APPLICATION OF MACHINE LEARNING APPROACHES FOR COVID-19 SPREAD PREADICTION IN ETHIOPIA

BY

SENEDU GEBEREMARIAM DEKEBO

OFFICE OF GRADUATE STUDIES
ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

ADAMA, ETHIOPIA
JANUARY 2021

Application of Machine Learning Approaches for COVID-19 Spread Prediction in Ethiopia

By Senedu Geberemariam Dekebo

Advisor: Teklu Urgessa (PhD.)

A Thesis Submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfilment for the Degree of Masters of Science in

Computer Science and Engineering

Office of Graduate Studies
Adama Science and Technology University

Adama, Ethiopia
January 2021

Declaration

I hereby declare that this MSc thesis is my original and has not been presented for a degree in any other university, and all sources of material used for this thesis have been duly acknowledged.

Name: Senedu Geberemariam Dekebo

Signature: _________________________

This MSc thesis has been submitted for examination with my approval as thesis advisor.

Name: Teklu Urgessa (PhD.)

Signature: _________________________

Date of Submission: ___________________

Approval of Board of Examiners

We, the undersigned, members of the Board of Examiners of the final open defense by Senedu Geberemariam Dekebo have read and evaluated his/her thesis entitles "Application of Machine Learning Approaches for COVID-19 Spread Prediction in Ethiopia" and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirement of the Degree of Masters of Science in Computer Science and Engineering.

__________________	__________________	__________________
Supervisor / Advisor	Signature	Date
__________________	__________________	__________________
Chairperson	Signature	Date
__________________	__________________	__________________
Internal Examiner	Signature	Date
__________________	__________________	__________________
External Examiner	Signature	Date

ACKNOWLEDGMENT

First, I would like to praise and thank God, the Almighty, for His blessing throughout my life and all steps in this research work.

My deep and genuine gratitude to my research advisor, Dr. Teklu Urgessa, for his friendly approach, readiness, and willingness to advise on the research, his critical comments, commitment to guide, and support greatly improved this thesis work.

I would also like to acknowledge the members of the Artificial Intelligence and Data Science Special Interest Group of the Computer Science and Engineering department, and to all members who sat on each phase presentation panel for their inputs which greatly shaped and improved the work. I would like to thank all my classmates who are working on their master's thesis for their timely help, idea, and support until the completion of my thesis.

Finally, my thanks go to my wonderful parents and siblings for their every support and advice in my life. May God always Bless and keep them safe.

Table of Contents

List of Tables

List of Figures

List of Equations

List of Abbreviations

AI	Artificial Intelligence
ACF	Auto Correlation Function
ANFIS	Adaptive Network-Based Fuzzy Inference System
AR	Auto-Regressive
ARIMA	Auto-Regressive Integrated Moving Average
ARMA	Auto-Regressive Moving Average
Bi- LSTM	Bi-directional LSTM
CDC	Center for Disease Control
COVID-19	Corona Virus Disease 2019
CP	Pearson's correlation coefficient
CPU	Central Processing Unit
FBProphet	Facebook Prophet
FMOH	Federal Ministry of Health
GRU	Gated Recurrent Unit
HWES	Holt Winter's Exponential Smoothing
LSTM	Long-Short Term Memory
MA	Moving Average
MAE	Mean Absolute Error

MAPE	Mean Absolute Percentage Error
MERS	Middle East Respiratory Syndrome
MLP	Multi-Layer Perceptron
MPE	Mean Percentage Error
MSE	Mean Squared Error
MS-Excel	Microsoft-Excel
NARNN	Nonlinear Autoregression Neural Network
NRMSE	Normalized Root Mean Squared Error
PPE	Personal Protective Equipment
PSNR	Peak Signal to Noise Ratio
R^2	Coefficient of Determination
RMSE	Root Mean Squared Error
RMSE	Root Mean Square Relative Error
RNN	Recurrent Neural Network
RRSE	Root Relative Squared Error
SARS	Severe Acute Respiratory Syndrome
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SMAPE	Symmetric Mean Absolute Percentage Error
SVM	Support Vector Machine

UK	United Kingdom
USA	United States of America
WHO	World Health Organization

Abstract

COVID-19 cases are increasing at an alarming rate all over the world. In Ethiopia, COVID-19 has infected hundreds of thousands and took thousands of lives. It is very important to know how many new cases and deaths to expect in the future, to be prepared in advance to avoid deaths from COVID-19.

As a solution to this problem, this research proposed to build a prediction model for COVID-19 spread in Ethiopia using machine learning approaches. COVID-19 data was collected from the Federal Ministry of Health (FMOH), Ethiopia's official website. We utilized the data from March 13, 2020, to January 10, 2021. In this study, daily confirmed, death, and recovered COVID-19 cases of Ethiopia are modeled with Multi-Layer Perceptron (MLP), Auto-Regressive Integrated Moving Average (ARIMA), Long-Short Term Memory (LSTM), and Facebook Prophet (FBProphet) approach.

The performance Metrics used are Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2). FBProphet model obtained the following results: for daily confirmed case RMSE=92.65, MAE=62.44, and R^2 = 0.95. For daily death case RMSE=2.68, MAE=1.95, and R^2 = 0.83. For daily recovered case RMSE=75.67, MAE=44.69, and R^2 = 0.98.

According to the results of the first step of the study, FBProphet achieves slightly better performance than the other models, the MLP model achieves less performance in death and recovered cases, and the ARIMA model achieves less performance in confirmed cases. In the second stage of the study, the ARIMA model was provided to make predictions in a 7-day perspective that is yet to be known. Results of the second step of the study show that the total new confirmed and death cases decrease and Recovered cases increase in Ethiopia.

Generally, COVID-19 spread prediction with the FBProphet machine learning method achieves a better performance than the ARIMA, LSTM, and MLP prediction models.

Key Words: COVID-19 spread prediction, COVID-19, machine learning, regression

CHAPTER ONE

1 Introduction

1.1 Background of the Study

Health is a state of complete physical, mental and social well-being and not only the absence of disease or susceptibility. For the fullest achievement of health, the benefits of medical, psychological, and related knowledge are essential. Governments are responsible for the health of their peoples which can be fulfilled only by the provision of adequate health and social measures [1].

The health of mankind around the world has been shaken by different kinds of diseases that were declared as pandemics, including HIV/AIDS, Cholera, and Spanish flu. The recent pandemic which is still a trait for the health of every individual in the globe is COVID-19 [2].

COVID-19 (Coronavirus Disease 2019) which started in Wuhan, China in December 2019, has emerged as a global health and economic security threat with an overwhelming growing incidence worldwide. When the World Health Organization (WHO) declared the disease as a global public health emergency, different stakeholders have stepped up efforts to convince the world that the disease is a serious problem that needs strong containment measures.

As of January 10, 2021, there have been over 82.9 million confirmed cases, and over 1.2 million deaths reported worldwide [3]. On the African continent, over 2.8 million COVID-19 Cases, 531,000 deaths, and over 1.8 million recoveries have been reported [4].

COVID-19 caused panic in Africa where the healthcare system is fragile in surviving the pandemic. Governments in the continent responded quickly in the early days, while there are concerns as some countries have limited capacity for fighting the pandemic [5].

Ethiopia, the second populated sub-Saharan country in Africa and the twelfth globally with a population of about 116 million, confirmed its first case of COVID-19 on March 13, 2020, two days later the WHO declared the disease a pandemic. As of January 10, 2021, the country tested over 1.8 million suspects, of whom 128,316 cases had been confirmed positive and of these, 1,994 died and 113,374 recovered including 2 relocations [6]. The first case in Ethiopia was forty-eight years Japanese who traveled from Burkina Faso to Ethiopia, and the second report was one Ethiopian and two Japanese, who had contact with the first Japanese person.

Before community transmission started a large number of cases were reported from mandatory quarantines which were for people coming from abroad. Nowadays cases are reported from all ten regions and two City Administrations, though the majority of cases are reported from Addis Ababa, the capital city of Ethiopia.

The Ethiopian Government has implemented a lot of measures including closing schools and universities, stopping transportation, ban all public gatherings and sports events in the entire country [7]. But the government could not continue with closing transportations and schools forever as the economy may fall drastically. So, a practical solution was to prepare places for quarantine, so that the people affected by this virus will remain in the quarantine until they fully recover [8].

As per the World Health Organization, there are currently more than 50 COVID-19 vaccine candidates on trial [9]. As of December 21, 2020, the Food and Drug Administration has given emergency use authorization to two COVID-19 vaccines. Even though the development of COVID-19 vaccines is very quick than usual, vaccine safety is still a top priority in all phases of development, approval, and post-approval monitoring [9]. However, there will be a limitation on the COVID-19 vaccine for a while, therefore, the vaccine is offered in phases. As more COVID-19 vaccines become available in 2021, vaccination will expand to include more people.

Also, as the virus mutates, how long a vaccine would stay effective is unknown. All necessary actions to reduce the spread of the virus and prepare medical response systems to tackle the increase in patient loads and protect the front-line medical staff with personal protective equipment (PPE), masks, and other essentials have to be made. So, if the number of novel coronavirus cases for the coming few days beforehand is known, responses can be planned accordingly.

Nowadays, in Ethiopia, with all the precautions in place still the virus is increasing at high speed. Hospitals and health care centers are overwhelmed by COVID-19 patients. The Shortage of resources for COVID-19 treatment is also a big challenge. Therefore, it is important to predict COVID-19 spread in the country to control the virus, and possibly reduce the death rate from the virus. This research work focuses on addressing the problem of COVID-19 spread prediction using a new dataset of Daily Confirmed cases, Daily New Deaths, and Daily Recovered cases. Machine learning approaches are proposed to solve the COVID-19 spread prediction problem in most of the works done previously, which gives good prediction results.

The rate of COVID-19 contagion and spread is faster than other viral infections encountered until today. Due to its rapid progress across the world in a short time, it is necessary to carry out intensive studies on it. Similar to many other infectious disease outbreaks, the success of controlling the new COVID-19 infection is based on revealing significant information. Therefore, predicting the future trend of the virus spread and presenting useful recommendations on how to decrease the virus spread is very important.

Machine Learning is a convenient tool to assist in fighting this pandemic by predicting the number of new confirmed cases, new deaths, and recovered by obtaining knowledge and information from the daily, weekly, or monthly report. Machine learning has attracted significant efforts to improve its performance with different metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination(R^2). Previously, different machine learning approaches have been proposed to be used for COVID-19 spread prediction.

Prediction is used for understanding the upcoming situation and can help the authorities to take timely actions and make decisions to contain the COVID-19 spread especially, in developing countries like Ethiopia where the shortage of resources and health professionals is high. The predicted number of affected people from COVID-19 in the future will be used to design appropriate responses from the higher authorities.

In Ethiopia, machine learning research on health care is almost nonexistent but it is believed that if machine learning is applied in this area, it might be critically important in revealing decision-support information for both health care administration and treatment. Therefore; the researcher is motivated to see the potential applicability of machine learning models on COVID-19 data in the Ethiopian context by taking daily reports published by FMOH, Ethiopia. FMOH provides reports and statistical data on health-related issues in Ethiopia, which should be manipulated further to derive new knowledge. The researcher believes that if machine learning methods are applied to this data, there might be a new pattern or relationship among the data, and the knowledge obtained is used to decrease deaths and infections from COVID-19. The results obtained can be used to control the virus spread in Ethiopia and improves the success and efficiency of the pandemic fight.

Finally, this research study proposed COVID-19 Daily confirmed cases, Daily death cases, and Daily recoveries in Ethiopia prediction using machine learning and deep learning algorithms on the new dataset of each case from the daily reports published by the Federal Ministry of Health (FMOH), Ethiopia.

1.2 Motivation of the Study

The Novel Coronavirus (COVID-19) pandemic is getting worse and worse daily, killing millions of people worldwide. The researcher chose this topic because the COVID-19 pandemic is spreading fast in Ethiopia and there is no means of predicting the future. This is a huge problem since nearly half of the deaths reported are not from active cases but they are found after they died from the virus. If there were a means to know how fast the virus is spreading and how many to expect to be infected beforehand it will save many lives and makes the decision-making problem easy for management.

In Ethiopia, many people are suffering from the pandemic and the virus is spreading at a high rate. The researcher believes that if the result from this study is applied to predict COVID-19 spread in Ethiopia the number of peoples who will get infected will be estimated beforehand and proper measures will be taken and the time consumed by organizations for decision-making will be less. Also, it will increase the chance of controlling the pandemic.

1.3 Statement of the Problem

According to daily reports from the Ministry of Health, Ethiopia the number of new confirmed cases and new deaths are increasing. The virus is spreading at a high rate and there is no mechanism to predict the future of the virus in the country. Applying machine learning for predicting the number of new confirmed cases and new deaths helps to decrease the rate of the infection and possibly save lives.

There have been some attempts in applying machine learning models for fighting the pandemic in different areas. In the last twelve months from the first incident, different researches have been done on the area, but machine learning models have a significant effect on predicting the number of new confirmed cases and new deaths.

This study addresses these problems by building a new COVID-19 dataset of new confirmed cases, new death cases, and new recovered cases in Ethiopia, utilizing machine learning and deep learning methods; this study addressed the problem by answering the following questions:

- How to implement machine learning and deep learning models to accurately predict COVID-19 spread in Ethiopia?
- To what extent machine learning approaches predict COVID-19 cases in Ethiopia?
- Which machine learning approach is better for COVID-19 spread prediction in Ethiopia?

1.4 Objectives of the Study

1.4.1 General Objective

The main objective of this study is to design and develop a predictive model for COVID-19 spread in Ethiopia through different machine learning approaches.

1.4.2 Specific Objectives

The specific objectives of this study are:

- To review different kinds of literature that support the study in the area of applying machine learning approaches for COVID-19 spread prediction.

- To collect COVID-19 data from the daily reports of the Federal Ministry of Health, Ethiopia.
- To develop prediction models using different machine learning approaches.
- To evaluate the performance of the models.

1.5 Scope and Limitations of the study

1.5.1 Scope of the Study

This study focuses only on Predicting COVID-19 spread in Ethiopia by using machine learning approaches. A new dataset is built by collecting daily new confirmed cases, daily death cases, and daily recovered cases from FMOH, Ethiopia's official website from March 2020 to January 2021. This study implements machine learning models for prediction and evaluates the result of each model based on different performance evaluation techniques.

1.5.2 Limitations of the Study

This study comes across limitations on a different phase of the research process, since there is a lack of shared public dataset and other studies for comparison in COVID-19 spread prediction for Ethiopia. As a result, this study creates a new dataset. The other constraints of this study are listed as follows:

- Due to some security issues in the country reports were not published on the FMOH website, so getting those reports and building a new dataset was challenging.
- Due to the tight schedule, this study only implemented and evaluated the proposed machine learning approaches on confirmed, death, and recovered COVID-19 cases, but does not include cases in the Intensive Care Unit and cases still in treatment.

1.6 Application of the Study

The main beneficiaries of this study are the authorities that take action and make decisions to contain the COVID-19 pandemic. On the one hand, governments and other legislative bodies benefit by taking these work results as input for developing new policies and to assess the effectiveness of the enforced policies for reducing COVID-19 spread in Ethiopia. Also, it helps to reduce the mortality rate and helps to utilize resources properly in the right place depending on the outcome of this study. Additionally, researchers can replicate the proposed research for other infections in Ethiopia and can serve as a baseline for related work on this issue or use the dataset to improve the research.

1.7 Organization of the Thesis

This research paper is organized into seven chapters in the following ways:
Chapter one discussed above, and includes the introduction of the study, the motivation, and the statement of problems, the research questions that would be answered in the proposed solutions, the scope and limitation, the objective of the study, the methodology, and the application results of the study.

Chapter two, presents the literature review and related works on COVID-19 spread prediction, definition and behavior of COVID-19, COVID-19 worldwide, COVID-19 in Ethiopia, Methodologies use in COVID-19 spread prediction: it discussed the overview of COVID-19, and finally, it discussed the related work.

Chapter three, discusses the research methodology for this research work, methods used to build the dataset, methods used to develop COVID-19 spread prediction models which are the research design for the study, and machine learning regressor algorithms, and finally, it presents the evaluation methods.

Chapter four discusses the proposed solution for COVID-19 spread prediction, which is the proposed COVID-19 machine learning approaches, machine learning models training, and testing.

Chapter five discusses the implementation and experimentation of the proposed solution. Including working environment, dataset description, model implementations, and evaluation models.

Chapter six discusses the results of the proposed machine learning approaches and also discusses the major results obtained by comparing all models based on the performance metrics.

Chapter seven presents a conclusion, recommendation, and identifies future works for further research direction on this research topic.

CHAPTER TWO

2 Literature Review and Related Works

Intensive research work is going on to evaluate and contain the worldwide disaster of COVID-19 on the human race. This chapter reviews relevant literature to further comprehend the concept and investigate the research problem.

2.1 Conceptual Literature

A technical review on COVID-19 spread prediction and existing approaches used for prediction. Using prediction models to predict the impact of the disease on the community helps to control the pandemic. Starting by defining the term COVID-19, and followed by COVID-19 Worldwide and Ethiopia, a review of COVID-19 spread prediction, and approaches that have been used in COVID-19 spread prediction and related works previously studied on the problem.

2.1.1 COVID-19

Coronaviruses are a large family of viruses causing mild to severe illnesses. One of the coronaviruses, SARS coronavirus (SARS-CoV) caused severe acute respiratory syndrome (SARS), which lasted from November 2002 to 2004. Another type of coronavirus, Middle East respiratory syndrome (MERS) caused by MERS coronavirus (MERS-CoV), was identified in September 2012 and continues to cause local outbreaks. The third novel coronavirus (COVID-19) is an infectious disease called Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [10]. Since the outbreak of the virus in December 2019, cases have been identified in a growing number of countries[11].

According to WHO a pandemic is when a new disease spreads worldwide and sustains a risk of future global spread. On March 11, WHO declared COVID-19 as a pandemic causing over 118,000 cases in over 110 countries and territories around the world [12].

COVID-19 spreads between people who are at a close distance from one another and through respiratory droplets produced when an infected person breathes, coughs, sneezes, sings, or talks. To reduce the spread of the virus and to be safe everyone should wash hands often, avoid close contact, always wear a mask around people, cover nose and mouth when sneezing and coughing, clean and disinfect frequently touched surfaces, and monitor health daily[13].

According to the Center for Disease Control (CDC) update on May 13, 2020, anyone can have mild to severe symptoms. Adults and people with severe underlying medical conditions like diabetes, lung, or heart disease are at higher risk for developing more serious complications from COVID-19. After being exposed to the virus symptoms ranging from mild to severe illness may appear in 2-14 days. The most common symptoms are fever, dry cough, and tiredness. The less common symptoms include aches and pains, sore throat, diarrhea, headache, loss of test or smell [14].

2.1.2 COVID-19 Worldwide

Starting from the first case reported in China from December 19, 2020, to January 10, 2021, COVID-19 cases are reported from every continent in 220 countries, areas, or territories. Confirmed cases are 89,802,096 and 1,940,529 Confirmed deaths globally[15]. The number of confirmed cases and death cases in each continent are shown in Table 2-1 as reported by the European Centre for disease prevention and control as of January 10, 2021 [16].

Table 2-1 COVID-19 confirmed and death cases in Each Continent

Continents	***Confirmed cases***	***Death cases***
Africa	3,059,974	72,834
Asia	18,549,010	317,547
America	39,844,634	925,925
Europe	28,291,217	623,024
Oceania	56,556	1,193
Others (international transport in Japan)	705	6

From Table 2-1, Africa has over three million confirmed cases, the five countries reporting most cases in Africa are South Africa (1,231,597), Morocco (452,532), Tunisia (162,350), Egypt (149,792), and Ethiopia (128,616). The five countries reporting the highest number of deaths are South Africa (33,163), Egypt (8,197), Morocco (7,743), Tunisia (5,284), and Algeria (2,807) [16].

New infections and deaths worldwide from COVID-19 are increasing each day, even though the development of the vaccine is continuing and showed good results, the infection and death cases are still high. Developing a way for predicting future cases might help in saving lives worldwide.

2.1.3 COVID-19 in Case of Ethiopia

The original incident of the COVID-19 virus in Ethiopia was conveyed on March 13, 2020. As of January 10, 2021, the Federal Ministry of Health, Ethiopia has established a total of 128,316 confirmed cases, 113,374 recovered including 2 relocations, and 1,994 deaths [6]. Ethiopia is the fifth country in Africa reporting the highest number of confirmed cases which is 4.2% of the total cases in Africa.

After the first incident different measures have been taken to decrease the spread of the COVID-19 and to save lives. The prevention measures implemented include temporally closing all schools, changing work times to work in shifts to make the work area safe, the six feet social distancing rule, and wearing a mask all the time unless in-home [17]. After a while, all the prevention measurements have been slightly removed and schools have been open and transportations are good to go with proper protection [18]. Even though different preventive actions have been taken, the number of new confirmed cases is increasing at a high rate.

COVID-19 cases have been reported in all regions and city administrations of the country. A large number of confirmed and death cases are reported from Addis Ababa city administration, Oromia and Amhara regions are the second and third regions reporting the highest cases in Ethiopia [6].

2.2 Spread Prediction Approaches in HealthCare

Until today spread prediction approaches in different areas of health care have been studied by different researchers. Some of the areas of healthcare in which mathematical and machine learning prediction approaches have been applied include infectious disease prediction and control [19], prediction about future infectious disease risks [20], and predicting the future of influenza [21] [22].

2.2.1 COVID-19 spread prediction Approaches

COVID-19 spread prediction has been studied by different researchers, and different approaches have been used to predict the virus spread globally and at the country level. In all approaches, COVID-19 spread prediction begins with collecting data from different platforms and government databases and testing the approach that best fits the scenario [23],[24],[25].

Almost all previous researches reviewed for this study uses a machine learning approach to predict COVID-19 Spread.

2.2.2 Machine Learning for COVID-19 spread prediction

Most state-of-the-art COVID-19 prediction techniques involve machine learning methods for regression. After preparing the dataset the regression algorithms perform the prediction task.

2.2.2.1 Machine Learning

Artificial Intelligence (AI) is applied through machine learning which provides the ability for systems to learn and improve from experience without being explicitly programmed. Machine learning develops programs that can learn from observations, data, direct experience, or instructions to find patterns in data and make better decisions in the future based on the samples provided. The main goal of machine learning is to allow computers to learn automatically without human involvement and adjust actions accordingly. Machine learning approaches can be categorized into supervised, unsupervised, and semi-supervised approaches [26].

Supervised machine learning: algorithms take a known set of input data and known response to the data or output and train a model using labeled samples to predicts future events. After sufficient training, the system can give targets for any new input. To modify the model, the learning algorithm compares its output values with the correct output value and finds errors [27].

Unsupervised machine learning: algorithms are trained with unclassified or unlabeled data, only training data is provided. The model is not supervised by the user, it allows the model to discover patterns and information that was previously undetected. Unsupervised learning describes how a model can infer a function from unlabeled data to describe a hidden pattern. The algorithm explores the data and draws inferences from datasets to describe hidden structures from the unlabeled data [27].

Semi-supervised machine learning: is a combination of unsupervised and supervised machine learning methods. This machine learning is used when labeled data is not enough to produce an accurate result and when there is a lack of ability or resource to get more labeled data, then semi-supervised techniques are used to increase the size of training data with unlabeled data [27].

The most common approach used for COVID-19 spread prediction is the supervised machine learning approach. This approach relies on the given dataset. Most of the research works reviewed for this research used a regression algorithm such as Long-Short Term Memory (LSTM), Auto-Regressive Integrated Moving Average (ARIMA), Multi-Layer Perceptron (MLP), and Facebook Prophet (FbPh) methods. Almost all of the research studies for COVID-19 spread prediction use more than two regression algorithms for computation, comparison, and are used to suggest which algorithm has high performance and prediction result for their proposed prediction model [28], [29], [30], [31].

Auto-Regression Integrated Moving Average (ARIMA): is a supervised machine learning model for solving regression problems. ARIMA algorithm performs regression by removing non-stationary values from the series using seasonal differencing of an appropriate order. A first-order seasonal difference is a difference between an observation and the corresponding observation from the previous year, month, or day. The model will iteratively estimate the three parameters which are the level of difference (d), the autoregressive order (p), and the moving average order (q) for better prediction [32].

Multilayer Perceptron (MLP): is a known and most used neural network which can be used for regression problems. This model has multiple layers consisting of neurons. Learning in this model is achieved in a supervised manner. The power of MLP comes from the non-linear activation function. There activation functions for updating the weight in each layer. The three layers in the network are the input layer, hidden layers, and the output layer. The choice of activation function for the output layer depends on the nature of the problem to be solved. For the hidden layers of neurons, sigmoid functions are preferred, because they have the advantage of both non-linearity and differentially. for output neurons, the activation function adapted to the distribution of the output data is recommended[33].

Long-Short Term Memory (LSTM): is one of the most powerful types of Recurrent Neural Network (RNN). Depending on the complexity of the built network LSTMs are capable of learning more than a thousand-time step. Access to the cells is provided by multiplicative gate units, which learn when to grant access. The three passes of LSTMs Forward pass, Forget gate, and Backward pass. The forget gate learns to reset the internal state of the memory cell when the stored information is no longer needed. According to [34], LSTM excels on tasks in which a limited amount of data must be remembered for a long time. Memory blocks have access control in the form of input and output gates; which prevent irrelevant information from entering or leaving the memory block. Memory blocks also have a forget gate which weights the information inside the cells, so whenever previous information becomes irrelevant, the forget gate can reset the state of the different cells inside the block. The most commonly used LSTM models are Bidirectional LSTM, Gated Recurrent Unit (GRU), and Grid LSTM (or N-LSTM). LSTM introduces a constant error flow through the internal states of a special memory cell as a possible solution for the vanishing error problem of RNN. LSTM can free the memory of irrelevant information [34].

Facebook Prophet (FBProphet): is a procedure developed by Facebook for predicting time series data. It's based on an additive model where non-linear trends are fit with daily seasonality, weekly seasonality, yearly seasonality, and holiday effects. The four main components are a linear or logistic growth curve trend, yearly seasonality component, weekly seasonal component, user-provided list of holidays. This model has better performance on time series data that have strong seasonal effects and several seasons of historical data. It can make time-series predictions with good accuracy using simple intuitive parameters. Prophet handles outliers, missing data, and shifts in trend well [35].

Table 2-2 Frequently Used Machine Learning Algorithms in COVID-19 spread prediction

Algorithm	***Pros***	***Cons***
LSTM	has a memory that can store previous timestep information, can handle noise, distributed representation, and continues value, works well over a broad range of parameters.	take longer to train, require more memory to train, are easy to overfit, Dropout is much harder to implement in LSTMs, sensitive to different random weight initializations.
ARIMA	can reduce non-stationary series to stationary series, handles seasonal data better.	needs large data, has no automatic updating feature, high cost.
MLP	used as a metaphor for human neural networks, Computationally efficient, Universal computing machines.	Convergence can be slow, local minima affect the training process, hard to scale.
FBProphet	it is fast, performs well on data with daily periodicity, handles missing data, shifts in the trend, and large outliers better.	Need manual tuning of each parameter, Exogenous variables cannot be taken care of, Data need to be feed in a pre-defined format.
SVM	Memory-efficient works well when there is a clear margin of separation, more effective in high dimensional spaces.	not suitable for large data sets, does not perform very well when the data set has more noise, doesn't directly provide probability estimates.
RNN	Model size does not increase, weights can be shared across the time steps, process inputs of any length, remember each information throughout the time.	Using activation functions makes the training and computation slow, processing long sequences is difficult, gradient vanishing and exploding are problems.

2.2.2.2 Deep Learning Used in COVID-19 spread prediction

Deep learning is a machine learning method based on artificial neural networks, teaching machines to do what comes naturally to humans. Learning can be supervised, unsupervised, or semi-supervised. Nowadays, Deep Learning models show promising results in COVID-19 spread prediction. It depends on the artificial neural networks but with extra depth. It tries to mimic the event in layers of neurons and attempt to learn in a real sense to identify patterns in the provided data. However, deep learning approaches are not always better than traditional machine learning supervised approaches. The performance of deep learning is subject to the right choice of algorithm and number of hidden layers as well as the learning rates. For instance, research work [29] and [36] propose a deep learning neural network-based on COVID-19 spread prediction using Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM) respectively. Also, their results showed a promising regression performance and accurate result for COVID-19 spread prediction.

2.3 Related Works

Prediction and analysis of COVID-19 positive cases using deep learning models, [37] Parul Arora et al. performed a study on prediction for COVID-19 positive cases in 32 states and union territories of India, Creating a dataset of 32 individual time-series data of confirmed COVID-19 cases in 28 states and 4 union territories since March 14, 2020, then the linear weighted moving average technique is used for handling missing values in each series, Recurrent Neural Network (RNN) based long short term memory (LSTM) cells are used as prediction models and LSTM variants such as deep LSTM, convolutional LSTM, and Bi-directional LSTM models are tested. Based on absolute error Bi- LSTM achieved very accurate results (error less than 3%) for short-term prediction (1–3 days) and convolutional LSTM gives the worst Daily and weekly predictions. The authors conclude that the result is helpful for state and national government authorities, researchers, and planners for managing services and arranging medical infrastructure accordingly, and the proposed model can be followed by other nations.

Ismail Kirbas et al. [23], proposed LSTM, ARIMA, and NARNN approaches for comparative analysis and forecasting COVID-19 cases in Eight European countries. The dataset was obtained from European Center for Disease Prevention and Control, it contains, confirmed COVID-19 cases of Denmark, Belgium, Germany, France, United Kingdom, Finland, Switzerland, and Turkey, the data covers 67, 90, 97, 100, 94, 90, 68 and 55 days respectively and ends on 3 May 2020. Experiments were done on each countries data by ARIMA, NARNN, and LSTM algorithms, and MSE, PSNR, RMSE, NRMSE, MAPE, and SMAPE performance metrics are used to select the most accurate model. According to the results, it was determined that the LSTM approach has the best performance compared to NARNN and ARIMA. Forward estimations were made with LSTM with high performance. Among the countries studied, in the UK highest rate of increase is observed, while the lowest number of cases was observed in Finland. According to the two-week future estimation study, in many countries, the total cases rate is expected to decrease slightly. The authors believe that the model performances will increase when the number of days in which the outbreak data is diversified and the data collection is increased. In this, study the number of days will increase and the data will be diversified so the results will be improved in this study.

Saleh I. Alzahrani et al. [30] used the ARIMA model for COVID-19 spread forecasting in Saudi Arabia under current public health interventions. The dataset contains the daily confirmed cases of COVID-19 across 31 affected cities from March 2, 2020, to April 20, 2020. In this work, four statistical models (AR, MA, ARMA, and ARIMA models) are used to predict the spread of COVID-19. The performance of each model was evaluated using (RMSE, MAE, R^2, MAPE, and RMSRE) performance metrics. The results obtained are for ARIMA (RMSE=21.17, MAE=14.93, R^2 =0.99, MAPE =2.16, RMSRE=0.024), for ARMA (RMSE=117.22, MAE=73.33, R^2 =0.87, MAPE =8.34, RMSRE=0.12), for AR (RMSE=180.29, MAE=160.4, R^2 = 0.69, MAPE =32.32, RMSRE=0.25), for MA (RMSE=241.05, MAE=190.87, R^2 = 0.46, MAPE =32.80, RMSRE= 0.26). From the results obtained, the ARIMA model has the best RMSE, MAPE, RMSRE results, and highest R^2 value. Therefore, the ARIMA model performs better than the other models proposed. The ARMA model performs better in second place, followed by AR then MA models. The authors believe that these findings can potentially help in promoting policies to address the COVID-19 pandemic and put viable strategies to contain it through introducing new movement restrictions as well as conducting proactive mass testing for potential COVID-19 cases across Saudi Arabia.

Sina F. Ardabili et al. propose Machine Learning for COVID-19 Outbreak Prediction [24], which presents a comparative analysis of soft computing models and machine learning to predict the COVID-19 outbreak. Among a wide range of machine learning models studied, Multi-Layer Perceptron, and adaptive network-based fuzzy inference systems showed promising results and reported a high generalization ability for long-term prediction. The dataset used in this study is the total case number (cumulative statistic) for five countries, including Italy, Germany, Iran, the USA, and China on total cases over 30 days. Due to the highly complex nature and variations in the behavior of the virus from nation-to-nation, this study recommends machine learning as an effective tool to model the outbreak.

Naresh Kumar and Seba Susan [38], conducted a study on COVID-19 Pandemic Prediction using Time Series Forecasting Models. In this paper, analysis and prediction are performed using ARIMA and FBProphet forecasting models. The dataset used in this study is a day level information of COVID-19 spread for cumulative cases from the whole world and ten most affected countries; US, Spain, Italy, France, Germany, Russia, Iran, United Kingdom, Turkey, and India from January 22, 2020, to May 20, 2020. For most of the countries' data, ARIMA has better performed compared to Prophet on the scale of MAE, RMSE, RRSE, and MAPE error matrices. The trend analysis shows rapid growth in the infected cases, and the prediction study shows a great rise in the expected active, recovered, and death cases worldwide. The authors suggest that the obtained forecasting results further can be improved by taking various variables into account like population density, weather, health system, patient history using deep learning techniques, and AI.

Tarik Chafiq et al., [31] conducted a study on forecasting COVID-19 in Morocco using the FBprophet Facebook framework in Python. The dataset used is daily level information on the number of affected cases by COVID-19 from march 02, 2020 to August 29, 2020, in Morocco from the Johns Hopkins University website. FBprophet algorithm for forecasting is applied on the dataset and performance validation is done using RMSE, MSE, MAE, and MAPE. The authors believe that this might be helpful for decision-makers and health authorities to evaluate risks and take proper precautions and strategies to control the spread of infectious diseases.

Mario Jojoa and Begoña Garcia-Zapirain, [29] analyzed machine learning for forecasting COVID-19 confirmed cases by taking the case of America. Support Vector Machine and Multilayer Perceptron algorithms are used to predict the number of COVID-19 infections in Brazil, Colombia, Mexico, Chile, Peru, and the United States. The dataset is open data from the European Union repository where a time series of confirmed contagious cases was modeled until May 25, 2020, it’s composed of 11 fields and has 19,248 records. The metrics used for evaluation are Pearson's correlation coefficient (CP), Mean Absolute Error (MAE), and Mean Percentage Error (MPE). This research work demonstrates that an MLP can predict better when an optimization algorithm to determine the hyperparameter (number of layers and neurons per layer) is applied. The results prove that MLP performs well for Chile, Mexico and the USA and SVM performs well for Brazil, Colombia, and Peru. The authors concluded that machine learning-based prediction systems are fundamental tools in decision-making to save thousands of lives.

Mehdi Azarafza et al. [36] proposed deep learning for COVID-19 Infection Forecasting for Iran. The dataset used contains COVID-19 confirmed cases from February 19, 2020, to March 22, 2020, at the regional level, and from February 19 to May 13, 2020, national level, collected from nationally recognized sources in Iran. From the various deep learning methods, LSTM is employed for forecasting time-series data. RNN, SARIMA, HWES, and moving averages approaches are compared to the LSTM model. The mean absolute percentage error, mean squared error, and mean absolute error metrics were used for comparisons. The results of this study show that the LSTM model has better performance than the other methods and gave fewer error values for infection development in Iran.

Abdelhafid Zeroual et al. [39] conducted a comparative study on deep learning for COVID-19 forecasting on time-series data. The dataset used in this study is the number of recovered and confirmed cases collected daily from six countries (Italy, Spain, China, the USA, and Australia) starting from January 22, 2020, till June 17, 2020. Five deep learning methods (Recurrent Neural Network, Long Short-term memory, Bidirectional LSTM, Gated recurrent units, and Variational AutoEncoder) are used for forecasting the number of new cases and recovered cases. The results from this study show the potential of the deep learning model in forecasting COVID-19 cases and highlight the performance of the VAE compared to the other five algorithms.

Qiuying Yang et al. [40] studied the COVID-19 epidemic in Italy based on the ARIMA model taking Hubei, China as an example. The dataset used was from the daily number of confirmed cases and new deaths of COVID-19 cases in Hubei, China from January 27, 2020, to March 17, 2020, the validated models were used to predict the trend of the epidemic of COVID-19 in Italy. ARIMA model is applied to predict the development of the Italian epidemic in the next 10 days and performance is evaluated by MAE criteria, result shows the models fit well and are more suitable for short-term prediction.

Roseline Oluwaseun et al. [41] propose Machine Learning for COVID-19 pandemic prediction in India [24], which used a simple average aggregated machine learning method for predicting the number, size, and length of COVID-19 cases level and windup period across India. The dataset is gathered were in a monthly form that is from January 2020 to April 2020. As of April 25, 2020, the COVID-19 dataset includes accumulated 408,658 total samples, confirmed cases of 24,942, recovered cases of 5,209, death cases of 779, and one migration. Auto-Regressive Integrated Moving Average (ARIMA) model is applied and also built a simple mean aggregated method established on the performance of three regression techniques such as Support Vector Regression, Neural Network, and Linear Regression. The results showed that COVID-19 disease can correctly be predicted and Postulated Aggregated Method predicted better.

2.3.1 Summary of Related Works

The next Table 2-3 presents a summary of related works in COVID-19 spread prediction research.

Table 2-3 Summary of COVID-19 Spread Prediction Related Works

Authors & year	***Title***	***Model***	***Result***
Parul Arora et al. (2020) [37]	Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India	- Stacked LSTM - Convolutional LSTM - Bi-directional LSTM	- Stacked LSTM (MAPE = 4.81%) - Bi-directional LSTM (MAPE = 3.22%) - Convolutional LSTM (MAPE = 5.05%)
Ismail Kırbas et al. (2020) [23]	Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN, and LSTM approaches	- ARIMA - NARNN - LSTM	LSTM makes the most successful estimation in Switzerland with MAPE of 2.5025 and the UK with a MAPE of 2.5025.
Saleh I. Alzahrani et al. (2020) [30]	Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions	- ARIMA - ARMA - AR - MA	- **ARIMA** (RMSE=21.17, MAE=14.93, R^2 =0.99, MAPE =2.16, RMSRE=0.024) - **ARMA** (RMSE=117.22, MAE=73.33, R^2 =0.87, MAPE =8.34, RMSRE=0.12) - **AR** (RMSE=180.29, MAE=160.4, R^2 = 0.69, MAPE =32.32, RMSRE=0.25)

			- **MA** (RMSE=241.05, MAE=190.87, R^2 = 0.46, MAPE =32.80, RMSRE= 0.26)
Sina F. Ardabili et al. (2020) [24]	COVID-19 Outbreak Prediction with Machine Learning	- MLP - ANFIS	Both models showed good results and reported a high generalization ability for long-term prediction.
Naresh Kumar and Seba Susan (2020) [38]	COVID-19 Pandemic Prediction using Time Series Forecasting Models	- ARIMA - Prophet	ARIMA has better performed compared to Prophet on the scale of error matrices.
Tarik Chafiq et al. (2020) [31]	COVID-19 forecasting in Morocco using FBprophet Facebook's Framework in Python	- FBprophet	FBprophet has good MSE, RMSE, MAE, and MAPE results.
Mario Jojoa and Begoña Garcia-Zapirain (2020) [29]	Forecasting COVID 19 Confirmed Cases Using Machine Learning: The Case of America	- MLP - SVM	MLP can predict better than SVM when optimization is applied.
Mehdi Azarafza et al. (2020) [36]	COVID-19 Infection Forecasting based on Deep Learning in Iran	- LSTM - RNN - SARIMA - HWES - Moving averages	LSTM model performed better than the other methods

Abdelhafid Zeroual et al. (2020) [39]	Deep learning methods for forecasting COVID-19 time-series data: A Comparative study	- RNN - LSTM - Bi-LSTM - GRU - VAE	VAE model performed better than the other proposed methods.
Qiuying Yang et al. (2020) [40]	Research on COVID-19 based on ARIMA model - Taking Hubei, China as an example to see the epidemic in Italy	- ARIMA	the model fits well and is more suitable for short-term prediction.
Roseline Oluwaseun et al. (2020) [41]	Machine Learning Prediction for COVID-19 Pandemic in India	- ARIMA - Postulated aggregated method of SVR, NN, and LR	Postulated Aggregated Method predicted better

The findings from this review are helpful to understand different aspects posed by the research on the use of machine learning approaches for COVID-19 spread predictions in different areas. This is significant because different researchers have a different approaches to COVID-19 pandemic prediction. There has been some research and discussion conducted on the prediction of COVID-19 cases worldwide and at the country level, including FBProphet, ARIMA resulting in promising performance. Most of the research found was on the prediction of new confirmed cases globally and for some selected countries. More research and testing are required to gain a better understanding of which machine learning approach predicts new confirmed cases, new death cases, and new recovered cases in Ethiopia. It is important to conduct more studies on COVID-19 cases and predictions using machine learning and have a better prediction model for Ethiopia.

CHAPTER THREE

3 Methodology

In this chapter, the research methodology to build the datasets and techniques to achieve research objectives and answer the research question are discussed. This chapter explains and justifies the methodologies used in conducting the study on COVID-19 Spread Prediction in Ethiopia.

3.1 Research Design of this Study

The research design for this study is Experimental. A literature review is done for understanding the problem and formulating the research questions. After research question formulation the data collection is performed. The solution for solving the problem is proposed and designed. Finally, an evaluation of the solution is performed and results from the evaluation are analyzed.

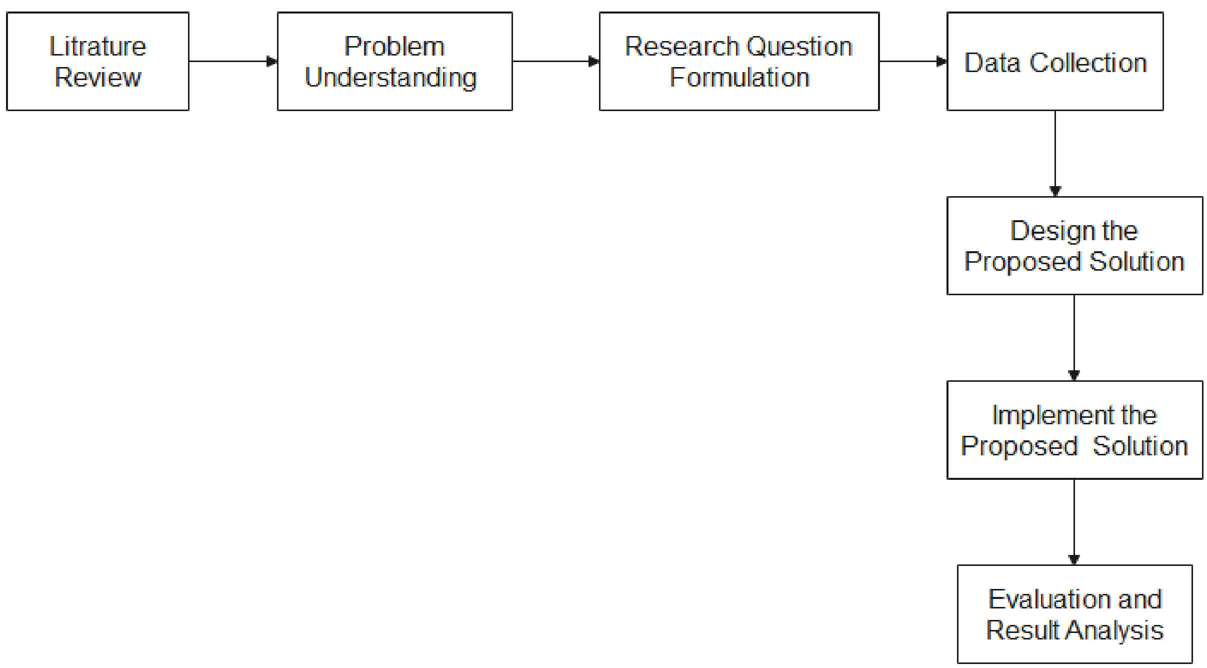

Figure 3-1 Research design for this study

3.2 Building a Dataset

The objectives of this study are to Predict COVID-19 Spread in Ethiopia. So, it needs a new COVID-19 daily confirmed cases, daily death cases, and daily recovered cases in Ethiopia datasets built. This new dataset needed because there are no published datasets for this purpose. The process of building the datasets for COVID-19 Spread Prediction in Ethiopia consists of two main steps:

1. Gathering the daily confirmed, death, and recovered cases from the Federal Ministry of Health official website and daily published reports of FMoH.
2. Preparing and consolidating gathered data into the appropriate dataset.

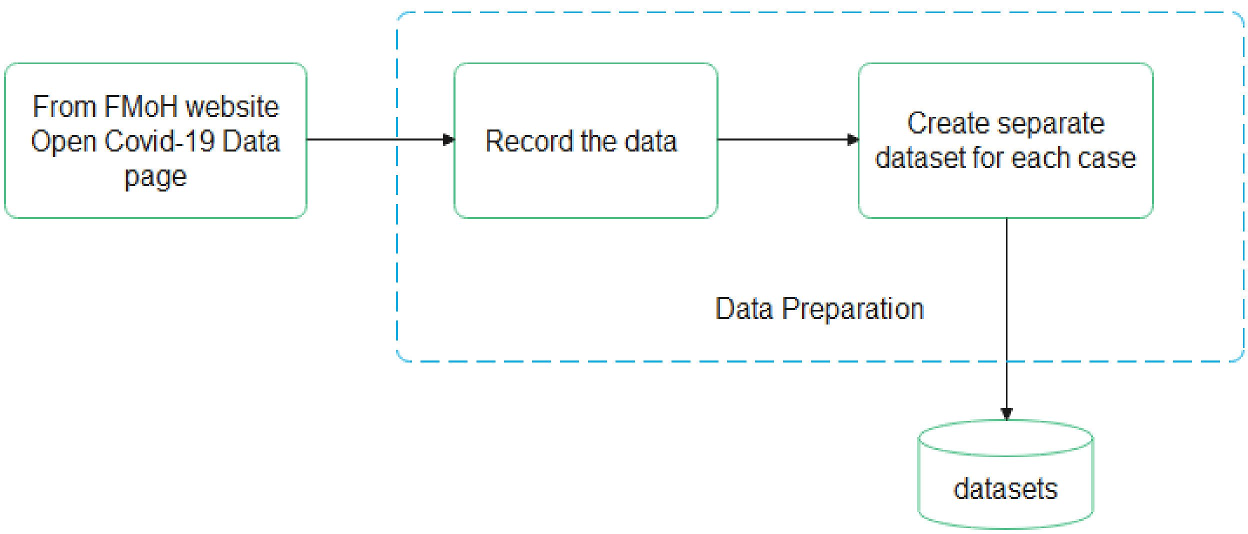

Figure 3-2 Method for Building COVID-19 in Ethiopia Datasets

3.2.1 Data Collection

This study gathers Daily confirmed cases, Daily death cases, and Daily recovered cases of COVID-19 in Ethiopia from FMoH's official website. This study aims to have a different dataset for each case. The researcher decided to use FMoH official website from other sources because of the trustworthiness of the platform, and according to the Ethiopian government, FMoH is the responsible office in health and working with the Ethiopian Public Health Institute which is tracking and fighting COVID-19 in Ethiopia [42].

This study collects daily Confirmed cases, Death cases, and Recovered cases from COVID-19 in Ethiopia from the first incident as of March 13, 2020, till January 10, 2021.

3.2.2 Dataset Preparation

After the data is collected, a preparation process follows, which is checking null values and unnecessary characters. This process used MS-Excel for dataset Collection and preparation, then used for designing a training model for COVID-19 Spread Prediction.

3.3 COVID-19 Spread Prediction Modeling

3.3.1 Machine Learning Algorithm

The objective of this study is to predict the spread of COVID-19 in Ethiopia using machine learning algorithms on the daily reported confirmed, death, and recovered cases datasets. RMSE, MAE, and R^2 performance evaluation metrics for comparison and accuracy of the prediction are used. The Auto-Regressive Integrated Moving Average (ARIMA), Multilayer Perceptron (MLP), Long-Short Term Memory (LSTM), and Facebook Prophet (FBProphet) algorithms are selected because of their good regression performance, and their popularity in solving time-series prediction problems on previous research work results in related work for other Countries [30], [31], [29], [36]. The following four machine learning algorithms are used to build prediction models.

3.4 Performance Evaluation

Evaluation for this study has been conducted on the prediction models for COVID-19 spread prediction. After all the evaluations have been conducted, results are discussed and concluded by suggesting a model that predicts COVID-19 spread in Ethiopia better.

3.4.1 Prediction Model Performance Evaluation Metrics

Prediction Model evaluation is required to quantify the prediction model performance because regression algorithms give a biased result. This study uses Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of determination (R^2).

Terms associated with performance evaluation metrics:

- **Actual values:** are the values from the dataset.
- **Predicted values:** are the outputs from the prediction model.
- **N:** are the total number of observations/rows in the dataset.

3.4.1.1 Mean Absolute Error

Mean Absolute Error is the total sum of the absolute difference between actual and predicted values and used to evaluate the regression problem's accuracy. MAE is based on the absolute value of error and will tell us how big of an error can we expect from the forecast on average. It is quite robust to outliers. MAE value lies between zero to infinity, treats large and small errors equally, and not sensitive to outliers. Small value indicates better model [43], [44].

$$MAE = |\, actual\ values - predicted\ values| \quad (3\text{-}1)$$

3.4.1.2 Root Mean Squared Error

RMSE is the standard deviation of errors when a prediction is made on a dataset. It is used as it punishes large errors and results in a score that is in the same unit as the forecast data. In this metric, the root of the values is considered while determining the accuracy of the model [45]. The value lies between zero to infinity, small value indicates a better model. RMSE is always greater than or equal to MAE. The greater difference between them indicates a greater variance in individual errors in the sample [43].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(predicted\ values - actual\ values)^2}{N}} \quad (3\text{-}2)$$

3.4.1.3 Coefficient of Determination (R^2)

The R-squared metric is a statistical measure of the closeness of the data to the fitted regression line. R- squared is always between 0 and 1. When the R^2 value is higher the model is a better fit for the data [46]. The higher the coefficient, the higher percentage of points the line passes through when the data points and line are plotted. When the value is negative, this means the model is not a good fit for the data [47].

CHAPTER FOUR

4 The proposed solution for COVID-19 Spread Prediction in Ethiopia

This chapter discusses the proposed solution for COVID-19 spread prediction in Ethiopia using machine learning approaches suitable to solve the problem. This study creates a new COVID-19 cases dataset of daily confirmed cases, daily death cases, and daily recovered cases and used it as the main component in the proposed solution to predict the spread of COVID-19 in Ethiopia.

4.1 The proposed COVID-19 Spread Prediction Framework

The proposed framework is used to predict COVID-19 Spread in Ethiopia as daily new cases, death cases, and recovery cases. The proposed solution is based on the framework shown in Figure 4-1. It takes COVID-19 daily confirmed cases, daily death cases, and daily recovered cases in Ethiopia dataset as input. After collecting the data, models are developed using ARIMA, MLP, LSTM, and FBProphet machine learning algorithms and trained on a training set of the whole dataset. The models are evaluated using RMSE, MAE, and R^2 metrics. The evaluation results were used to select the best prediction model. The outcome of these tasks is a model used for COVID-19 Spread Prediction in Ethiopia. Finally, the prediction model is evaluated and selected based on the results obtained using the model evaluation method discussion in chapter three. The final selected prediction model was used to develop a prediction model for COVID-19 spread in Ethiopia.

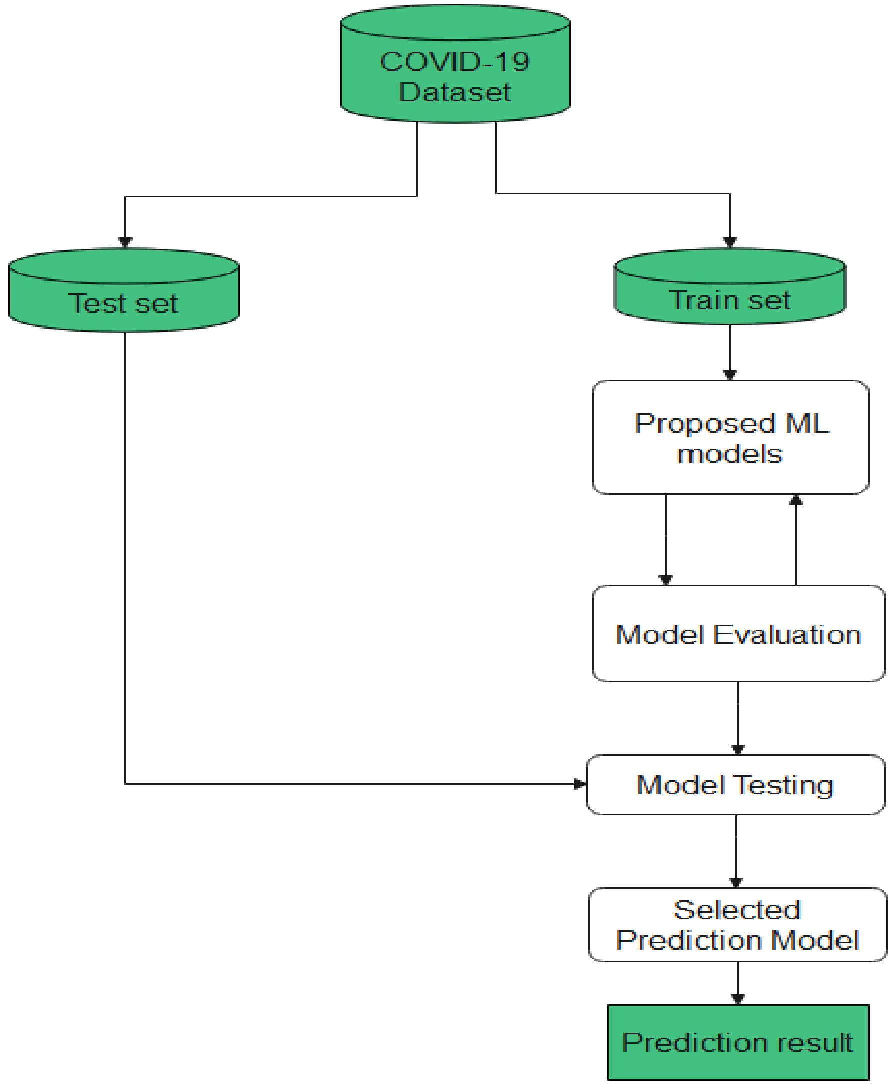

Figure 4-1 COVID-19 Spread Prediction Framework

4.2 Machine Learning Models Building

This subcomponent of the proposed framework for COVID-19 Spread Prediction performs machine learning regression. This study builds a regression model mainly by using machine learning regression algorithms ARIMA, MLP, LSTM, and FBProphet on the dataset. Which could be selected for the prediction of the COVID-19 spread prediction. The process of modeling is to find patterns in the training set of the dataset for prediction using machine learning algorithms. The output of these modeling is a trained model that can be used for predicting COVID-19 spread, making predictions on new confirmed, death, and recovered cases.

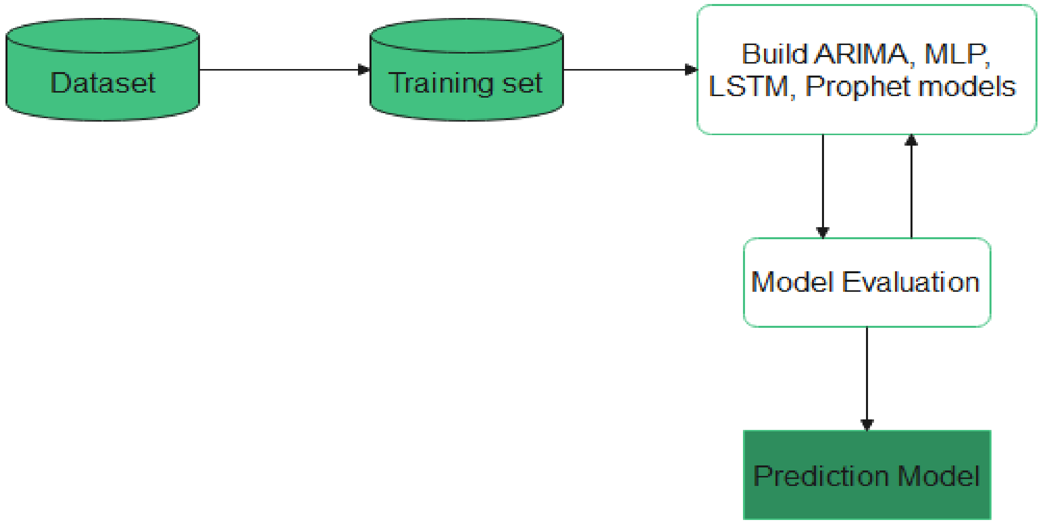

Figure 4-2 Model Building Flow Diagram

This study proposed the non-seasonal type of Auto-Regressive Integrated Moving Average (ARIMA) machine learning model. It integrated auto-regression and moving- average. For accurate prediction, grid parameter search is applied to find the best values for the level of difference, the autoregressive order, and the moving average order parameters of the model. The configuration with a lower value of mean square error is used.

The proposed regression algorithm, Long-Short Term (LSTM) is a special kind of RNN machine learning model that's capable of learning long-term dependencies. This study uses classic LSTM for modeling. LSTM model is one type of artificial neural network which uses a special gate for removing unnecessary information in each layer.

The proposed Multi-Layer Perceptron (MLP) model is a feed-forward artificial neural network having multiple layers and non-linear activation. MLP users a backpropagation supervised learning technique for training.

The proposed Facebook Prophet (FBProphet) model is a machine learning model from Facebook for time-series forecasting that's capable of handling seasonality and holiday effects better. This study uses Prophet's for modeling. FBProphet model is a type of Prophet which handles seasonality in time-series data better. For better model performance a grid parameter searches to find the best value for changepoint_prior_scale and seasonality_prior_scale is performed. The values for the parameters which give a better model performance and lower mean square error are used.

4.3 Models Evaluation and Testing

To test and estimate the learning ability of machine learning models trained on the COVID-19 cases dataset. The basic concern of the machine learning model is to find an accurate estimation of the generalization error of the trained models on finite datasets. Because of this reason this study used different performance evaluation metrics appropriate for models, these metrics are MAE, RMSE, and R^2, as discussed in chapter three.

CHAPTER FIVE

5 Implementation and Experimentation

This chapter presents the implementation of the proposed solution for COVID-19 Spread Prediction in Ethiopia by using the methodology discussed in chapter three. Also, implement the proposed solution in chapter four. This study follows an experimental approach to determine the best machine learning algorithm for COVID-19 spread in Ethiopia prediction. Moreover, performs experiments by using each dataset.

5.1 Implementation Environment

This study used several development tools and packages to implement the proposed solution for COVID-19 spread prediction. This study uses python programming language for implementing and experimenting with each proposed solution from loading the data to the model building and evaluation phase. Also, to evaluate the implemented proposed regression model. Python is used because of its programming language choice for developers, researchers, and data scientists who need to work in machine learning models. Table 5-1 shows the list of tools and python packages with their version and description which is used in this study.

Table 5-1 Description of the Tools and Python Packages Used

Tools		
Name	***Version***	***Description***
Anaconda Navigator	1.9.12	allows us to launch development applications and easily manage conda packages, environments without the need to use commands.
Jupyter notebooks	6.0.3	has a web application and notebook document components for creating and sharing notebook documents containing live codes, visualizations, equations, and narrative text. Used for data

		cleaning, data transformation, statistical modeling, numerical simulation, data visualization, and machine learning.
Python	3.8.3	easy to learn, powerful programming language to develop a machine learning application.
Microsoft Excel	2016	Used for data preparation tasks in cleaning filtering and sorting the gathered data.
Packages		
Scikit-learn	0.23.1	A set of simple and efficient python modules for machine learning and data mining. This study uses it for training and testing models. The name of the package is called sklearn.
pandas	1.0.5	High-performance, easy-to-use data structures, and data analysis tools. This study uses it for data reading, manipulation, writing, and handling the data frame.
NumPy	1.18.5	Array processing for numbers, strings, and objects. This study uses it for handling numeric data for training and testing the model.
matplotlib	3.2.2	Publishing quality figures in python. This study uses it for data and results visualization.
statsmodels	0.11.1	Provides functions and classes for estimation of statistical models, statistical data exploration, and conducting statistical tests.
Keras	2.4.3	Provides the required abstraction and building blocks for developing machine learning solutions with high iteration velocity.
fbprophet	0.7.1	Used for forecasting time-series data based on an additive model with non-linear trends, seasonality, and holiday effects. This study uses it for forecasting with Prophet.

5.2 Deployment Environment

The tools which are discussed above in section 5.1 have been deployed on a personal computer equipped with a processor Intel® Core™ i7-75000U, CPU 2.90GHz, 2 Core(s), 8 Gigabytes of physical memory, 250 Gigabyte hard disk storage capacities with Windows 10 Pro, 64 bits operating system.

5.2.1 Dataset Description

The datasets used in this study are taken from the Federal Ministry of Health, Ethiopia [6]. This data is not constant as the number of cases increases or decreases daily. It consists of 304 individual time-series data of confirmed, death, and recovered cases in the country since March 13, 2020. Data from March 13, 2020, to January 10, 2021, is taken for study purposes. Data is split into training data from March 13, 2020, to November 10, 2020, and testing data from November 11, 2020, to January 10, 2021.

The detail for building the datasets is discussed in chapter three. The dataset contains the Date, New confirmed cases, New deaths, and New recovered columns which are used for modeling the prediction. Each dataset contains the date column as common and has different columns in each dataset. The columns are:

- **Date** is the date COVID-19 cases are reported
- **Daily_new_cases** contain the number of new cases reported on a particular day.
- **Daily_death_cases** are the number of people who died from COVID-19 on that day.
- **Daily_recovered_cases** are the number of patients who recovered that day.

5.2.2 COVID-19 Trends in Ethiopia

We study confirmed cases, recovered cases, and death cases from COVID-19 in Ethiopia. The new confirmed cases of coronavirus in Ethiopia from 13th March 2020 to 10th January 2021 are shown in Figure 5-1. The figure clearly shows that Ethiopia has the highest confirmed cases. The graph is not flattening yet, this shows the cases from the virus are still increasing.

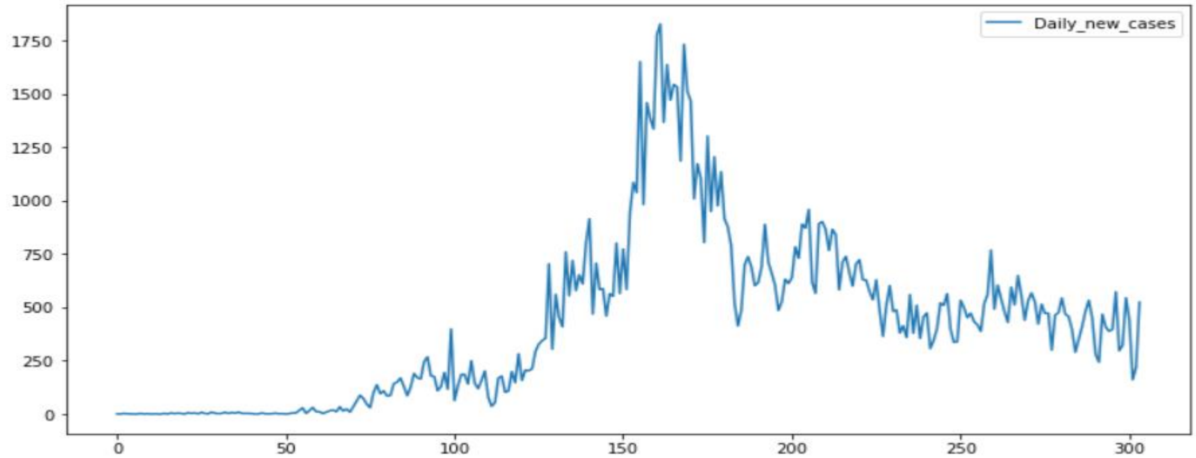

Figure 5-1 Trend of COVID-19 confirmed cases in Ethiopia.

Figure 5-1 shows the trend of the confirmed cases of coronavirus spread from 13 March 2020 to 10 January 2021in Ethiopia. The trend explains that the growth of the virus is almost exponential after mid-August 2020.

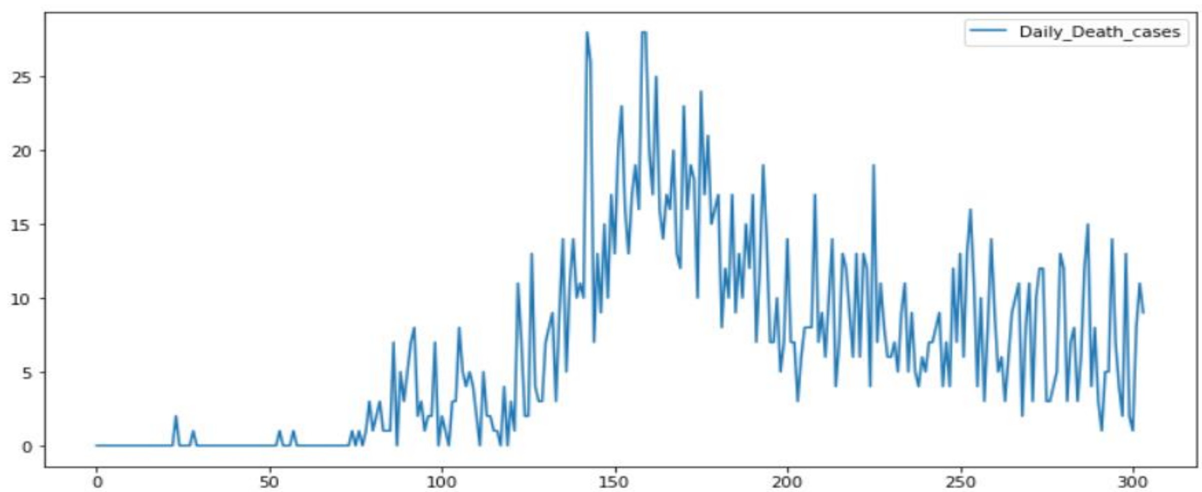

Figure 5-2 Trend of the Death rate of COVID-19 cases in Ethiopia.

Figure 5-2 above shows the death rate of COVID-19 patients in Ethiopia from 13 March 2020 to 10 January 2021. The figure shows that the death rate from the virus is still high. The virus has taken the lives of many people in Ethiopia.

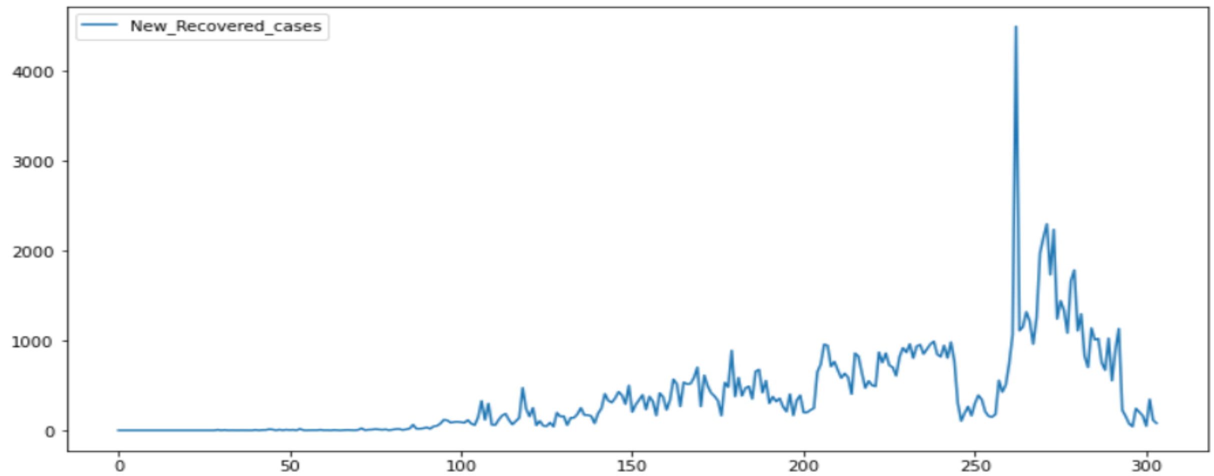

Figure 5-3 Recovery rate of COVID-19 case in Ethiopia

Figure 5-3 above shows the recovery rate of COVID-19 in Ethiopia from 13 March 2020 to 10 January 2021. From the figure, the highest recovery is recorded around the middle of November 2020. Various studies [48], [49] show that the disease recovers automatically after some time but causes major health problems that can lead to death, if not taken care of.

5.3 Loading and Reading the Data Implementation

To implement the COVID-19 spread data loading and reading, the study used a python programming language and modules. To perform the whole implementation of the methods, we perform the following common activities, which are importing important libraries packages and load the dataset file to the disk using the code in figure 5-4. The loaded dataset using panda is used throughout the whole implementation of this study.

```
df= pd.read_csv('C:/Users/ADONI/Desktop/New folder/dataset.csv',header=0,
                parse_dates=[0], index_col=0, date_parser=parser)
```

Figure 5-4 Python Code for Loading the Dataset

5.3.1 Implementation to Read the DateTime Column

To use the dataset with a date, month, and year column, we implement a method that performs strip time to use the date column. The method is written using the Python strptime module. The method makes the date column readable for the methods. The output from this process is a readable column. The COVID-19 data loading method uses a Python DateTime module.

```
def parser(x):
    return datetime.strptime(x, '%m/%d/%Y')
```

Figure 5-5 Python Code for Reading the date column

5.4 Machine Learning Models Implementations

To build a machine learning model using the proposed algorithm, we used a python pandas library packages and followed the conventional method that is importing the important library package for modeling and metrics for model evaluation.

```
from math import sqrt
from numpy import array
from numpy import mean
from numpy import std
from datetime import datetime
import numpy as np
from pandas import DataFrame
from pandas import concat
from pandas import read_csv
from sklearn.metrics import mean_squared_error
from keras.models import Sequential
from keras.layers import Dense
from matplotlib import pyplot
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
from statsmodels.tools.eval_measures import rmse
from sklearn.metrics import r2_score
from fbprophet import Prophet
from fbprophet.diagnostics import cross_validation
from fbprophet.diagnostics import performance_metrics
from fbprophet.plot import plot_cross_validation_metric
from statsmodels.tsa.arima_model import ARIMA
from pandas.plotting import autocorrelation_plot
```

Figure 5-6 Importing the Important Package for Modeling

After loading, the data is classified as a training set and test set. Then the model is trained by using the training dataset, testing is performed using the testing dataset.

```
X = newcase.values
train = X[0:249]
test = X[249:]
```

Figure 5-7 Code for Preparing Dataset with train and test split for Models Training

To find the best value for ARIMA model parameter values a grid search is performed, the result gives the values of the level of difference, the autoregressive order, and the moving average order in the model prediction accuracy is high. It is clear from the trends in section 5.3.1 that data of confirmed, death and recovered cases are not in stationary form. A grid parameter search for the ARIMA model is performed as shown in Appendix A.1.

```
# evaluate combinations of p, d and q values for an ARIMA model
def evaluate_models(dataset, p_values, d_values, q_values):
    dataset = dataset.astype('float32')
    best_score, best_cfg = float("inf"), None
    for p in p_values:
        for d in d_values:
            for q in q_values:
                order = (p,d,q)
                try:
                    mse = evaluate_arima_model(dataset, order)
                    if mse < best_score:
                        best_score, best_cfg = mse, order
                    print('ARIMA%s MSE=%.3f' % (order,mse))
                except:
                    continue
    print('Best ARIMA%s MSE=%.3f' % (best_cfg, best_score))
```

Figure 5-8 Code for Grid Parameter search for ARIMA model

This study used the ARIMA() method of the statsmodels package to building the ARIMA model. This method is instantiated with the training dataset and order which have a level of difference, the autoregressive order, and the moving average order to predict the future.

```python
model = ARIMA(newcase, order=(1,0,1))
model_fit = model.fit(disp=0)
```

Figure 5-9 Instantiating ARIMA and Fit the Model

To implement the LSTM model, the study uses a Keras *LSTM* () method. This model is suitable for time-series data. This method has four parameters which are training dataset, batch size, epochs (how many times the learning algorithm will work through the entire training dataset), and neurons.

```python
def fit_lstm(train, batch_size, nb_epoch, neurons):
    X, y = train[:, 0:-1], train[:, -1]
    X = X.reshape(X.shape[0], 1, X.shape[1])
    model = Sequential()
    model.add(LSTM(neurons, batch_input_shape=(batch_size, X.shape[1], X.shape[2])))
    model.add(Dense(1))
    model.compile(loss='mean_squared_error', optimizer='adam')
    for i in range(nb_epoch):
        model.fit(X, y, epochs=1, batch_size=batch_size, verbose=0, shuffle=False)
        model.reset_states()
    return model
```

Figure 5-10 Instantiating LSTM and Fit the Model

To implement the MLP model, the *fit_model()* function is defined which has training data, batch size, epoch, and the number of neuron parameters. It uses relu (rectified linear unit) activation function and adam optimizer.

```python
# fit an MLP network to training data
def fit_model(train, batch_size, nb_epoch, neurons):
    X, y = train[:, 0:-1], train[:, -1]
    model = Sequential()
    model.add(Dense(neurons, activation='relu', input_dim=X.shape[1]))
    model.add(Dense(1))
    model.compile(loss='mean_squared_error', optimizer='adam')
    model.fit(X, y, epochs=nb_epoch, batch_size=batch_size, verbose=0, shuffle=False)
    return model

model = fit_model(train_trimmed, batch_size, epochs, neurons)
```

Figure 5-11 Instantiating MLP and Fit the Model

To find the best value for FBProphet model parameters a grid search is performed, the result gives the values of the level of changepoint_prior_scale and seasonality_prior_scale in which the model prediction accuracy is high and RMSE is low. The grid parameter search is performed as shown in Appendix A.2.

```
param_grid = { 'changepoint_prior_scale': [0.001, 0.01, 0.1, 0.5],
               'seasonality_prior_scale': [0.01, 0.1, 1.0, 10.0], }
all_params = [dict(zip(param_grid.keys(), v)) for v in itertools.product(*param_grid.values())]
rmses = []
df = pd.read_csv('C:/Users/ADONI/Desktop/Covid/My DS/New2.csv')
df.rename(columns={'Date':'ds','Daily_new_cases':'y'},inplace=True)
cutoffs = pd.to_datetime(['2020-03-20', '2020-07-20', '2020-11-20'])
for params in all_params:
    m = Prophet(**params).fit(df)
    df_cv = cross_validation(m, cutoffs=cutoffs, horizon='7 days', parallel="processes")
    df_p = performance_metrics(df_cv, rolling_window=1)
    rmses.append(df_p['rmse'].values[0])
tuning_results = pd.DataFrame(all_params)
tuning_results['rmse'] = rmses
print(tuning_results)
```

Figure 5-12 Code for Grid Parameter search for FBProphet model

To implement the Prophet model, the *Prophet()* method of the fbprophet package is used. This study used the *Prophet() method* of the fbprophet package for building the Prophet model. This method has growth, seasonality_mode, parameters.

```
# Initialize the Model
model = Prophet(growth='linear',seasonality_mode='multiplicative',
                daily_seasonality=True,weekly_seasonality=False,
                yearly_seasonality=False )
```

Figure 5-13 Instantiating prophet and Fit the Model

5.5 Model Testing and Evaluation

The study used *rmse(), mean_absolute_error(),* and *r2_score()* methods, which are used to evaluate the performance of the built models using the entire dataset. These methods give a result of evaluation metrics such as coefficient of determination, root mean squared error, mean squared error, and mean absolute error value of the model under testing.

CHAPTER SIX

6 Result and Discussions

In this chapter, the result of the experiments of the proposed solution for COVID-19 spread prediction using machine learning is discussed. Here the result of each model in each dataset is discussed. Finally, we discuss predicting the accuracy of adopted models for daily confirmed cases, daily death cases, and daily recovered cases.

6.1 Model Evaluation Results

The dataset is split into training and test set. Train the models using the training dataset and testing is done on the testing set. The result from the models is evaluated using RMSE, MAE, and R^2 performance metrics. The results obtained by each model on each COVID-19 case are presented below.

6.1.1 Predicting Daily Confirmed Cases

To predict and analyze the daily confirmed cases of the disease, we have performed an evaluative study of the adopted models using confirmed data of the 304 days in Ethiopia. So, we have performed stationarity techniques to evaluate the ARIMA model. We have applied LSTM, MLP, and FBProphet directly on actual data to fit the model and generated the forecasting results. Table 6.1-1 shows the performance results of the models for the confirmed cases. Best MAE results are 92.65, 122.61, 158.35, 288.77 for Ethiopia COVID-19 confirmed cases data by FBProphet, MLP, LSTM, and ARIMA respectively. Results show that FBProphet prediction almost matches the actual values, MLP results in the second-best result whereas ARIMA and LSTM did not perform as well. We can see that the R^2 value of FBProphet is 0.95 which is very much acceptable to generate forecasting results, whereas the R^2 for ARIMA, LSTM, MLP are 0.55, 0.82, and -0.27 respectively.

The prediction models are trained on the training set of the dataset. The result of each performance measure for ARIMA, LSTM, MLP, and FBProphet models for daily confirmed cases are presented in the table below.

Table 6-1 Performance results of the models for COVID-19 Confirmed cases in Ethiopia

Model	***RMSE***	***MAE***	R^2
ARIMA (0,0,1)	288.77	212.68	0.55
LSTM	158.35	108	0.82
MLP	122.61	94.31	-0.27
FBProphet	92.65	62.44	0.95

The results in Table 6-1 show the performance measures of the models for the daily confirmed cases column in the dataset. The worst performance results (RMSE= 288.77, MAE=212.68, and R^2=0.55) recorded for the ARIMA model, and the best performance result (RMSE=92.65, MAE=62.44, and R^2=0.95) recorded for the FBProphet model.

Predicted and actual data are plotted together in the figures below to visualize the fitting accuracy of the models. Figure 6-1, Figure 6-2, Figure 6-3, and Figure 6-4 show the predicted and actual values of ARIMA, LSTM, MLP, and FBProphet models respectively. We can see that the FBProphet model can fit well in the case of daily confirmed cases data in Ethiopia as shown in Figure 6-4.

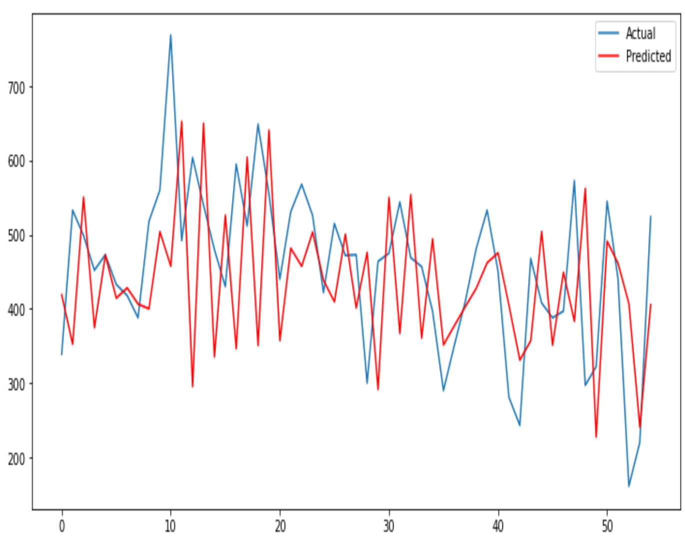

Figure 6-1 ARIMA Forecasting for Daily confirmed cases

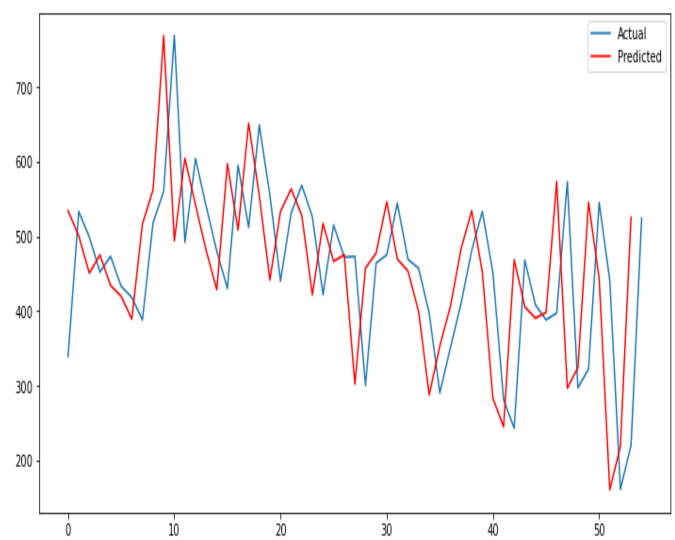

Figure 6-2 LSTM Forecasting for Daily confirmed cases

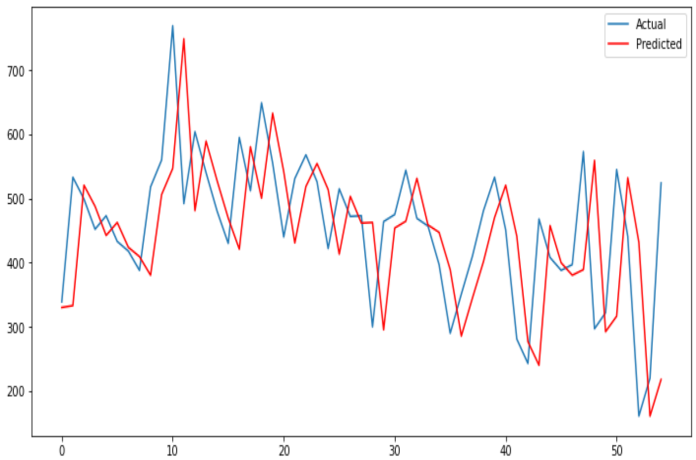

Figure 6-3 MLP Forecasting for Daily confirmed cases

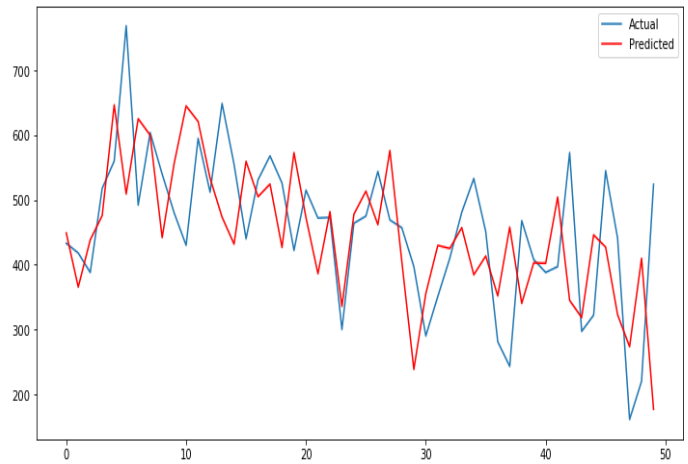

Figure 6-4 FBProphet Forecasting for Daily confirmed cases

6.1.2 Predicting Daily Death Cases

Coronavirus has taken many lives. So, it is necessary to analyze the death rate of the virus and predicting future cases that can guide the government to take measures in advance. In this section, we have evaluated the forecasting models for death cases in Ethiopia. We have converted the non-stationary death cases dataset into a stationary form to fit the ARIMA model. FBProphet, LSTM, MLP model is applied to the actual data to forecast the prediction results. Table 6-2 shows the prediction results of the models for death cases in Ethiopia. We can see that prediction errors of FBProphet are very less, LSTM and ARIMA are intermediate whereas MLP prediction has a high error factor in the results. The results suggest that FBProphet can be used for actual forecasting of the death cases to plan the facilities accordingly.

ARIMA, LSTM, MLP, and FBProphet models have been applied to the daily death cases column of the dataset. The models are trained on the training set from the dataset. The result of each performance measure for ARIMA, LSTM, MLP, and FBProphet models are presented in the table below.

Table 6-2 Performance results of the models for COVID-19 death cases in Ethiopia

Model	***RMSE***	***MAE***	R^2
ARIMA (4,0,1)	3.59	2.41	0.72
LSTM	4.97	3.71	0.37
MLP	4.95	4.18	-0.48
FBProphet	2.68	1.95	0.83

The results in Table 6-2 show the performance measures of the models on the daily death cases column of the dataset. The best performance results (RMSE=2.68, MAE=1.95, and R^2=0.83) were recorded for the FBProphet model, and the worst performance result (RMSE=4.95, MAE=4.18, and R^2=-0.48) recorded for the MLP model.

Figure 6-5, Figure 6-6, Figure 6-7, and Figure 6-8 below show the results by all models in death cases data. Forecasted and actual data are plotted together to visualize the fitting accuracy of the models. We can see that the FBProphet model can fit well as shown in Figure 6-8. FBProphet handles outliers during modeling and forecasting better. The results also depict that FBProphet can fit well to model and predict the results.

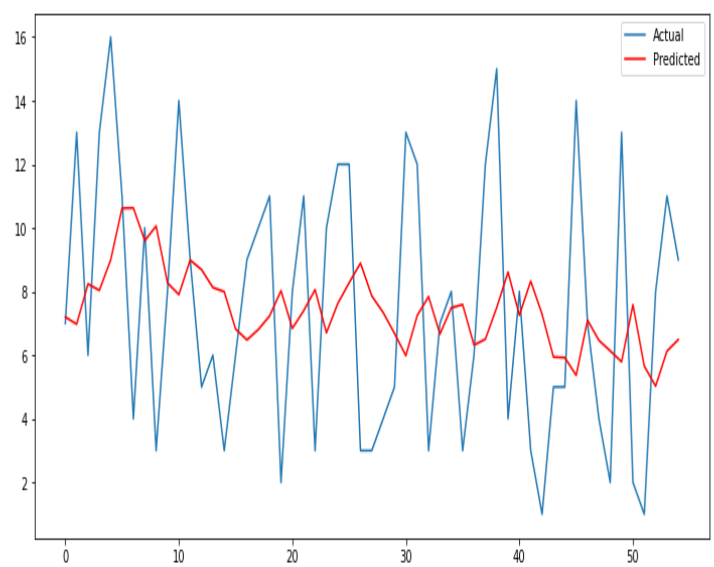

Figure 6-5 ARIMA Prediction for Daily death cases

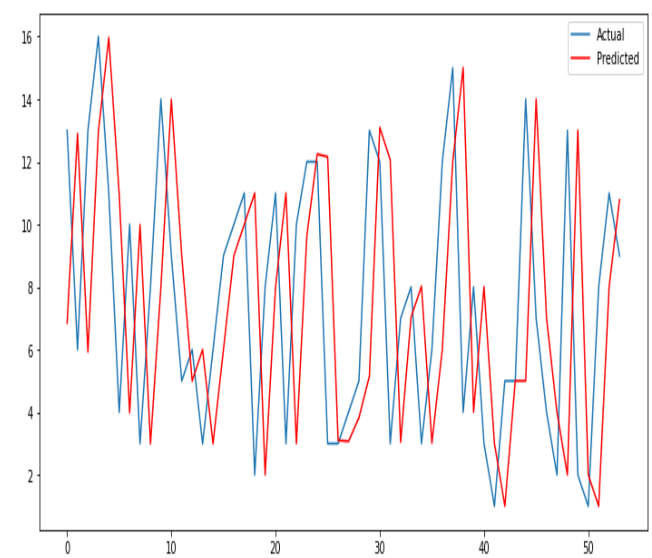

Figure 6-6 LSTM Prediction for Daily death cases

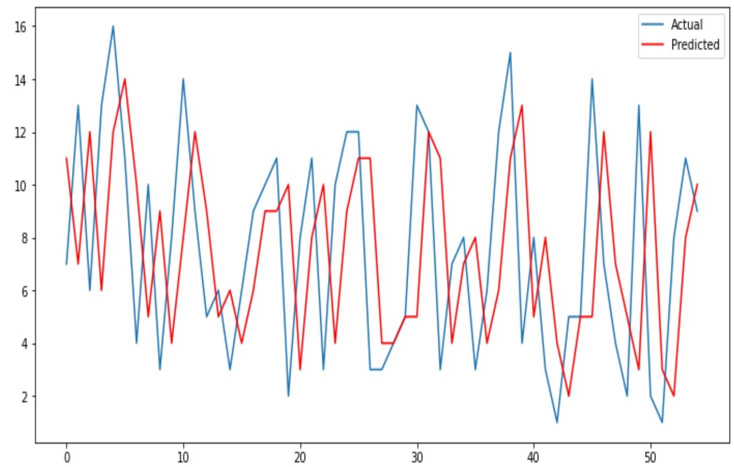

Figure 6-7 MLP Prediction for Daily death cases

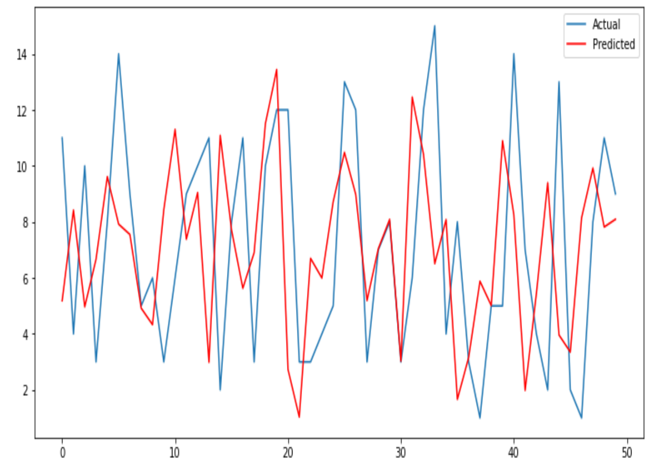

Figure 6-8 FBPhrophet Prediction for Daily death cases

6.1.3 Predicting Daily Recovered Cases

Recovered cases are the number of people who were confirmed before and were under medical supervision and are free from the virus. We use ARIMA, LSTM, MLP, and FBProphet models to predict the recovered cases. ARIMA can be used for prediction if data is stationary. The recovered case data is converted to stationary form. We also have used ACF plots to identify appropriate values of q and p order of ARIMA. The FBProphet, LSTM, and MLP are applied directly to actual data. Performance results for predicting recovered cases of Ethiopia are shown in Table 6-3. We have mentioned the order of ARIMA along with the performance results in the tables. The FBProphet performed better than the other models in fitting the data. The Best MAE scores are 44.69 and 77.83 FBProphet and ARIMA respectively. From the results, we can clearly say that FBProphet has far better performance as compared to the other models for all types of error measures which are RMSE, and R^2. MLP prediction has a high error factor in the results. The results suggest that FBProphet can be used for actual forecasting of the cases to plan the services accordingly.

The result of ARIMA, LSTM, MLP, and FBProphet models have been applied to the daily recovered cases column of the dataset. The models are trained on the daily recovered cases column of a training set from the dataset. The result of each performance measure for ARIMA, LSTM, MLP, and FBProphet models are presented in the table below.

Table 6-3 Performance results of the models for COVID-19 recovered cases in Ethiopia

Model	***RMSE***	***MAE***	***R^2***
ARIMA (1,0,0)	123.24	77.83	0.82
LSTM	360.57	149.54	0.49
MLP	705	352.6	0.18
FBProphet	75.67	44.69	0.98

The results in Table 6-3 show the performance measures of the models on the daily recovered cases dataset. The best performance results (RMSE=75.67, MAE=44.69, and R^2=0.98) were recorded for the FBProphet model, and the worst performance result (RMSE=705, MAE=352.6, and R^2=0.18) recorded for the MLP model.

Predicted and actual data are plotted together to visualize the fitting accuracy of the models in Figure 6-9, Figure 6-10, Figure 6-11, and Figure 6-12 by all models in recovered cases data. We can see that the FBProphet model can fit well as shown in Figure 6-12 below. The results also depict that FBProphet can fit well to model and predict the results.

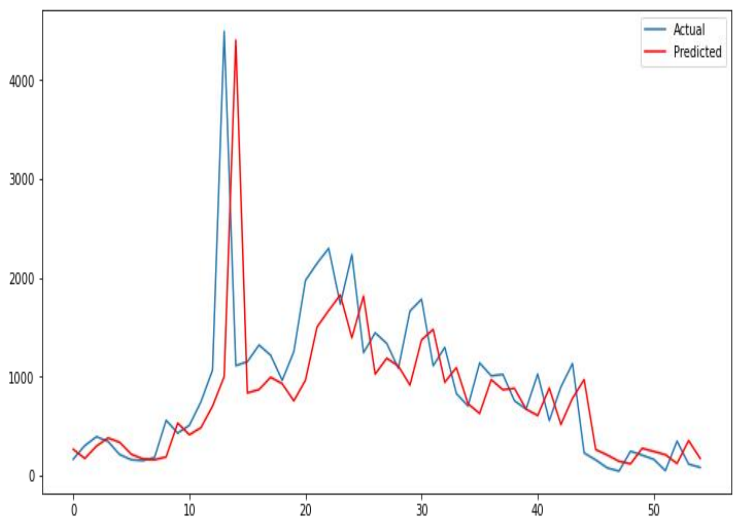

Figure 6-9 ARIMA model performance for Daily recovered cases

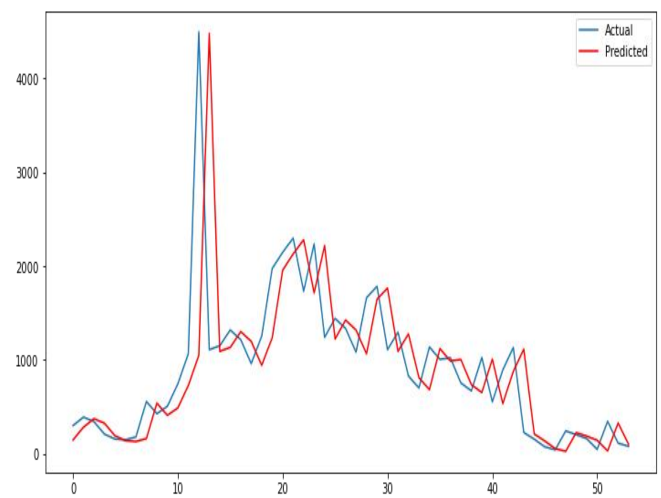

Figure 6-10 LSTM model performance for Daily recovered cases

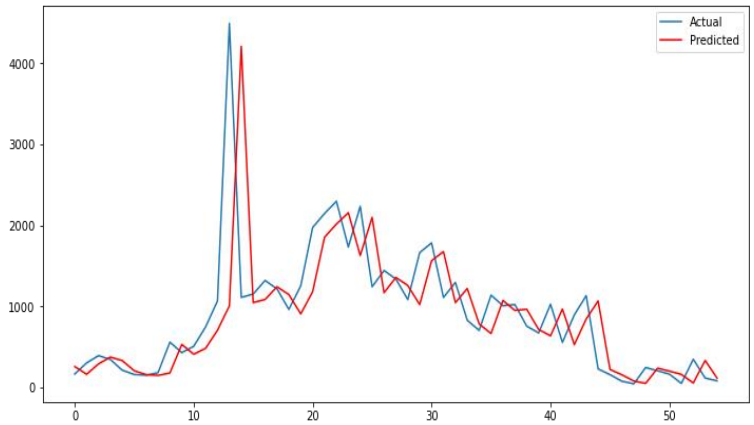

Figure 6-11 MLP model performance for Daily recovered cases

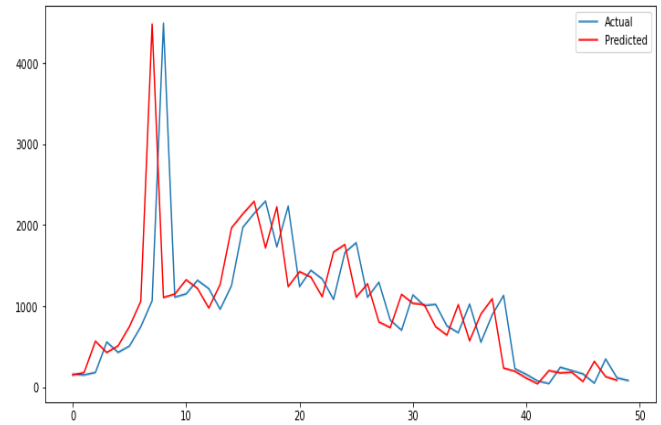

Figure 6-12 FBProphet model performance for Daily recovered cases

This study uses a different model performance evaluation metrics, as discussed in chapters three and four. These metrics are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of determination (R^2). Table 6-4 below shows the performance of all performance metrics on each dataset.

Table 6-4 Summary of all models results in each dataset

Cases	***Model***	***Performance Metrics***		
		RMSE	***MAE***	***R²***
Confirmed Cases	ARIMA (0,0,1)	288.77	212.68	0.55
	LSTM	158.35	108	0.82
	MLP	122.61	94.31	-0.27
	FBProphet	92.65	62.44	0.95
Death Cases	ARIMA (4,0,1)	3.59	2.41	0.72
	LSTM	4.97	3.71	0.37
	MLP	4.95	4.18	-0.48
	FBProphet	2.68	1.95	0.83
Recovered Cases	ARIMA (1,0,0)	123.24	77.83	0.82
	LSTM	360.57	149.54	0.49
	MLP	705	352.6	0.18
	FBProphet	75.67	44.69	0.98

Table 6-4 shows the results of the evaluation metric for each model on the dataset. MAE metrics selected to compare the models based on the reason that it shows how much is the error than RMSE when the dataset contains outliers as described in chapter three.

Based on the performance measures, the FBProphet model on New death cases data obtains a higher score of 1.95 than LSTM, MLP, and ARIMA models. As shown in Table 6-1 above the FBProphet model on the New confirmed cases data obtains a higher score of 62.44 than LSTM, MLP, and ARIMA models. For recovered cases, the FBProphet model obtained the best result of 44.69 in MAE measurement. A slightly higher result is recorded for ARIMA than LSTM and LSTM models.

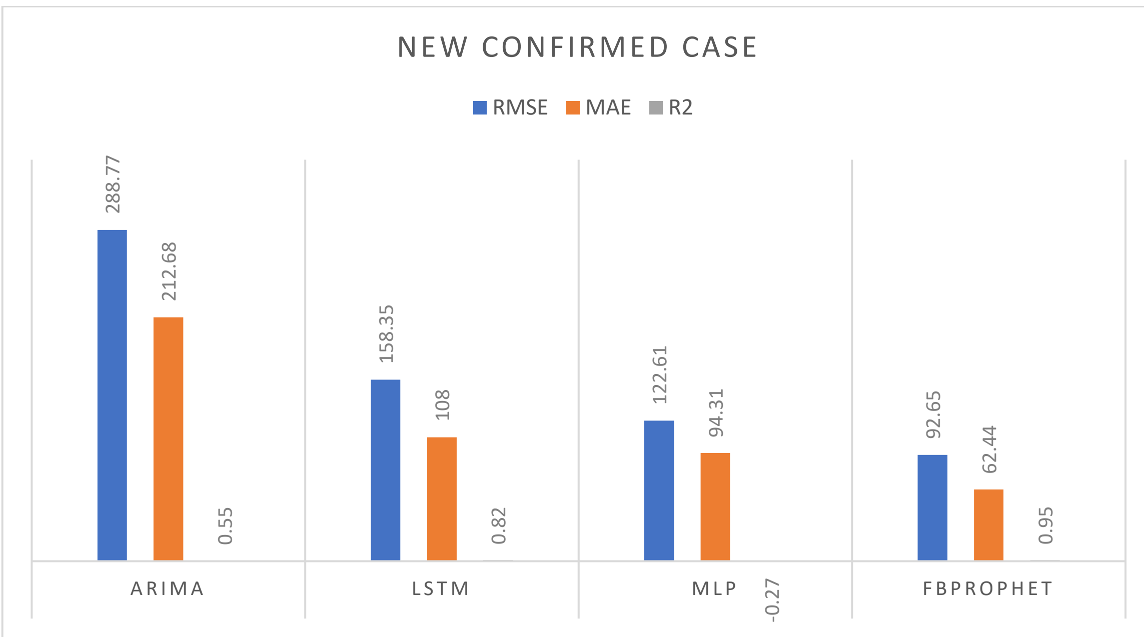

Figure 6-13 Comparison of all methods on New confirmed cases

From Figure 6-13 above it is seen that the FBProphet model shows better performance for predicting daily confirmed cases. Above all the coefficient of determination also known as the R_squared evaluation metric is 0.95 which is the highest, this indicates that the selected model fits better for the daily confirmed cases data.

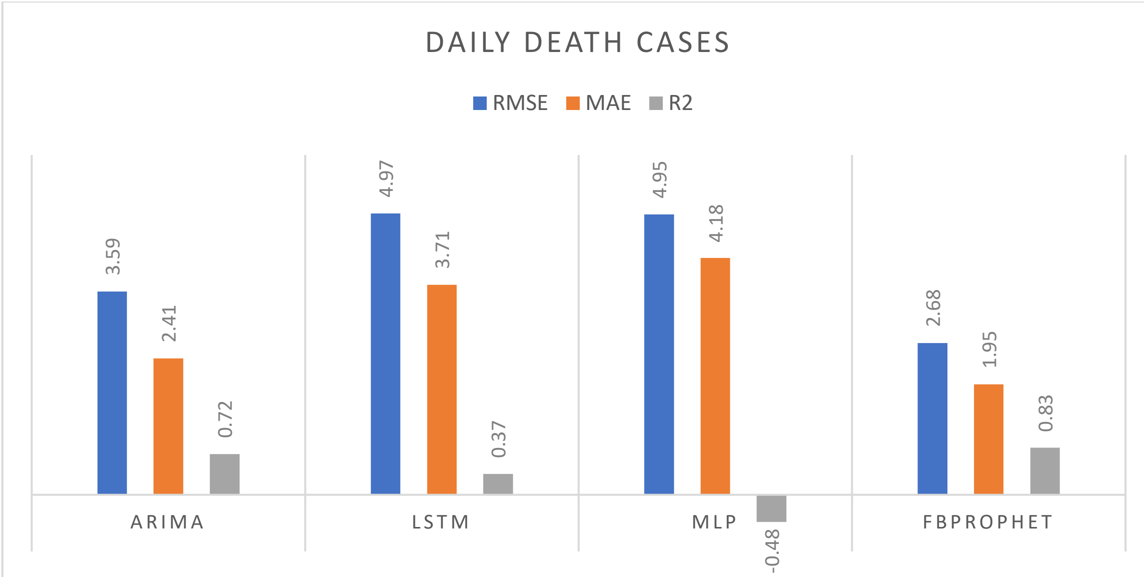

Figure 6-14 Comparison of all methods on daily death cases

The above Figure 6-14, proves that the FBProphet model gives better prediction results and fits the data better than ARIMA, LSTM, or MLP models tested for predicting daily death cases from COVID-19 in Ethiopia.

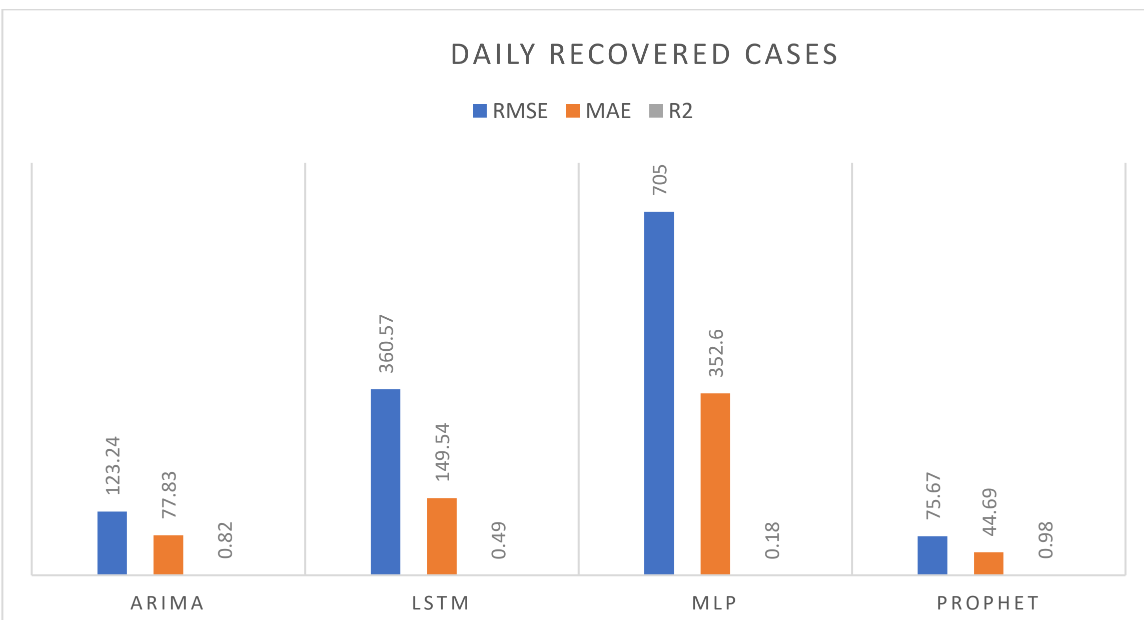

Figure 6-15 Comparison of all methods on daily recovered cases

As shown in Figure 6-15 above FBProphet model has better evaluation results for predicting daily recovered cases than the other three models.

The results of this study show that FBProphet performs better than LSTM, MLP, and ARIMA for all columns (new confirmed cases, new death cases, and recovered cases) of the dataset. ARIMA model gives the second-best performance result for daily death and daily recovered cases. For daily confirmed cases LSTM model is the second-best algorithm. The problem of error in prediction occurs in all proposed models, as we have discussed earlier. This occurs due to several reasons but the major one is the nature of the data which is outliers and zero-valued entry. Finally, these discussed problems are reflected in the prediction performance of the proposed models.

6.1.4 Predicting the future 7-day confirmed Cases

For forecasting the future 7-days confirmed cases, training is performed on the whole dataset which is from March 13, 2020, to January 10, 2021. Forecasting is done by ARIMA, LSTM, and FBProphet models. The result obtained from each model is shown in the table below. As shown in Table 6-5 below the ARIMA model gives forecasting close to the actual value for daily confirmed cases in Ethiopia.

Table 6-5 Confirmed cases future prediction results

Date	***Actual Value***	***ARIMA predicted value***	***LSTM predicted value***	***FBProphet predicted value***
1/11/2021	300	388	223	177
1/12/2021	376	387	477	259
1/13/2021	463	386	584	346
1/14/2021	467	386	122	687
1/15/2021	404	385	422	840
1/16/2021	446	385	796	1036
1/17/2021	423	384	193	1215

6.1.5 Predicting the future 7-day Death Cases

Training for forecasting the coming seven days' daily death cases is done using the entire dataset with 304 instances. ARIMA, LSTM, and FBProphet models are used for forecasting the upcoming week's death cases. The result obtained from each model is shown in the table below. As shown in Table 6-6 below the ARIMA model gives forecasting close to the actual value for daily confirmed cases in Ethiopia. The second-best forecasts are given by the FBProphet model. LSTM gives forecasts which are very far from the actual value.

Table 6-6 Death cases future prediction results

Date	***Actual Value***	***ARIMA predicted value***	***LSTM predicted value***	***FBProphet predicted value***
1/11/2021	9	6	10	8
1/12/2021	1	7	-2	17
1/13/2021	2	7	2	15
1/14/2021	2	7	12	14
1/15/2021	15	7	18	16
1/16/2021	6	7	-7	24
1/17/2021	1	7	1	22

6.1.6 Predicting the future 7-days Recovered Cases

Future recovered cases forecasting for 7-days is performed using ARIMA, LSTM, and FBProphet models trained on all instances of the dataset. ARIMA, LSTM, and FBProphet models are used for forecasting the upcoming week recovered cases. The result obtained from each model is shown in the table below. As shown in Table 6-7 below the ARIMA model gives forecasting close to the actual value for daily confirmed cases in Ethiopia. LSTM model forecasted the future poorly including negative prediction.

Table 6-7 Daily recovered cases future prediction results

Date	***Actual Value***	***ARIMA predicted value***	***LSTM predicted value***	***FBProphet predicted value***
1/11/2021	189	144	822	562
1/12/2021	699	194	1050	497
1/13/2021	305	233	194	0
1/14/2021	182	263	356	255
1/15/2021	679	286	34	510
1/16/2021	617	304	-7	1128
1/17/2021	102	318	129	1165

6.2 Discussion

MSE and MAE performance evaluation metrics are used for comparing the performance on the previous work by Ismail et al. [23] that used seven countries' daily cases dataset for forecasting COVID-19 spread and also to compare the ability of models to predict the spread. This study experimented on ARIMA, LSTM, MLP, and FBProphet machine learning algorithms to model COVID-19 spread prediction using daily confirmed, daily death, and daily recovered cases of the dataset of Ethiopia evaluated by RMSE, MAE, and R^2.

FBProphet, LSTM, and MLP models are applied directly to actual data. The ARIMA model needs the data to be stationary before used for prediction, so the data in each dataset is converted to stationary form. The ACF plots are used to identify appropriate values of the number of lagged forecast errors (q) and the number of autoregressive terms (p) of ARIMA. A grid parameter search is done for finding the best values for FBProphet parameters (change prior scale and seasonality prior scale). FBProphet adopts successively progression and avoids outliers during modeling and forecasting. The results also show that FBProphet fits well in case of fewer data whereas the other models require sufficient data to model and predict.

From the performance measures, the R^2 value tells us how the model fits with the data. The FBProphet model fits better than all the other models with an R^2 value of 0.95, 0.83, and 0.98 for daily confirmed cases, death cases, and recovered cases. Tarik Chafiq et al. [31], claim that they obtain R^2=0.99 for 49 days cumulative daily cases. The large range of the coefficient of determination (R^2) result shows that the model fitting the data changes as the size of the dataset grows. When we looked at the other model's coefficient of determination results, it indicates that the number of data makes a great difference in the prediction performance of the model.

As shown in the results section of this study, Machine learning approaches are effective for the daily confirmed, death, and recovered COVID-19 cases prediction in Ethiopia. Among all the models proposed FBProphet is the best model for predicting COVID-19 cases in Ethiopia, LSTM is the second-best in confirmed cases prediction and ARIMA model is the second-best in death cases and recovered cases prediction as measured by the performance evaluation metrics.

From the results, for COVID-19 Spread prediction in Ethiopia, the Facebook Prophet model has far better performance as compared to ARIMA, LSTM, and MLP models for all types of performance measures which are MAE, RMSE, and R^2. ARIMA has low performance on daily confirmed cases prediction and MLP has low performance on daily death and recovered cases prediction. The results suggest that FBProphet can be used for the actual prediction of the cases to plan the services accordingly.

For forecasting the future seven days the whole dataset from March 13, 2020, to January 10, 2021, is used for training. The results show that the ARIMA model gives a better forecast for daily confirmed, death, and recovered COVID-19 cases in Ethiopia. The FBProphet model gives the second-best forecast and LSTM gives a very far forecast from the actual value for all cases.

COVID-19 spread prediction performance can be improved by increasing the quality and size of the data to be trained on and performing hyperparameter tuning for all models.

CHAPTER SEVEN

7 Conclusion, Recommendation, and Future work

In this Chapter, the proposed COVID-19 spread prediction using machine learning and the finding of this research is concluded. It also described the recommendation and future research directions.

7.1 Conclusion

This research proposed to develop a solution to COVID-19 spread prediction using machine learning approaches. This study attempted to implement and compare machine learning and deep learning methods specifically for COVID-19 spread prediction in Ethiopia. To successfully execute the study, it was essential to understand and define COVID-19, explore existing various techniques used to tackle the problem, understand the COVID-19 spread, as discussed in chapter two. Also, the different methods followed to implement and design models that have the capability of predicting the spread. These methods include collecting daily new confirmed, new death, and new recovered cases for building the dataset, model training using ARIMA, LSTM, MLP, and FBProphet models, and testing the performance of the models. Finally performs model comparisons based on evaluation metric results.

In this study, we manually collected daily reports of COVID-19 new cases, new deaths, and new recovered cases in Ethiopia. This process resulted in a dataset with 304 instances of confirmed, death, and recovered cases. Based on the dataset, the models developed using ARIMA, LSTM, MLP, and FBProphet are implemented. The experiment is performed using the dataset. The FBProphet model on New confirmed cases, daily death cases, and daily recovered cases resulted in better performance than ARIMA, LSTM, and MLP. ARIMA performs slightly better than LSTM, MLP, and FBProphet models on daily death and daily recovered cases and LSTM performs slightly better in confirmed cases.

Finally, the FBProphet model performs better than LSTM, MLP, and ARIMA models in the new confirmed case, new death, and recovered cases data of the dataset used in this research. So FBProphet model is the best of our knowledge to use in COVID-19 spread prediction in

Ethiopia study. For forecasting, the future 7-days confirmed, death, and recovered cases the ARIMA model gives better results compared to the other models.

7.2 Recommendations

The proposed prediction models have shown some capacity of predicting confirmed, death, and recovered COVID-19 cases in Ethiopia. As a result, the researcher recommends COVID-19 treatment centers to develop future forecasting models that assist the healthcare system to avoid the deaths and infections that are forecasted.

This study proposed a solution using machine learning approaches to build a model for COVID-19 spread in Ethiopia. It's better to build the prediction models for the ten regions and two city administrations.

7.3 Future works

This study is limited only to the confirmed, death, and recovered cases, for a more inclusive prediction model, future research can focus on developing a model for some other cases such as cases in the Intensive care unit and active cases (currently in treatment) in treatment. Future research could focus on such types of content.

* * *

*This research was funded by Adama Science and Technology University with a grant number **ASTU/SM-R/097/19**.*

References

[1] “Constitution.” [Online]. Available: https://www.who.int/about/who-we-are/constitution. [Accessed: 15-Feb-2021].

[2] “Outbreak: 10 of the Worst Pandemics in History.” [Online]. Available: https://www.mphonline.org/worst-pandemics-in-history/?T. [Accessed: 17-Feb-2021].

[3] “WHO Coronavirus Disease (COVID-19) Dashboard | WHO Coronavirus Disease (COVID-19) Dashboard.” [Online]. Available: https://covid19.who.int/. [Accessed: 02-Dec-2020].

[4] “Coronavirus Disease 2019 (COVID-19) – Africa CDC.” [Online]. Available: https://africacdc.org/covid-19/. [Accessed: 03-Dec-2020].

[5] “WHO Director-General’s opening remarks at the media briefing on COVID-19 - 20 April 2020.” [Online]. Available: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---20-april-2020. [Accessed: 03-Dec-2020].

[6] “Covid-19! |.” [Online]. Available: https://www.covid19.et/covid-19/. [Accessed: 11-Jan-2021].

[7] “Summary of COVID 19 Economic Response Measures,” pp. 1–2, 2020.

[8] Z. Shigute, A. D. Mebratie, G. Alemu, and A. Bedi, “Containing the spread of COVID-19 in Ethiopia,” *J. Glob. Health*, vol. 10, no. 1, pp. 1–4, 2020, doi: 10.7189/JOGH.10.010369.

[9] “COVID-19 vaccines.” [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines. [Accessed: 06-Jan-2021].

[10] "Coronaviruses | NIH: National Institute of Allergy and Infectious Diseases." [Online]. Available: https://www.niaid.nih.gov/diseases-conditions/coronaviruses. [Accessed: 17-Jan-2021].

[11] "What Is COVID-19? | coronavirus." [Online]. Available: https://coronavirus.dc.gov/page/what-covid-19. [Accessed: 25-Nov-2020].

[12] "WHO | What is a pandemic?," *WHO*, 2020.

[13] "How to Protect Yourself & Others | CDC." [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html. [Accessed: 25-Nov-2020].

[14] "Symptoms of Coronavirus | CDC." [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html. [Accessed: 25-Nov-2020].

[15] "WHO Coronavirus Disease (COVID-19) Dashboard | WHO Coronavirus Disease (COVID-19) Dashboard." [Online]. Available: https://covid19.who.int/. [Accessed: 25-Nov-2020].

[16] "COVID-19 situation update worldwide, as of week 1 2021." [Online]. Available: https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases. [Accessed: 20-Jan-2021].

[17] "COVID-19: Ethiopia closes schools, bans public events." [Online]. Available: https://www.aa.com.tr/en/africa/covid-19-ethiopia-closes-schools-bans-public-events/1767683. [Accessed: 16-Feb-2021].

[18] "ኮቪድ 19 መከላከያ ቅድመ ሁኔታዎችን በ3 ሳምንታት ያሟሉ ትምህርት ቤቶች በጊዜ ሰሌዳው መሰረት ትምህርት መጀመር እንደሚችሉ ተገለፀ - Welcome to Fana Broadcasting Corporate S.C." [Online]. Available: https://www.fanabc.com/ትምህርት-ቤቶችን-ዳግም-መክፈት-በሚቻልበ/. [Accessed: 16-Feb-2021].

[19] N. B. Dimitrov and L. A. Meyers, "Mathematical Approaches to Infectious Disease Prediction and Control," *Risk Optim. an Uncertain World*, pp. 1–25, 2010, doi: 10.1287/educ.1100.0075.

[20] M. Woolhouse, "How to make predictions about future infectious disease risks," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 366, no. 1573, pp. 2045–2054, 2011, doi: 10.1098/rstb.2010.0387.

[21] C. S. Lutz *et al.*, "Applying infectious disease forecasting to public health: A path forward using influenza forecasting examples," *BMC Public Health*, vol. 19, no. 1, pp. 1–12, 2019, doi: 10.1186/s12889-019-7966-8.

[22] C. Viboud and A. Vespignani, "The future of influenza forecasts," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 8, pp. 2802–2804, 2019, doi: 10.1073/pnas.1822167116.

[23] İ. Kırbaş, A. Sözen, A. D. Tuncer, and F. Ş. Kazancıoğlu, "Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches," *Chaos, Solitons and Fractals*, vol. 138, 2020, doi: 10.1016/j.chaos.2020.110015.

[24] S. F. Ardabili *et al.*, "COVID-19 Outbreak Prediction with Machine Learning," *Algorithms*, vol. 13, no. 10, p. 249, 2020, doi: 10.3390/a13100249.

[25] F. Rustam, A. A. Reshi, and A. Mehmood, "COVID-19 Future Forecasting Using Supervised Machine Learning Models," vol. 8, pp. 101489–101499, 2020, doi: 10.1109/ACCESS.2020.2997311.

[26] "What is Machine Learning? A definition - Expert System | Expert.ai." [Online]. Available: https://www.expert.ai/blog/machine-learning-definition/. [Accessed: 16-Feb-2021].

[27] "Supervised and Unsupervised Machine Learning Algorithms." [Online]. Available: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/. [Accessed: 16-Feb-2021].

[28] G. R. Shinde, A. B. Kalamkar, P. N. Mahalle, N. Dey, J. Chaki, and A. E. Hassanien, "Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art," *SN Comput. Sci.*, vol. 1, no. 4, 2020, doi: 10.1007/s42979-020-00209-9.

[29] M. Jojoa and B. Garcia-Zapirain, "Forecasting COVID 19 Confirmed Cases Using Machine Learning: the Case of America," no. September, 2020, doi: 10.20944/preprints202009.0228.v1.

[30] S. I. Alzahrani, I. A. Aljamaan, and E. A. Al-Fakih, "Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions," *J. Infect. Public Health*, vol. 13, no. 7, pp. 914–919, 2020, doi: 10.1016/j.jiph.2020.06.001.

[31] T. Chafiq, M. Ouadoud, and K. Elboukhari, "Covid-19 forecasting in morocco using fbprophet facebook's framework in python," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 5, pp. 8654–8660, 2020, doi: 10.30534/ijatcse/2020/251952020.

[32] A. E. Permanasari, I. Hidayah, and I. A. Bustoni, "SARIMA (Seasonal ARIMA) implementation on time series to forecast the number of Malaria incidence," *Proc. - 2013 Int. Conf. Inf. Technol. Electr. Eng. "Intelligent Green Technol. Sustain. Dev. ICITEE 2013*, no. October, pp. 203–207, 2013, doi: 10.1109/ICITEED.2013.6676239.

[33] P. Marius-Constantin, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Trans. Circuits Syst.*, vol. 8, no. 7, pp. 579–588, 2009.

[34] A. M. Journal, "1 8 9 0," vol. 1, pp. 9–11, 2013.

[35] E. Žunić, K. Korjenić, K. Hodžić, and D. Đonko, "Application of Facebook'S Prophet Algorithm for Successful Sales Forecasting Based On Real-World Data," *arXiv*, vol. 12, no. 2, pp. 23–36, 2020, doi: 10.5121/ijcsit.2020.12203.

[36] M. Azarafza, M. Azarafza, and J. Tanha, "COVID-19 Infection Forecasting based on Deep Learning in Iran," pp. 1–7, 2020, doi: 10.1101/2020.05.16.20104182.

[37] P. Arora, H. Kumar, and B. K. Panigrahi, “Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India,” *Chaos, Solitons and Fractals*, vol. 139, p. 110017, 2020, doi: 10.1016/j.chaos.2020.110017.

[38] N. Kumar and S. Susan, “COVID-19 Pandemic Prediction using Time Series Forecasting Models,” pp. 1–7, 2020, doi: 10.1109/icccnt49239.2020.9225319.

[39] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun, “Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study,” *Chaos, Solitons and Fractals*, vol. 140, 2020, doi: 10.1016/j.chaos.2020.110121.

[40] Q. Yang, J. Wang, H. Ma, and X. Wang, “Research on COVID-19 based on ARIMA modelΔ—Taking Hubei, China as an example to see the epidemic in Italy,” *J. Infect. Public Health*, vol. 13, no. 10, pp. 1415–1418, 2020, doi: 10.1016/j.jiph.2020.06.019.

[41] R. O. Ogundokun and J. B. Awotunde, “Machine Learning Prediction for Covid 19 Pandemic in India,” vol. 2019, 2020, doi: 10.1101/2020.05.20.20107847.

[42] “Covid-19! |.” [Online]. Available: https://www.covid19.et/covid-19/. [Accessed: 28-Aug-2020].

[43] “Mean Square Error: Definition, Applications and Examples.” [Online]. Available: https://www.mygreatlearning.com/blog/mean-square-error-explained/. [Accessed: 27-Nov-2020].

[44] “What is Mean Squared Error, Mean Absolute Error, Root Mean Squared Error and R Squared? - Studytonight.” [Online]. Available: https://www.studytonight.com/post/what-is-mean-squared-error-mean-absolute-error-root-mean-squared-error-and-r-squared#. [Accessed: 27-Nov-2020].

[45] “Root-Mean-Square Error (RMSE) | Machine Learning.” [Online]. Available: https://www.includehelp.com/ml-ai/root-mean-square error-rmse.aspx. [Accessed: 28-Nov-2020].

[46] "Coefficient of Determination (R Squared): Definition, Calculation - Statistics How To." [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/coefficient-of-determination-r-squared/. [Accessed: 28-Nov-2020].

[47] "Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?" [Online]. Available: https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit. [Accessed: 28-Nov-2020].

[48] M. Haghani, M. C. J. Bliemer, F. Goerlandt, and J. Li, "Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID- 19 . The COVID-19 resource centre is hosted on Elsevier Connect , the company ' s public news and information ," no. January, 2020.

[49] L. Pérez-Campos Mayoral, M. T. Hernández-Huerta, G. Mayoral-Andrade, E. Pérez-Campos Mayoral, and E. Pérez-Campos, "A letter to the editor on 'World Health Organization declares global emergency: A review of the 2019 novel Coronavirus (COVID-19),'" *Int. J. Surg.*, vol. 79, no. January, pp. 163–164, 2020, doi: 10.1016/j.ijsu.2020.05.066.

Appendixes

Appendix A: A Sample Source Code

Appendix A.1 A Sample Code for ARIMA Grid Search

```
import warnings

from sklearn.metrics import mean_squared_error

# evaluate an ARIMA model for a given order (p,d,q)

def evaluate_arima_model(X, arima_order):

    # prepare training dataset

    train_size = int(len(X) * 0.80)

    train, test = X[0:train_size], X[train_size:]

    history = [x for x in train]

    # make predictions

    predictions = list()

    for t in range(len(test)):

        model = ARIMA(history, order=arima_order)

        model_fit = model.fit(disp=0)

        yhat = model_fit.forecast()[0]

        predictions.append(yhat)

        history.append(test[t])

    # calculate out of sample error
```

```
        error = mean_squared_error(test, predictions)
        return error
# evaluate combinations of p, d and q values for an ARIMA model
def evaluate_models(dataset, p_values, d_values, q_values):
        dataset = dataset.astype('float32')
        best_score, best_cfg = float("inf"), None
        for p in p_values:
                for d in d_values:
                        for q in q_values:
                                order = (p,d,q)
                                try:
                                        mse = evaluate_arima_model(dataset, order)
                                        if mse < best_score:
                                                best_score, best_cfg = mse, order
                                        print('ARIMA%s MSE=%.3f' % (order,mse))
                                except:
                                        continue
        print('Best ARIMA%s MSE=%.3f' % (best_cfg, best_score))
# load dataset
def parser(x):
```

```
    return datetime.strptime('190'+x, '%Y-%m')
#series = read_csv('shampoo-sales.csv', header=0, parse_dates=[0], index_col=0, squeeze=True, date_parser=parser)
# evaluate parameters
p_values = range(0, 5)
d_values = range(0, 3)
q_values = range(0, 3)
warnings.filterwarnings("ignore")
evaluate_models(newcase.values, p_values, d_values, q_values)
```

Appendix A.2 A Sample Code for FBProphet Grid Search

```
# Python
import itertools
import numpy as np
import pandas as pd
from fbprophet import Prophet
from fbprophet.diagnostics import cross_validation
from fbprophet.diagnostics import performance_metrics
import warnings
warnings.filterwarnings("ignore")
param_grid = { 'changepoint_prior_scale': [0.001, 0.01, 0.1, 0.5],
```

```
                'seasonality_prior_scale': [0.01, 0.1, 1.0, 10.0], }
all_params = [dict(zip(param_grid.keys(), v)) for v in itertools.product(*param_grid.values())]
rmses = []
df = pd.read_csv('C:/Users/ADONI/Desktop/Covid/My DS/New2.csv')
df.rename(columns={'Date':'ds','Daily_new_cases':'y'},inplace=True)
cutoffs = pd.to_datetime(['2020-03-20', '2020-07-20', '2020-11-20'])
for params in all_params:
    m = Prophet(**params).fit(df)
    df_cv = cross_validation(m, cutoffs=cutoffs, horizon='7 days', parallel="processes")
    df_p = performance_metrics(df_cv, rolling_window=1)
    rmses.append(df_p['rmse'].values[0])
tuning_results = pd.DataFrame(all_params)
tuning_results['rmse'] = rmses
print(tuning_results)
```