: *Journal of Pure and Applied Microbiology* 14 (Apr. 2020). DOI: 10.22207/JPAM.14.SPL1.40.

[62] Tiago Tiburcio da Silva, Rodrigo Francisquini, and Mariá C.V. Nascimento. “Meteorological and Human Mobility Data on Predicting COVID-19 Cases by a Novel Hybrid Decomposition Method with Anomaly Detection Analysis: A Case Study in the Capitals of Brazil”. In: *Expert Syst. Appl.* 182.C (Nov. 2021). ISSN: 0957-4174. DOI: 10.1016/j.eswa.2021.115190. URL: https://doi.org/10.1016/j.eswa.2021.115190.

[63] Taylor SJ and Letham B. “Forecasting at scale”. In: *PeerJ Preprints* (2017), 5:e3190v2. DOI: https://doi.org/10.7287/peerj.preprints.3190v2.

[64] Aman Swaraj et al. “Implementation of stacking based ARIMA model for prediction of Covid-19 cases in India”. In: *Journal of Biomedical Informatics* 121 (2021), p. 103887. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2021.103887. URL: https://www.sciencedirect.com/science/article/pii/S1532046421002161.

[65] Noshin Tasnia, Shakik Mahmud, and M. F. Mridha. "COVID-19 Future Forecasting Tool: Infected Patients Recovery and Hospitalization Trends Using Deep Learning Models". In: (2021), pp. 1–6. DOI: 10.1109/ICSCT53883.2021.9642691.

[66] Ahmet Tekkeşin. "Artificial Intelligence in Healthcare: Past, Present and Future". In: *The Anatolian Journal of Cardiology* 22 (Oct. 2019). DOI: 10.14744/AnatolJCardiol.2019.28661.

[67] Mesut Toğaçar, Burhan Ergen, and Zafer Cömert. "COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches". In: *Computers in Biology and Medicine* (May 2020), p. 103805. DOI: 10.1016/j.compbiomed.2020.103805.

[68] Anuradha Tomar and Neeraj Gupta. "Prediction for the spread of COVID-19 in India and effectiveness of preventive measures". In: *Science of The Total Environment* 728 (Apr. 2020), p. 138762. DOI: 10.1016/j.scitotenv.2020.138762.

[69] Oskar Triebe et al. *NeuralProphet: Explainable Forecasting at Scale*. 2021. DOI: 10.48550/ARXIV.2111.15397.

[70] Jagadishwari V. "Time series Covid 19 Predictions with Machine Learning Models". In: *2021 Emerging Trends in Industry 4.0 (ETI 4.0)*. 2021, pp. 1–4. DOI: 10.1109/ETI4.051663.2021.9619334.

[71] Parth Wadhwa et al. "Predicting the Time Period of Extension of Lockdown due to Increase in Rate of COVID-19 Cases in India using Machine Learning". In: *Materials Today: Proceedings* 37 (Aug. 2020). DOI: 10.1016/j.matpr.2020.08.509.

[72] Peipei Wang et al. "Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran". In: *Chaos, Solitons Fractals* 140 (Aug. 2020), p. 110214. DOI: 10.1016/j.chaos.2020.110214.

[73] Qiuwei Wang. "Performance of Different Models of Machine Learning in Predicting the COVID-19 Pandemic". In: *2020 International Conference on Public Health and Data Science (ICPHDS)*. 2020, pp. 218–226. DOI: 10.1109/ICPHDS51617.2020.00050.

[74] Milind Yadav, Murukessan Perumal, and M Srinivas. "Analysis on novel coronavirus (COVID-19) using machine learning methods". In: *Chaos, Solitons & Fractals* 139 (2020), p. 110050. ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2020.110050.

[75] Jamileh Yousefi and Andrew Hamilton-Wright. "Characterizing EMG Data using Machine-Learning Tools". In: *Computers in Biology and Medicine* 51 (Aug. 2014). DOI: 10.1016/j.compbiomed.2014.04.018.

[76] Seid Miad Zandavi, Taha Hossein Rashidi, and Fatemeh Vafaee. "Dynamic Hybrid Model to Forecast the Spread of COVID-19 Using LSTM and Behavioral Models Under Uncertainty". In: *IEEE Transactions on Cybernetics* (2021), pp. 1–13. DOI: 10.1109/TCYB.2021.3120967.

[77] Abdelhafid Zeroual et al. "Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study". In: *Chaos, Solitons Fractals* 140 (2020), p. 110121. ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2020.110121.

[78] Nanning Zheng et al. "Predicting COVID-19 in China Using Hybrid AI Model". In: *IEEE Transactions on Cybernetics* 50.7 (2020), pp. 2891–2904. DOI: 10.1109/TCYB.2020.2990162.