Disease outbreak prediction from time - series data using machine learning and deep learning approaches

Master of Computer Application

In the Faculty of Engineering & Technology

Jadavpur University

By

KASTURI GANGULY

Exam Roll No : MCA2340022

Registration No : **160140 of 2021 - 2022**

Under the Guidance of

Prof. Ram Sarkar

Department of Computer Science & Engineering

Department of Computer Science & Engineering

Jadavpur University

Kolkata - 700 032

2023

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

FACULTY OF ENGINEERING & TECHNOLOGY

JADAVPUR UNIVERSITY

CERTIFICATE

I hereby recommend that the project entitled "**A time series analysis-based Disease outbreak prediction using machine learning and deep learning models**" prepared under my supervision by **KASTURI GANGULY** , Exam Roll No : MCA2340022 , be accepted for the degree of **Master of Computer Application** of **Jadavpur University, Kolkata**.

Supervisor

_____

Prof. Ram Sarkar
Department of Computer Science & Engineering
Jadavpur University, Kolkata – 32

_____

Prof. Nandini Mukhopadhyay
Head of the Department of
Computer Science & Engineering
Jadavpur University, Kolkata – 32

_____

Prof. Ardhendu Ghoshal
Dean,
Faculty of Engineering & Technology,
Jadavpur University, Kolkata-32

JADAVPUR UNIVERSITY

FACULTY OF ENGINEERING & TECHNOLOGY

CERTIFICATE OF APPROVAL *

The foregoing project "**A time series analysis-based Disease outbreak prediction using machine learning and deep learning models**" at instance is hereby approved as a creditable study of an engineering subject carried out and presented in a manner of satisfactory to warrant its acceptance as pre-requisite to the degree for which it has been submitted. It is notified to be understood that by this approval, the undersigned do not necessarily endorse or approve any statement made, opinion expressed and conclusion drawn there in but approve the project only for the purpose for which it has been submitted.

Final Examination for the Evaluation of Project

Board of Examiners

_____

_____

(Signature of Examiners)

* Only in case project is approved

ACKNOWLEDGEMENTS

I express my profound gratitude and sincere thanks to Prof. Ram Sarkar for his valuable suggestions, guidance, constant encouragement and intent supervision at every stage of project work. It has been a great learning process for me. It is in his association that gave me many opportunities to ameliorate my skills and knowledge.

I am also thankful to Mr. Neelotpal Chakraborty, Assistant Professor, Institute of Engineering & Management (IEM) and Research Scholar, JU for providing his support and encouragement that I received to complete my project.

Last but not the least, I express my gratitude to my friends and family for their constant support and unfailing guidance in whatever I did.

_____

KASTURI GANGULY

M.C.A. (C.S.E)

Exam Roll No : MCA2340022

Date :

Place : Kolkata

Contents

List of figures

List of Tables

ABSTRACT

Time series analysis is an important way for understanding complex systems and making informed decisions. Tracking and predicting disease outbreaks all over the world is a complex problem, but time series analysis can be helpful for solving this issue. There are many ways to perform time series analysis like Descriptive analysis, Visual analysis, Frequency domain analysis, machine learning based analysis and many more.

In this project, machine learning regression models like Linear regression, Random Forest and deep learning model LSTM(Long short-term memory) is used to forecast time series based disease datasets. Three different datasets(COVID-19 in Kerala, Hungarian chickenpox and Adenovirus in Bay Area) are considered here. From this study, the most accurate model for Covid 19 in Kerala dataset is founded to be the Multiple linear regression model with the RMSE value 2910.811. The most accurate model for Hungarian chickenpox dataset is Random forest(RMSE value 24.799). The most accurate model for Adenovirus in Bay Area dataset is Multiple linear regression (RMSE value 3.1).

Chapter 1

1. Introduction

Time series forecasting has been a prosperous filed of science due to its popularity in real world applications. Machine learning is increasingly being used in time series forecasting to develop more accurate and robust models. In this project, forecasting is done by machine learning models on time series data of disease outbreaks.

1.1. About Time series forecasting

A time-series is a discrete sequence of a time-valued function ordered over time and most forecasting problems involve the use of time series data.

Forecasting of a time-series is useful for its wide-spread real-world applications. For example, the population growth in a country from the measure of its current population is useful to determine the future prospect of the citizens [1].

Examples of time series forecasting

- Forecasting the closing price of a stock each day.
- Forecasting product sales in units sold each day for a store.
- Forecasting unemployment for a state each quarter.
- Forecasting the average price of gasoline each day.

1.2. Machine Learning Approach to Time-Series forecasting

Time series analysis forecasting using machine learning was shown to be the most successful in identifying patterns in both structured and unstructured data.

There are many techniques used in time series forecasting that try to achieve better precision and reduce errors.

With machine learning, time series forecasting becomes faster, more precise, and more efficient in the long run.

Machine learning autonomously defines points of interest in the unlimited flow of data to then align them with customer data insights at hand and conduct *what-if* analysis. This

results in particularly efficient takes on stimulating the demand in the commercial sector, for instance.

Some applications of machine learning time series forecasting is Stock prices forecasting, Demand and sales forecasting, Web traffic forecasting, Climate and weather prediction, Demographic and economic forecasting, Scientific studies forecasting etc.

There are several time series forecasting machine learning methods like Artificial neural network, LSTM(Long short term memory based neural network), Random forest, KNN(K-nearest neighbors regression), SVM(Support vector regression) etc [2]

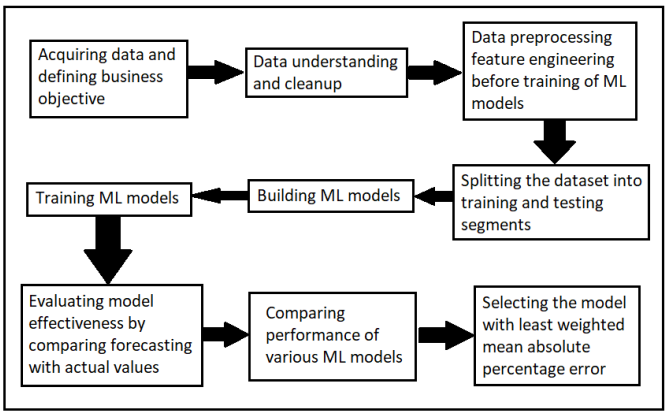

FIGURE 1 : THE MACHINE LEARNING FORECASTING PROCESS

1.3. About the diseases:

1.3.1. COVID - 19:

At saturday April 8, 2023 covid-19 virus affects 761,926,704 confirmed cases and has caused a total of 6,889,973 deaths worldwide. In India, total number of confirmed covid cases : 44,729,284 and total number of covid related deaths : 530,901 [3].

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).

COVID-19 was first reported in the Wuhan province of China in December of 2019 [4].

Although COVID-19 symptoms can vary, they frequently include fever, coughing, headaches, fatigue, breathing issues, loss of smell, and loss of taste [5] [6].

One to fourteen days after virus exposure, symptoms may appear. At least one-third of those who contract the disease experience no symptoms at all [7] [8].

When people breathe air contaminated by the virus's droplets and minute airborne particles, COVID-19 can spread.

With 44,729,284 reported cases of COVID-19 infection as of 9 April 2023, according to Indian government statistics, India has the second-highest number of confirmed cases in the world (after the United States of America) and the third-highest number of COVID-19 deaths (after the United States and Brazil), with 530,901 deaths [9]

On January 30, 2020, three towns in Kerala received reports of the country's first COVID-19 cases [10] [11]

1.3.2. Chickenpox:

The varicella zoster virus(VZV) is the primary cause of Chickenpox. The symptoms are a mild, self-limiting illness, characterised by low-grade fever, malaise, and a generalised, itchy, vesicular rash [12]. Chickenpox is extremely contagious [13]. The virus can be transmitted by close interaction or respiratory droplets [14]. In the US, one-dose varicella vaccination became available for kids in 1995. A second dose was advised in 2006 to further reduce varicella outbreaks and disease [15].

1.3.3. Adenovirus:

Adenovirus infection is a contagious viral disease. It is caused by Adenoviruses. The symptoms include fever, fatigue, muscle aches, headache, abdominal pain and swollen neck glands . The main way that infection spreads is through close contact between infected individuals [16]. Military personnel who may be more susceptible to infection from adenovirus types 4 and 7 can use a vaccine for those viruses [17].

1.4. Motivation:

Infectious diseases are a leading cause of death worldwide. Infectious disease outbreaks can be traumatic for everyone. So if you can predict the approximate onset of the disease, you can take measures in advance. The type and frequency of disease outbreaks Can be predicted. More importantly, it can predict whether the disease is likely to reoccur.

From the past experience of Covid 19, it can be said that proper forecasting is needed to prevent the fatalness of any disease. So Time series forecasting has played a crucial role in predicting and managing any pandemic. There have been a vast number of research studies conducted on COVID-19 since the start of the pandemic.

In this respect, medical time series data that describe the number of cases on each day or week are selected. Agenda is to prediction of future outbreak of those diseases.

In this project, some machine learning models are analysed to get the best machine learning model which gives higher accuracy and lower error value. The best machine learning model from the analysis can be used for forecasting of the diseases.

1.5. Scope of project:

Time series forecasting is an important tool for many industries like finance, healthcare, manufacturing, retail that need to make informed decisions based on historical data time series forecasting is also essential in weather forecasting as it allows meteorologists and weather forecasters to provide accurate and timely predictions of weather conditions, helping individuals and organizations make informed decisions and take necessary precautions to mitigate any potential risks. On the other hand, time series forecasting in stock price prediction is essential as it allows investors and traders to make informed decisions based on historical data trends and patterns. So time series forecasting becomes very relevant and important in the field of research work.

Time series forecasting in predicting and controlling disease outbreaks allows public health officials and healthcare professionals to identify potential outbreaks before they occur, plan interventions to prevent the spread of disease, and allocate resources more effectively.

The demand for time series forecasting in disease outbreak management has increased significantly in recent years, especially in light of the COVID-19 pandemic.

So for this project, time series data on disease outbreak is considered and machine learning models are trained by those data to manage disease outbreaks and identify potential outbreaks before they occur.

1.6. Organization of Project work

In the present work, time series prediction of disease outbreak is done and best working model among some machine learning model is proposed. The work can be divided into some parts. In this section, the purpose of each part is discussed.

Chapter 1 : Introduction

In the introduction part, some basics of time series forecasting and machine learning is discussed. The a brief description of the diseases are narrated. Main motivation and the scope of work are also discussed in this section.

Chapter 2 : Related work

This chapter describes some previous work regarding time series analysis, time series analysis of disease outbreak.

Chapter 3 : Working methodology

This section describes the theoretical concept of machine learning models which are used in the project like simple linear regression, Multiple linear Regression, Random forest and LSTM.

Chapter 4 : Experimental results

This section is about the experiments done for the project and the results of those experiments.

4.1. Data collection

This section describes how and from where the appropriate datasets are collected.

4.2. Data description

This section gives the proper description like length, features of whole datasets.

4.3. Data Preprocessing

Here, the essential features which are considered in experiment are selected from the main dataset.

4.4. Model training and testing

Total data is split into training and testing set and Training set is used to train the models and testing set is used to check the accuracy of the trained model.

4.5. Results analysis

In this section, results of the experiments are discussed. RMSE(Root Mean Squared Error) values and relevant graphs are shown for different datasets and different machine learning models.

4.6. Comparative discussion

Here, results from the experiments are compared and best machine learning model which provides the lowest RMSE is described.

Chapter 5 : Conclusion and future scope

Finally, the project work is concluded in this chapter and the future scope of the work is specified

Chapter 2

Related Work

In this chapter, some recent methods proposed by different researches have been discussed.

Shahid and Muneeb analyzed data consisting of COVID-19 one from Brazil, Germany, Italy, Spain, UK, China, India Israel, Russia, and the USA. This research utilized LSTM, GRU, and BI-LSTM, ARIMA, SVR with polynomial and RBF kernels. The highest scores of MAE and RMSE are 0.007 and 0.0077 respectively through Bi-LSTM [18].

Gupta et al. analyzed the dataset through the SVM, Prophet Forecasting Model, and Linear Regression model, the paper makes reliable predictions of COVID-19 measuring. The SVM method is categorized into Active Rate, Cured Rate, and Death Rate, and the Active rate has shown 266.82 in MSE and 16.22463 in RMSE. Cured rate and Death rate have marked 139229.8 and 17.12 in MSE, and 273.1351 and 4.137632 in RSME as well [19].

Smita, Alakananda and Alok Ranjan analyzed the daily statistics of people affected by Covid 19 and took into a account to predict the next days trend in the active cases in Odisha as well as India. Regression model such as linear and Multiple Linear Regression techniques are applied to the dataset to visualize the trend of the affected cases. A comparison of Linear Regression and Multiple Linear Regression model is performed where the score of the model R^2 tends to be 0.99 and 1.0 which indicates a strong prediction model to forecast the next coming days active cases [20].

Cafér Mert Yesilkanat investigated the performance of the Random Forest in estimating the near future case numbers for 190 countries in the world. And it is mapped in comparison with actual confirmed cases results. : it has been found that R^2 values for testing sub-data of RF model estimates range between 0.843 and 0.995 (average R^2= 0.959), and RMSE values between 141.76 and 526.18 (mean RMSE = 259.38); and that R^2 values for estimating sub-data range between 0.690 and 0.968 (mean R^2 = 0.914), and RMSE values between 549.73 and 2500.79 (mean RMSE = 909.37). These results show that the random forest machine learning algorithm performs well in estimating the number of cases for the near future in case of an epidemic like Novel Coronavirus [21].

Wadie, Arzu and Dániel use time-series forecasting techniques to model and predict the future incidence of chickenpox. To achieve this, we implement and simulate multiple models like ARIMA, SARIMA, SARIMAX, N-BEATS, DeepAR, LSTM, GRU(Gated recurrent unit), TFT(**Temporal Fusion Transformer)** and data preprocessing techniques on a Hungary-collected dataset [22].

Benedek, Paul, Oliver, Tamas and Rik propose the Chickenpox Cases in Hungary dataset as a new dataset for comparing graph neural network architectures. The time series analysis and forecasting experiments demonstrate that the Chickenpox Cases in Hungary dataset is adequate for comparing the predictive performance and forecasting capabilities of novel recurrent graph neural network architectures [23].

Vinay and LeiZhang evaluated the key features to predict the trends and possible stopping time of the current COVID-19 outbreak in Canada and around the world using LSTM , a deep learning approach. the RMSE error is 34.83 with an accuracy of 93.4% for short term predictions in Canada. Meanwhile the RMSE error is about 45.70 with an accuracy of 92.67% for long term predictions [24].

Chapter 3

Working Methodologies

3.1. Machine learning models

Machine learning is a technique to train machines to handle data more effectively and improve their performance through experiences [25].

This section is devoted to briefly describe the basic principle of three machine learning models(Simple linear regression, Multiple linear regression, Random Forest) that will be used later for time-series forecasting on different disease datasets.

3.1.1. Simple Linear Regression [26]

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables. If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression. The linear regression model provides a sloped straight line representing the relationship between the variables.

The formula(1) for linear regression is,

$$y = \beta_0 + \beta_1 X + \epsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

while y denotes the dependent variable for the independent variable X, β_0 denotes the intercept, the predicted value of y when the X is 0, β_1 denotes the regression coefficient, ϵ denotes the error of the estimate.

3.1.2. Multiple linear regression [27]

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Multiple Linear Regression is an extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable.

$$y = \beta_0 + \beta_1 x_1 + \ldots\ldots + \beta_n x_n + \varepsilon$$

Where,

y = dependent variable

x_i = independent variable

β_i = parameter

ε = error

3.1.3. Random Forest [28] [29]

Decision trees which have high variance are combined together in parallel so that the resultant variance becomes low. Decision trees are trained on particular sample data and the output is depend on multiple decision trees.

This forest involves m number of regression trees and the Random Forest algorithm averages the responses of grown trees $\{T(x)\}_1^M$

$$\widehat{f}_{RF}^{M}(x) = \frac{1}{m}\sum_{m=1}^{m} T(x)$$

In a random forest, aggregation (or "bagging") refers to the process of creating multiple decision trees by randomly selecting subsets of the original training data, and then aggregating the results of these trees to make a final prediction.

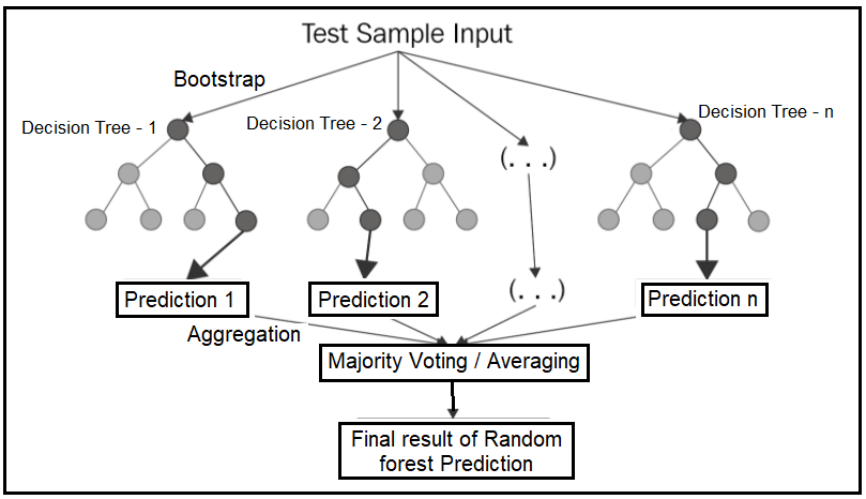

FIGURE 2 : BOOTSTRAP AND AGGREGATION IN RANDOM FOREST

3.2. Deep learning models

Deep learning is a subfield of machine learning that is based on artificial neural networks. These networks are composed of multiple layers of interconnected nodes (neurons) that process and transform input data to produce output [30].

This section is devoted to briefly describe the basic principle of one deep learning model(LSTM) that will be used later for time-series forecasting on different disease datasets.

3.2.1. LSTM (Long Short-Term Memory) [31] [32]

LSTM is a type of recurrent neural network (RNN) architecture that is widely used in deep learning for tasks such as natural language processing, speech recognition, and image captioning.

The main advantage of LSTM is its ability to handle long-term dependencies, which is a common problem in traditional RNNs. The architecture of LSTM includes a series of memory cells that can remember previous inputs and outputs over a long period of time. The memory cells are controlled by three gates: an input gate, an output gate, and a forget gate.

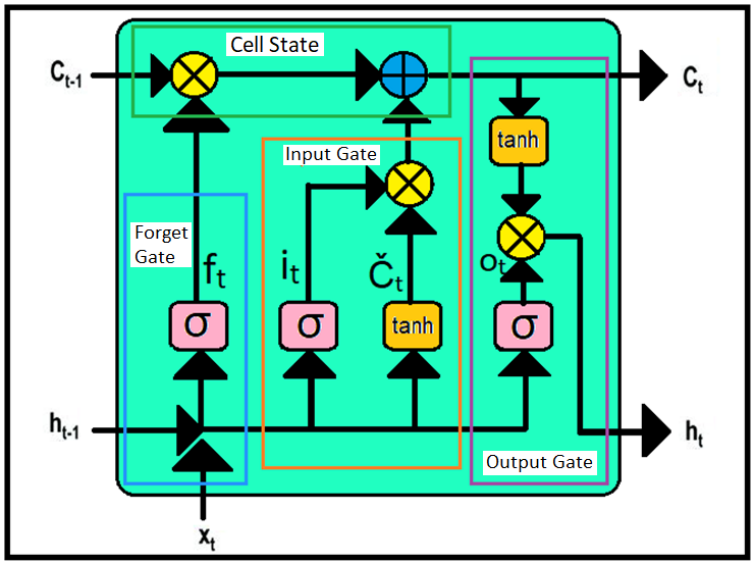

FIGURE 3 : ARCHITECTURE OF LSTM

The input gate determines which information should be stored in the memory cell, the forget gate determines which information should be discarded from the memory cell, and the output gate determines which information should be used to generate the output of the cell. These gates are controlled by a set of learnable parameters, which are updated during the training process using backpropagation.

The compact forms of LSTM can be illustrated mathematically as:

$$f_t = \sigma_g (\omega_f x_t + u_f h_{t-1} + b_f)$$

$$i_t = \sigma_g (\omega_i x_t + u_i h_{t-1} + b_i)$$

$$o_t = \sigma_g (\omega_o x_t + u_o h_{t-1} + b_o)$$

In which the bolder variables represent vectors, b is bias vector, matrices ω and u are the input and recurrent connection weights respectively, which will be adjusted from the training process. The subscript i, o, f and c indicate input gate, output gate, forget gate and the memory cell separately. t indicates the index of the time-step. x_t is the input vector, f_t is the activation vector of the forget gate, it is the activation vector of the input/update gate, o_t represents the activation vector of the output gate. h_t is the state vector in the hidden layer which is also the output vector of the LSTM unit. c_t is the cell state vector. σ_g is an activation function with a sigmoid function.

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c (\omega_c x_t + u_c h_{t-1} + b_c)$$

$$h_t = o_t \circ \sigma_h (c_t)$$

In which the operator o represents the Hadamard product, the initial values are $h_0 = 0$ and $c_0 = 0$. σ_c is an activation function with hyperbolic tangent function, while σ_h is an activation function with hyperbolic tangent function.

Chapter 4

Experimental Results

4.1. Dataset Location

The key aim of the project is to find out the best Machine learning or deep learning model which can be used for future prediction of a time series data. Three timeseries disease datasets are used that are mentioned in Table 1.

Dataset Name	URL	Location of the dataset
covid19-confirmed-cases-kerala	https://www.kaggle.com/datasets/anandhuh/covid19-confirmed-cases-kerala	kaggle.com
hungary chickenpox	https://archive.ics.uci.edu/ml/datasets/Hungarian+Chickenpox+Cases	uci machine learning repository
clinical-sentinel-laboratory-influenza-and-other-respiratory-virus-surveillance-data-by-region-and-influenza-season	https://data.world/chhs/fc544658-35c5-4be0-af20-fc703bc57c13	data.world

TABLE 1 : THE SOURCE OF THE DATASETS USED IN THE PROJECT FOR EXPERIMENTATION

4.2. Data description

4.2.1. covid19-confirmed-cases-Kerala

This dataset contains the confirmed COVID-19 cases in Kerala, India from **January 31, 2020** to **May 22, 2022**.

It contains dates and confirmed cases, which can be used for time series analysis.

4.2.2. Hungary chickenpox

A spatio-temporal dataset of weekly chickenpox (childhood disease) cases from Hungary. The dataset consists of a county-level adjacency matrix and time series of the county-level reported cases between 2005 and 2015.

Attributes are weekly counts of chickenpox cases in Hungarian counties.

4.2.3. clinical-sentinel-laboratory-influenza-and-other-respiratory-virus-surveillance-data-by-region-and-influenza-season

This dataset contains influenza and other respiratory virus surveillance data for different region in California from 10th October 2009 to 26th September 2020.

It contains season, date code, weekending, region, respiratory virus, number positive, specimens tested, percent positive.

4.3. Data preprocessing

Instead of processing the whole data, the main data is sorted or some features are selected so that it becomes easy to fit the data in the machine learning or deep learning models.

4.3.1. covid19-confirmed-cases-kerala

For covid19-confirmed-cases-kerala dataset, the whole data is taken for analysis.

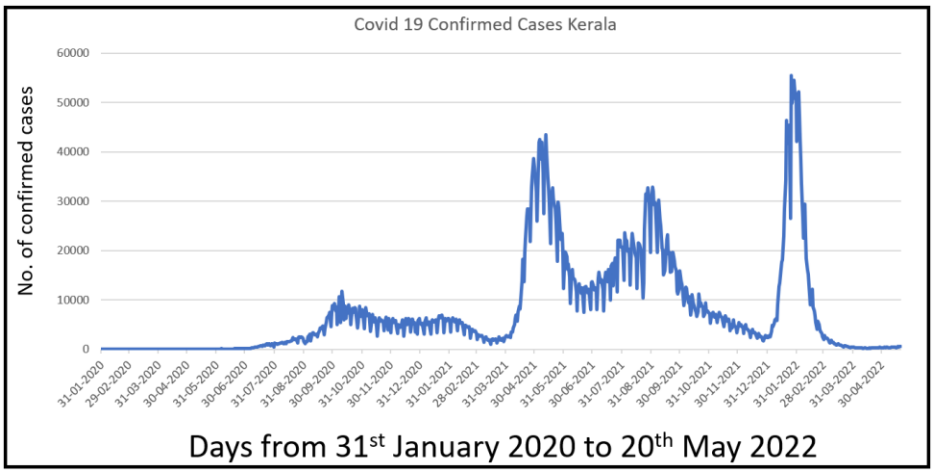

FIGURE 4 : PLOTTING THE WHOLE DATA OF COVID 19 IN KERALA

4.3.2. hungary chickenpox

As hungary chickenpox dataset contains data for 20 countries, Only one country's data which 'BARANYA' is considered here.

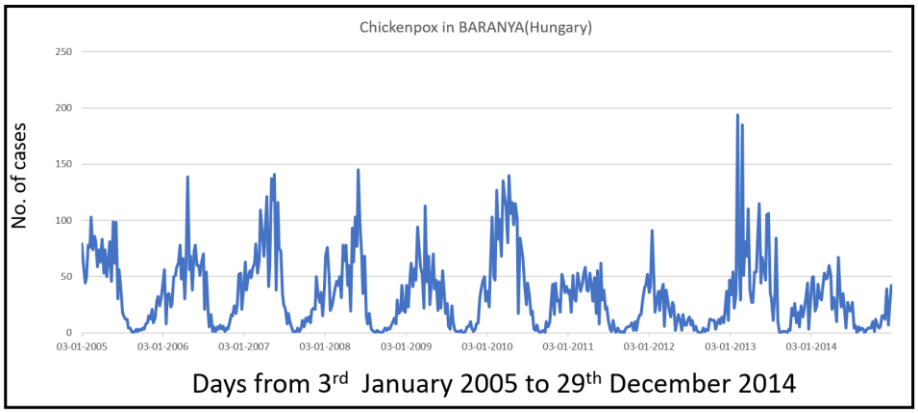

FIGURE 5 : PLOT THE WHOLE DATA OF CHICKENPOX IN BARANYA(HUNGARY)

4.3.3. clinical-sentinel-laboratory-influenza-and-other-respiratory-virus-surveillance-data-by-region-and-influenza-season

Among several respiratory virus, only Adenovirus is considered in Bay Area region only.

New dataset contains Date and number positive as columns.

Range of new dataset is 5th October 2013 to 26th September 2020.

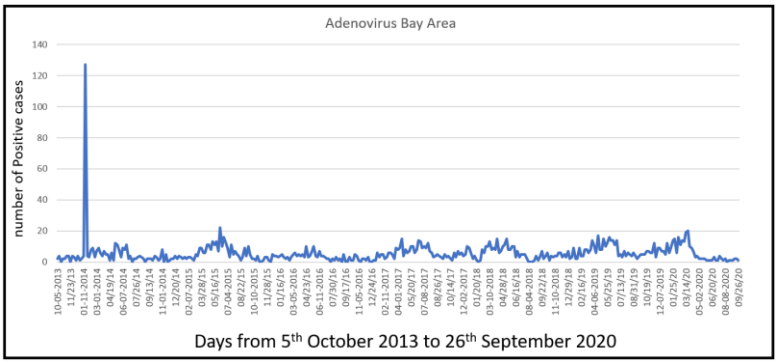

FIGURE 6 : PLOT THE WHOLE DATA OF ADENOVIRUS IN BAY AREA

4.4. Model training and testing

The learning phase refers to the stage of the process where a machine learning model is trained on a set of data to learn the features and attributes of the data. This set of data is

known as training set. Test set is a subset of data which is used to evaluate the performance of a trained model.

For every dataset, two types of splitting are done.

- 80% of the dataset has been used for training purpose and 20% for testing purpose.
- 70% of the dataset has been used for training purpose and 30% for testing purpose.

On this two types of splitting, four models (Simple linear regression, Multiple linear regression, Random forest, LSTM) are executed.

4.5. Results analysis

4.5.1. Prediction using Simple linear regression

4.5.1.1. Dataset – Adenovirus Bay area

Simple linear regression produces RMSE = 3.6123305769740273 where 80% of the dataset has been used for training purpose and 20% for testing purpose.

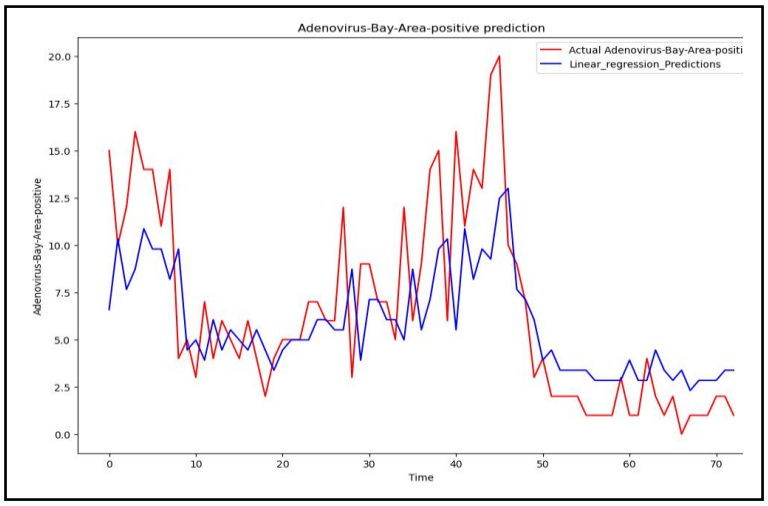

FIGURE 7 : POSITIVE CASES VS SIMPLE LINEAR REGRESSION PREDICTIONS FOR ADENOVIRUS IN BAY AREA(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Figure 7 represents the positive cases of Adenovirus in Bay Area from 11th May 2019 to 26th September 2020 and predicted cases by the Linear regression prediction algorithm where the x-axis shows the date of the occurrence of positive cases and y – axis shows the number of positive cases. Here, 80% of the dataset has been used for training purpose and 20% for testing purpose.

Simple linear regression produces RMSE = 3.5263457845865958 where 70% of the dataset has been used for training purpose and 30% for testing purpose.

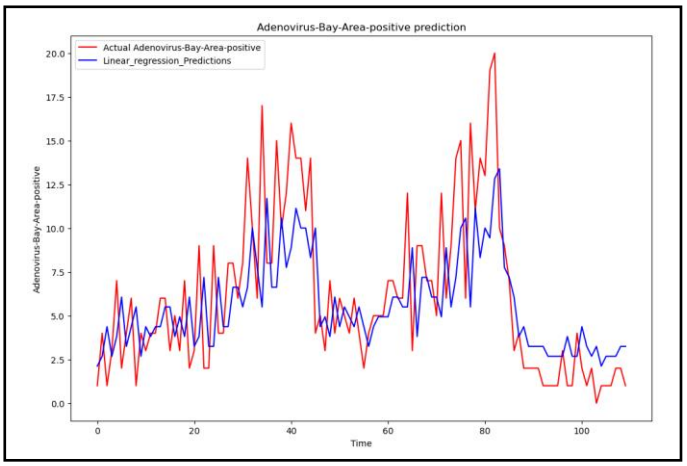

FIGURE 8 : POSITIVE CASES VS SIMPLE LINEAR REGRESSION PREDICTIONS FOR ADENOVIRUS IN BAY AREA(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Figure 8 represents the positive cases of Adenovirus in Bay Area from 25th August 2018 to 26th September 2020 and predicted cases by the Linear regression prediction algorithm where the x-axis shows the date of the occurrence of positive cases and y – axis shows the number of positive cases. Here, 70% of the dataset has been used for training purpose and 30% for testing purpose.

4.5.1.2. Dataset – Hungary chickenpox

Simple linear regression produces RMSE = 32.484645567112175 where 80% of the dataset has been used for training purpose and 20% for testing purpose.

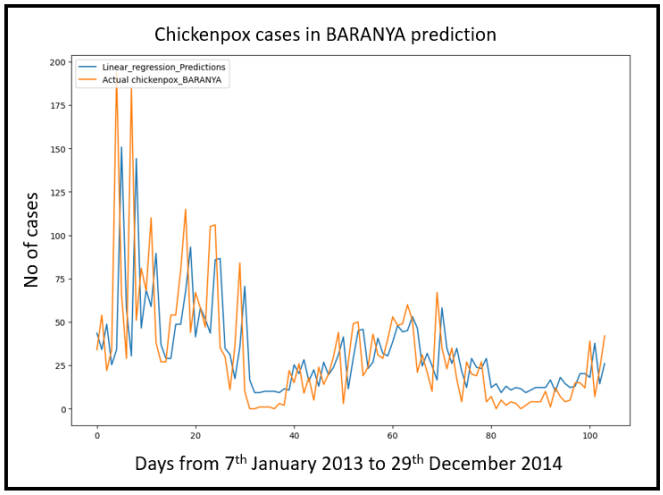

FIGURE 9 : ACTUAL CASES VS SIMPLE LINEAR REGRESSION PREDICTIONS FOR CHICKENPOX CASES IN BARANYA (TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Figure 9 represents the cases of Chickenpox in BARANYA from 7th January 2013 to 29th December 2014 and predicted cases by the simple Linear regression prediction algorithm where the x-axis shows the date of the occurrence of cases and y – axis shows the number of cases. Here, 80% of the dataset has been used for training purpose and 20% for testing purpose.

Simple linear regression produces RMSE = 27.826431981713657 where 70% of the dataset has been used for training purpose and 30% for testing purpose.

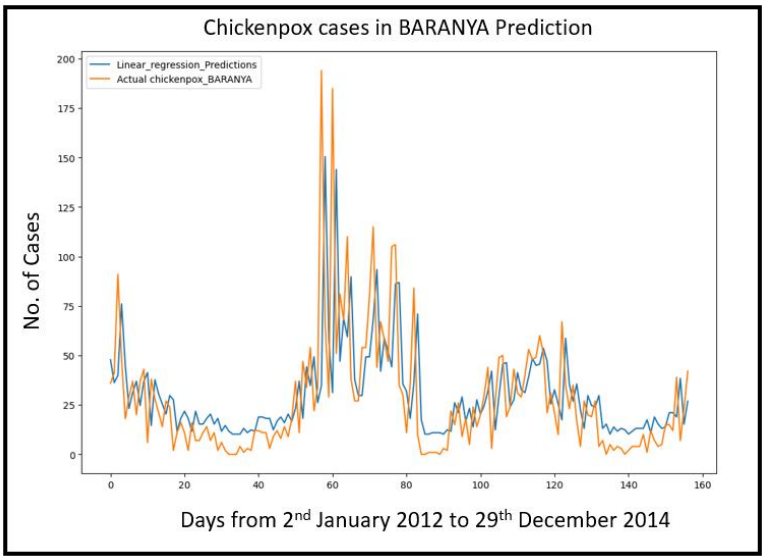

FIGURE 10 : ACTUAL CASES VS SIMPLE LINEAR REGRESSION PREDICTIONS FOR CHICKENPOX CASES IN BARANYA (TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Figure 10 represents the cases of Chickenpox in BARANYA from 2nd January 2012 to 29th December 2014 and predicted cases by the simple Linear regression prediction algorithm where the x-axis shows the date of the occurrence of cases and y – axis shows the number of cases. Here, 70% of the dataset has been used for training purpose and 30% for testing purpose.

4.5.1.3. Dataset – Covid kerala

Simple linear regression produces RMSE = 3532.3447049312704 where 80% of the dataset has been used for training purpose and 20% for testing purpose.

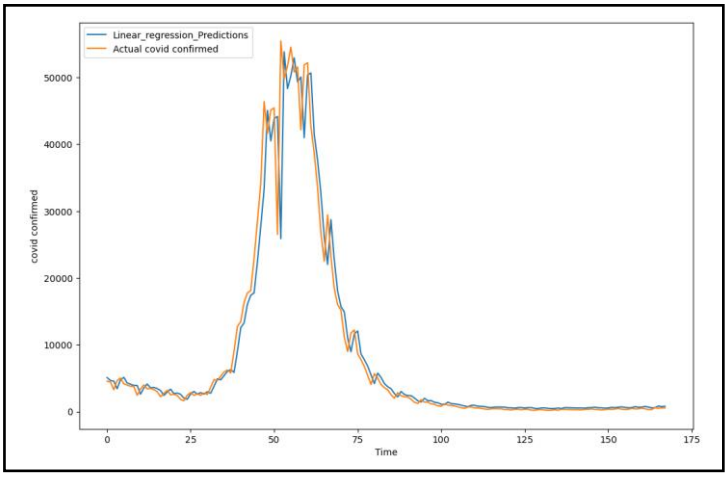

FIGURE 11 : CONFIRMED CASES VS SIMPLE LINEAR REGRESSION PREDICTIONS FOR COVID 19 CONFIRMED CASES IN KERALA (TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Figure 11 represents the confirmed cases of Covid 19 in Kerala from 4th December 2021 to 20th May 2022 and predicted cases by the simple Linear regression prediction algorithm where the x-axis shows the date of the occurrence of confirmed cases and y – axis shows the number of confirmed cases. Here, 80% of the dataset has been used for training purpose and 20% for testing purpose.

Simple linear regression produces RMSE = 3063.8380423113335 where 70% of the dataset has been used for training purpose and 30% for testing purpose.

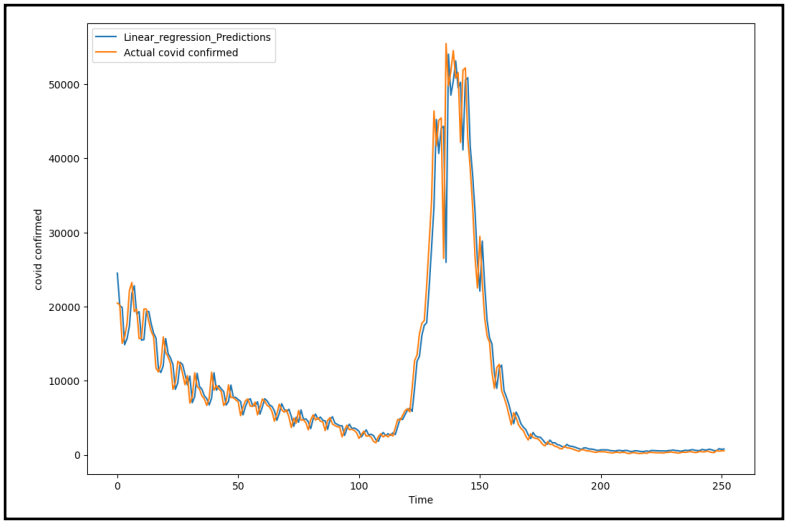

FIGURE 12 : CONFIRMED CASES VS SIMPLE LINEAR REGRESSION PREDICTIONS FOR COVID 19 CONFIRMED CASES IN KERALA (TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Figure 12 represents the confirmed cases of Covid 19 in Kerala from 11th September 2021 to 20th May 2022 and predicted cases by the simple Linear regression prediction algorithm where the x-axis shows the date of the occurrence of confirmed cases and y – axis shows the number of confirmed cases. Here, 70% of the dataset has been used for training purpose and 30% for testing purpose.

4.5.2. Prediction using Multiple linear regression

4.5.2.1. Dataset – Adenovirus Bay area

Multiple linear regression(using previous 3 weeks) produces RMSE = 3.106168140083212 where 80% of the dataset has been used for training purpose and 20% for testing purpose.

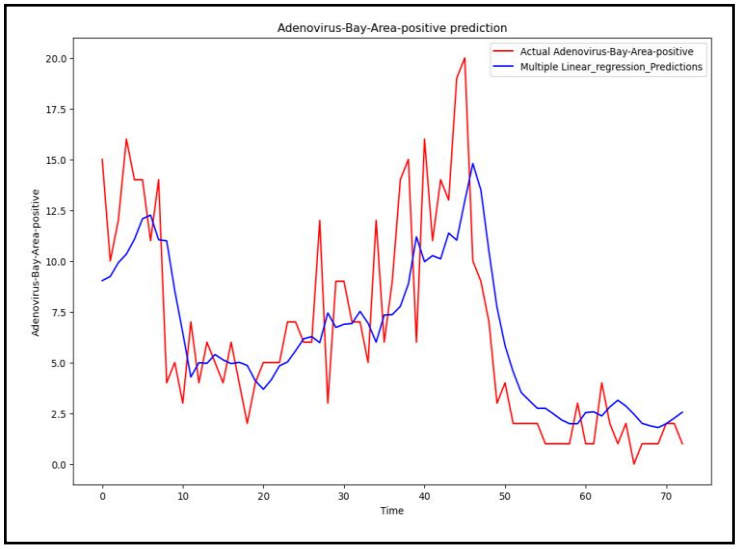

FIGURE 13 : POSITIVE CASES VS MULTIPLE LINEAR REGRESSION(USING PREVIOUS 3 WEEKS) PREDICTIONS FOR ADENOVIRUS IN BAY AREA(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Figure 13 represents the positive cases of Adenovirus in Bay Area from 11th May 2019 to 26th September 2020 and predicted cases by the Multiple Linear regression(using previous 3 weeks) prediction algorithm where the x-axis shows the date of the occurrence of positive cases and y – axis shows the number of positive cases. Here, 80% of the dataset has been used for training purpose and 20% for testing purpose.

Multiple Linear regression(using previous 3 weeks) produces RMSE = 3.107213906665679 where 70% of the dataset has been used for training purpose and 30% for testing purpose

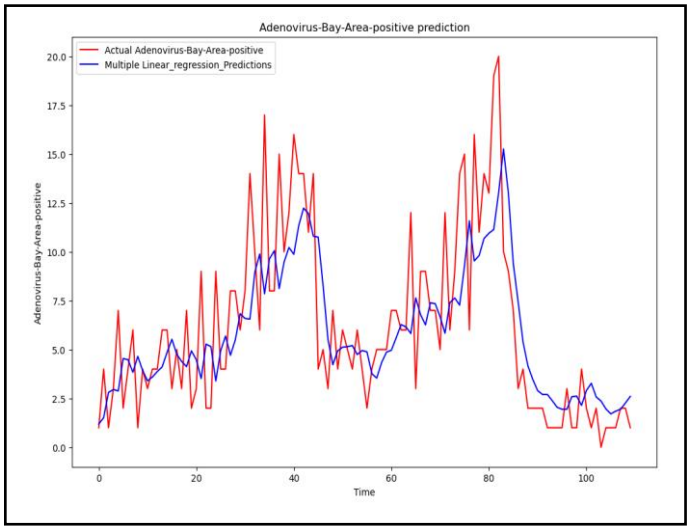

FIGURE 14 : POSITIVE CASES VS MULTIPLE LINEAR REGRESSION(USING PREVIOUS 3 WEEKS) PREDICTIONS FOR ADENOVIRUS IN BAY AREA(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Figure 14 represents the positive cases of Adenovirus in Bay Area from 25th August 2018 to 26th September 2020 and predicted cases by the Multiple Linear regression(using previous 3 weeks) prediction algorithm where the x-axis shows the date of the occurrence of positive cases and y – axis shows the number of positive cases. Here, 70% of the dataset has been used for training purpose and 30% for testing purpose.

4.5.2.2. Dataset – Hungary chickenpox

For dataset Hungary chickenpox, 5 and 10 previous data are used as predictor variables. The RMSE values are compared and then the graph which gives better result (less RMSE value) is plotted.

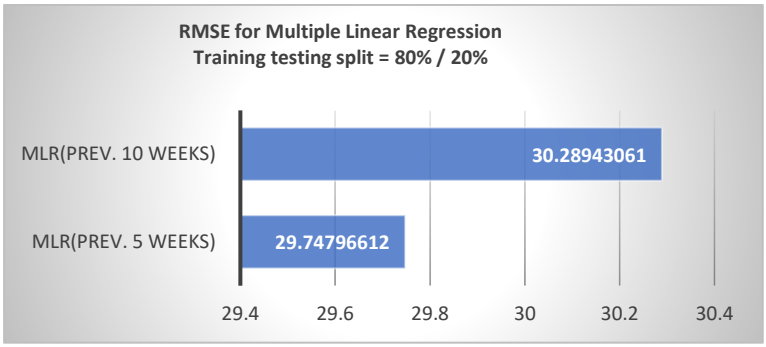

FIGURE 15 : SHOWS THE COMPARISON OF RMSE OF MULTIPLE LINEAR REGRESSION(CONSIDERING PREVIOUS 5 DATA AND 10 DATA) ON HUNGARY CHICKENPOX DATASET(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

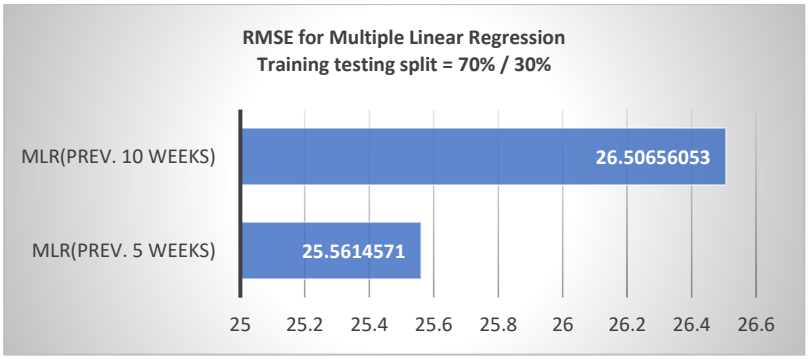

FIGURE 16 : SHOWS THE COMPARISON OF RMSE OF MULTIPLE LINEAR REGRESSION(CONSIDERING PREVIOUS 5 DATA AND 10 DATA) ON HUNGARY CHICKENPOX DATASET(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Multiple linear regression(using previous 5 days) produces RMSE = 29.747966116900173 where 80% of the dataset has been used for training purpose and 20% for testing purpose.

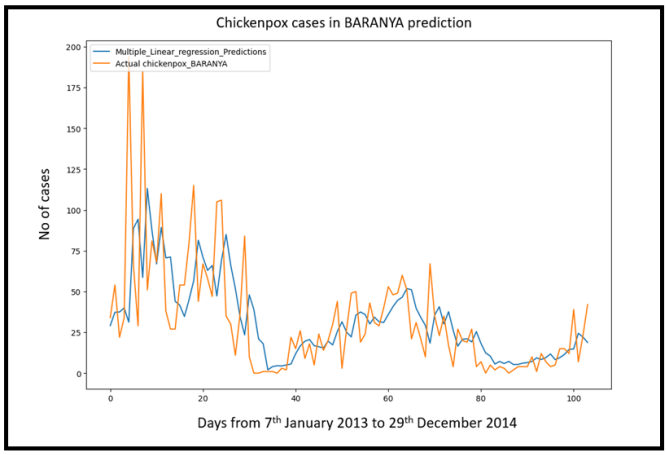

FIGURE 17 : POSITIVE CASES VS MULTIPLE LINEAR REGRESSION(USING PREVIOUS 5 DAYS) PREDICTIONS FOR CHICKENPOX IN BARANYA(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Figure 9 represents the cases of Chickenpox in BARANYA from 7th January 2013 to 29th December 2014 and predicted cases by the Multiple Linear regression(using previous 5 days) prediction algorithm where the x-axis shows the date of the occurrence of cases and y – axis shows the number of cases. Here, 80% of the dataset has been used for training purpose and 20% for testing purpose.

Multiple linear regression(using previous 5 days) produces RMSE = 25.561457096623982 where 70% of the dataset has been used for training purpose and 30% for testing purpose.

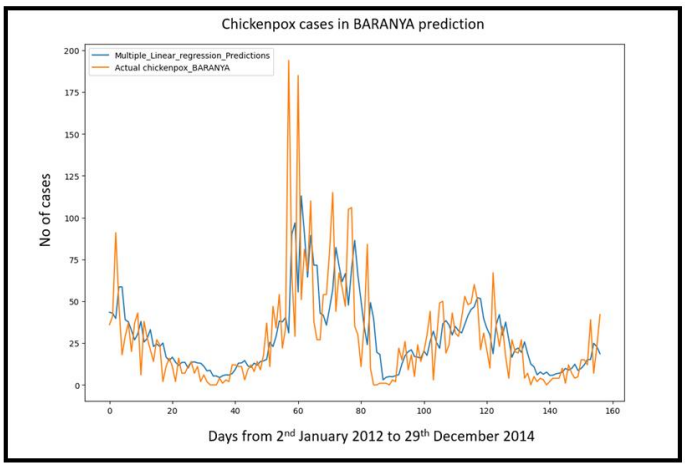

FIGURE 18 : POSITIVE CASES VS MULTIPLE LINEAR REGRESSION(USING PREVIOUS 5 DAYS) PREDICTIONS FOR CHICKENPOX IN BARANYA(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Figure 11 represents the cases of Chickenpox in BARANYA from 2nd January 2012 to 29th December 2014 and predicted cases by the Multiple Linear regression(using previous 5 days) prediction algorithm where the x-axis shows the date of the occurrence of cases and y – axis shows the number of cases. Here, 70% of the dataset has been used for training purpose and 30% for testing purpose.

4.5.2.3. Dataset – Covid kerala

For dataset Covid 19 cases in Kerala, 5 and 10 previous data are used as predictor variables. The RMSE values are compared.

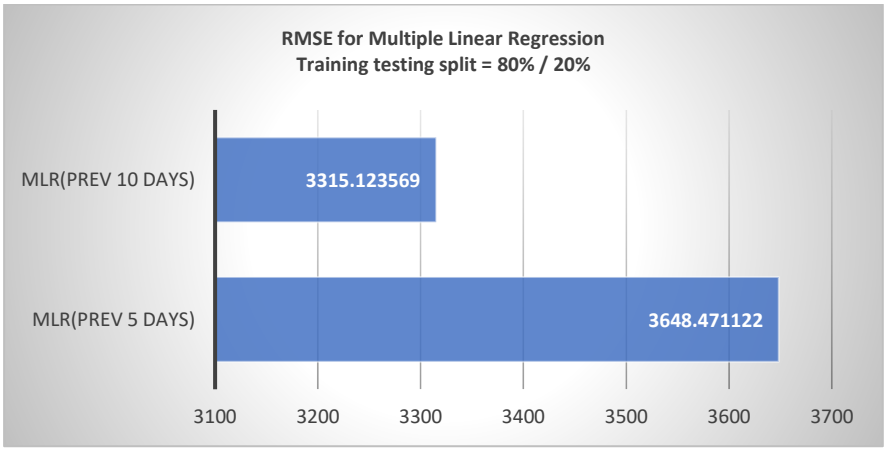

FIGURE 19 : SHOWS THE COMPARISON OF RMSE OF MULTIPLE LINEAR REGRESSION(CONSIDERING PREVIOUS 5 DATA AND 10 DATA) ON COVID 19 CONFIRMED CASES IN KERALA DATASET(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET

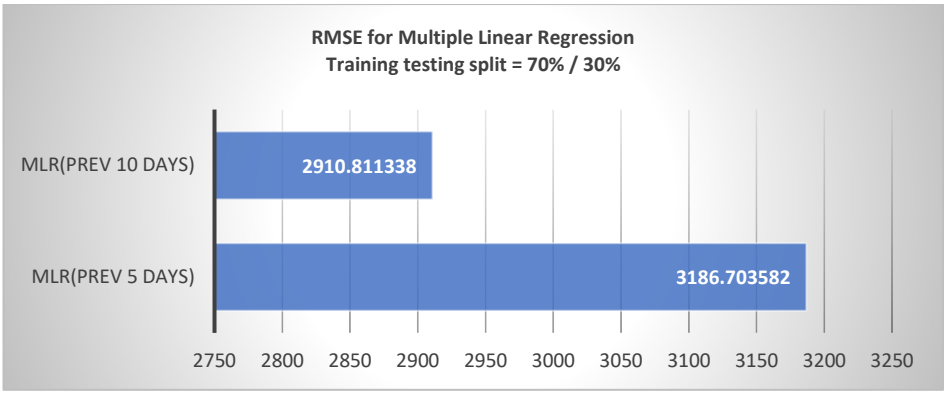

FIGURE 20 : SHOWS THE COMPARISON OF RMSE OF MULTIPLE LINEAR REGRESSION(CONSIDERING PREVIOUS 5 DATA AND 10 DATA) ON COVID 19 CONFIRMED CASES IN KERALA DATASET(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Multiple linear regression(using previous 10 days) produces RMSE = 3315.1235691891816 where 80% of the dataset has been used for training purpose and 20% for testing purpose.

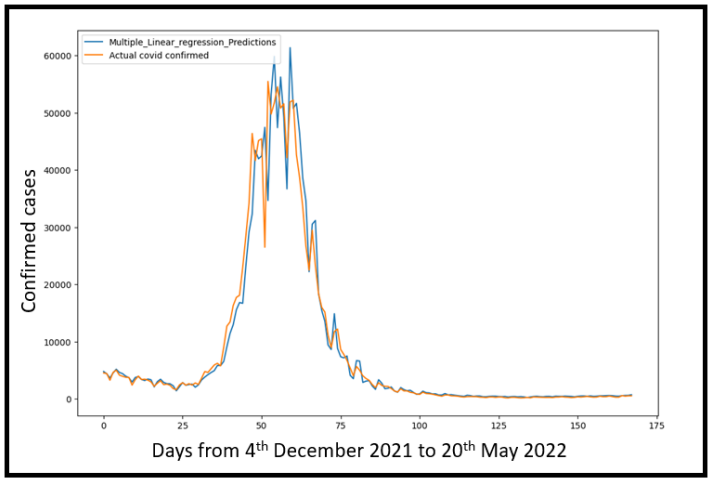

FIGURE 21 : CONFIRMED CASES VS MULTIPLE LINEAR REGRESSION(USING PREVIOUS 10 DAYS) PREDICTIONS FOR COVID 19 CONFIRMED CASES IN KERALA(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Figure 21 represents the cases of Covid 19 in Kerala from 4th December 2021 to 20th May 2022 and predicted confirmed cases by the Multiple Linear regression(using previous 10 days) prediction algorithm where the x-axis shows the date of the occurrence of confirmed cases and y – axis shows the number of confirmed cases. Here, 80% of the dataset has been used for training purpose and 20% for testing purpose.

Multiple linear regression(using previous 10 days) produces RMSE = 2910.8113378352673 where 70% of the dataset has been used for training purpose and 30% for testing purpose.

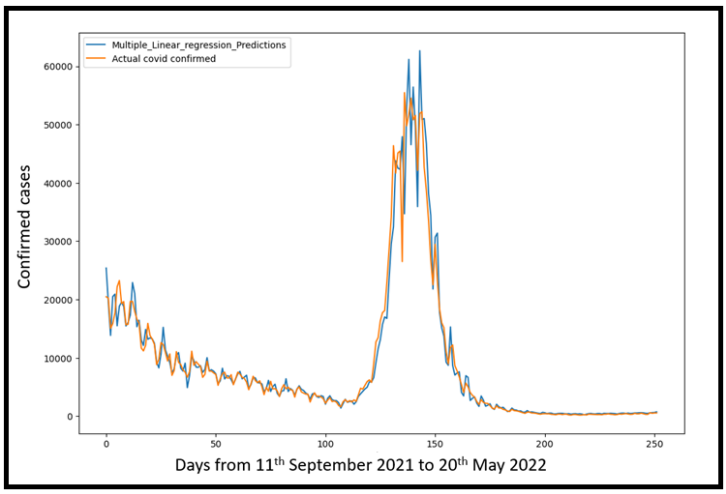

FIGURE 22 : CONFIRMED CASES VS MULTIPLE LINEAR REGRESSION(USING PREVIOUS 10 DAYS) PREDICTIONS FOR COVID 19 CONFIRMED CASES IN KERALA(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Figure 22 represents the cases of Covid 19 in Kerala from 11th September 2021 to 20th May 2022 and predicted confirmed cases by the Multiple Linear regression(using previous 10 days) prediction algorithm where the x-axis shows the date of the occurrence of confirmed cases and y – axis shows the number of confirmed cases. Here, 70% of the dataset has been used for training purpose and 30% for testing purpose.

4.5.3. Prediction using Random forest

4.5.3.1. Dataset – Adenovirus Bay area

Random forest regression(using previous 3 weeks) using parameters n_estimators=100, random_state=0 produces RMSE = 3.850452235342625 where 80% of the dataset has been used for training purpose and 20% for testing purpose.

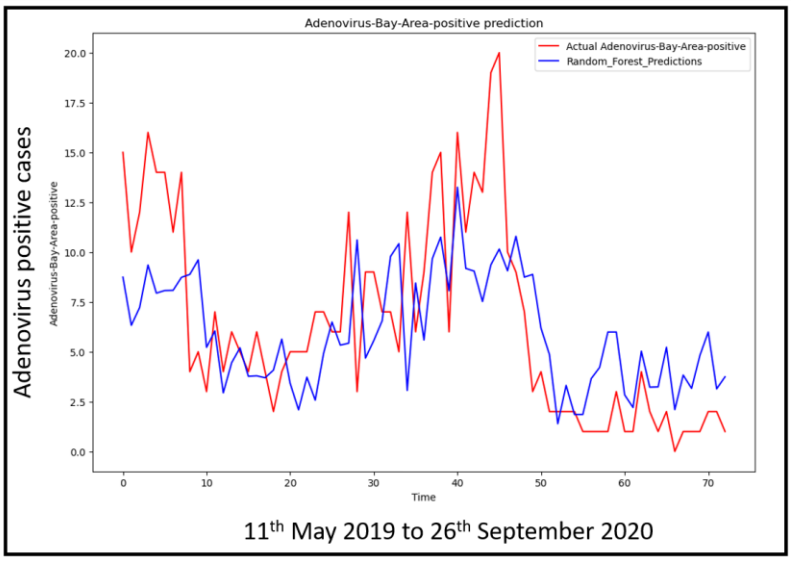

FIGURE 23 : POSITIVE CASES VS RANDOM FOREST REGRESSION(USING PREVIOUS 3 WEEKS)(USING PARAMETERS N_ESTIMATORS=100, RANDOM_STATE=0) PREDICTIONS FOR ADENOVIRUS IN BAY AREA(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Figure 23 represents the positive cases of Adenovirus in Bay Area from 11th May 2019 to 26th September 2020 and predicted cases by the Random Forest regression(using previous 3 weeks) prediction algorithm where the x-axis shows the date of the occurrence of positive cases and y – axis shows the number of positive cases. Here, 80% of the dataset has been used for training purpose and 20% for testing purpose.

Random Forest regression(using previous 3 weeks) using parameters n_estimators=100, random_state=0 produces RMSE = 3.590153632677852 where 70% of the dataset has been used for training purpose and 30% for testing purpose.

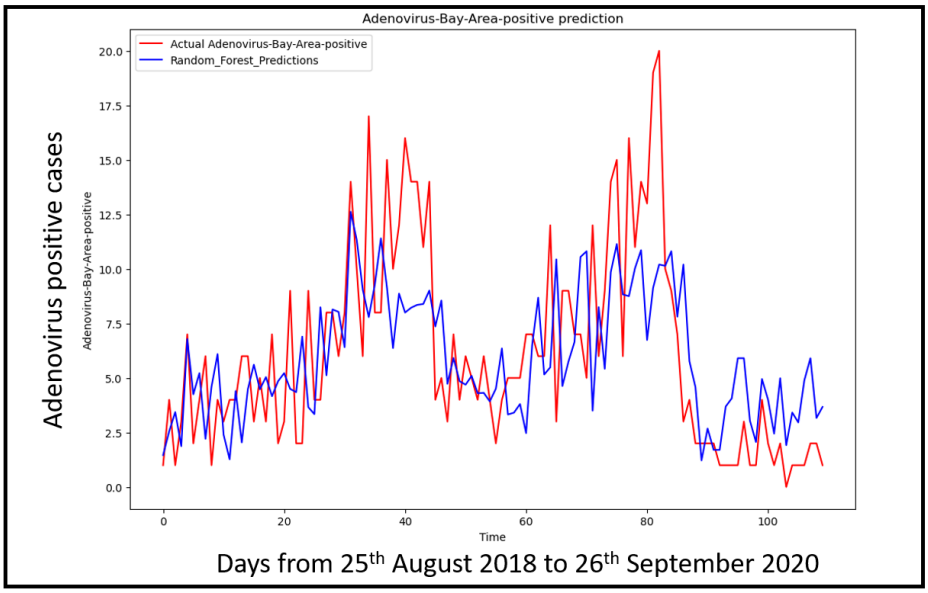

FIGURE 24 : POSITIVE CASES VS RANDOM FOREST REGRESSION(USING PREVIOUS 3 WEEKS) PREDICTIONS FOR ADENOVIRUS IN BAY AREA(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Figure 24 represents the positive cases of Adenovirus in Bay Area from 25th August 2018 to 26th September 2020 and predicted cases by the Random forest regression(using previous 3 weeks) prediction algorithm where the x-axis shows the date of the occurrence of positive cases and y – axis shows the number of positive cases. Here, 70% of the dataset has been used for training purpose and 30% for testing purpose.

4.5.3.2. Dataset – Hungary chickenpox

Random forest regression(using previous 10 days) using parameters n_estimators=100, random_state=0 produces RMSE = 29.19783272054491 where 80% of the dataset has been used for training purpose and 20% for testing purpose.

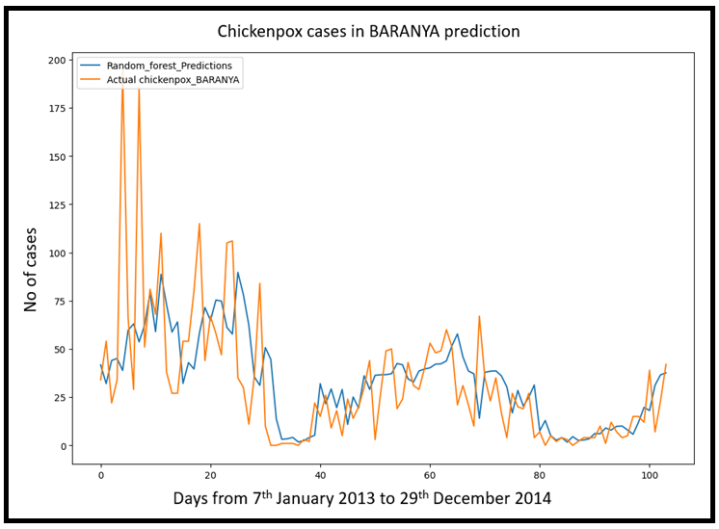

FIGURE 25 : POSITIVE CASES VS RANDOM FOREST REGRESSION(USING PREVIOUS 10 DAYS) PREDICTIONS FOR CHICKENPOX IN BARANYA(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Figure 25 represents the cases of Chickenpox in BARANYA from 7th January 2013 to 29th December 2014 and predicted cases by the Random Forest regression(using previous 10 days) prediction algorithm where the x-axis shows the date of the occurrence of cases and y – axis shows the number of cases. Here, 80% of the dataset has been used for training purpose and 20% for testing purpose.

Random forest regression(using previous 10 days) using parameters n_estimators=100, random_state=0 produces RMSE = 24.79959850455771 where 70% of the dataset has been used for training purpose and 30% for testing purpose.

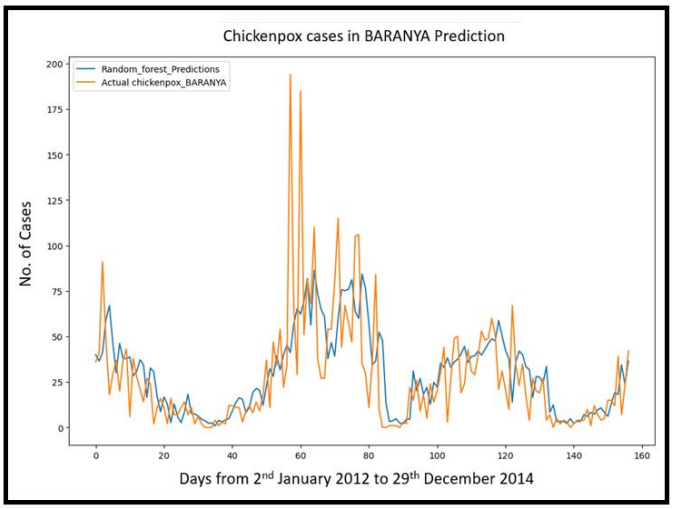

FIGURE 26 : POSITIVE CASES VS RANDOM FOREST REGRESSION(USING PREVIOUS 10 DAYS) PREDICTIONS FOR CHICKENPOX IN BARANYA(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Figure 26 represents the cases of Chickenpox in BARANYA from 2nd January 2012 to 29th December 2014 and predicted cases by the Random forest regression(using previous 10 days) prediction algorithm where the x-axis shows the date of the occurrence of cases and y – axis shows the number of cases. Here, 70% of the dataset has been used for training purpose and 30% for testing purpose.

4.5.3.3. Dataset – Covid kerala

Random forest regression(using previous 5 days) using parameters n_estimators=100, random_state=0 produces RMSE = 4202.757072430393 where 80% of the dataset has been used for training purpose and 20% for testing purpose.

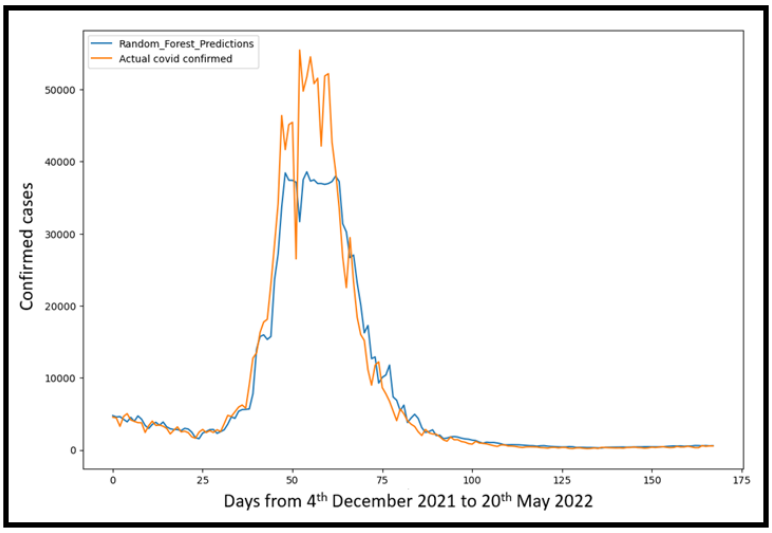

FIGURE 27 : CONFIRMED CASES VS RANDOM FOREST REGRESSION(USING PREVIOUS 5 DAYS) PREDICTIONS FOR COVID 19 CONFIRMED CASES IN KERALA(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Figure 27 represents the cases of Covid 19 in Kerala from 4th December 2021 to 20th May 2022 and predicted confirmed cases by the Random forest regression(using previous 5 days) prediction algorithm where the x-axis shows the date of the occurrence of confirmed cases and y – axis shows the number of confirmed cases. Here, 80% of the dataset has been used for training purpose and 20% for testing purpose.

Random forest regression(using previous 5 days) using parameters n_estimators=100, random_state=0 produces RMSE = 3601.5329058843618 where 70% of the dataset has been used for training purpose and 30% for testing purpose.

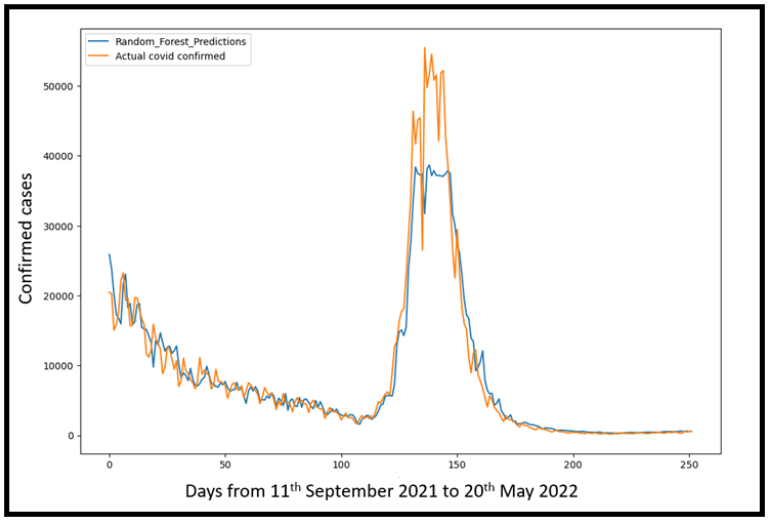

FIGURE 28 : CONFIRMED CASES VS RANDOM FOREST REGRESSION(USING PREVIOUS 5 DAYS) PREDICTIONS FOR COVID 19 CONFIRMED CASES IN KERALA(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Figure 28 represents the cases of Covid 19 in Kerala from 11th September 2021 to 20th May 2022 and predicted confirmed cases by the Random forest regression(using previous 5 days) prediction algorithm where the x-axis shows the date of the occurrence of confirmed cases and y – axis shows the number of confirmed cases. Here, 70% of the dataset has been used for training purpose and 30% for testing purpose.

4.5.4. Prediction using LSTM

4.5.4.1. Dataset – Adenovirus Bay area

For dataset Adenovirus Bay Area, LSTM is also used to see whether its works better than the machine learning regression models or not.

So for LSTM different timesteps and epochs are used to examine which parameter is working best.

For dataset Adenovirus Bay Area, 5 different epochs are used i.e. 20,25,30,50,100,150 and 2 different timesteps are used i.e. 60,90.

The RMSEs are compared and the best LSTM graph will be displayed.

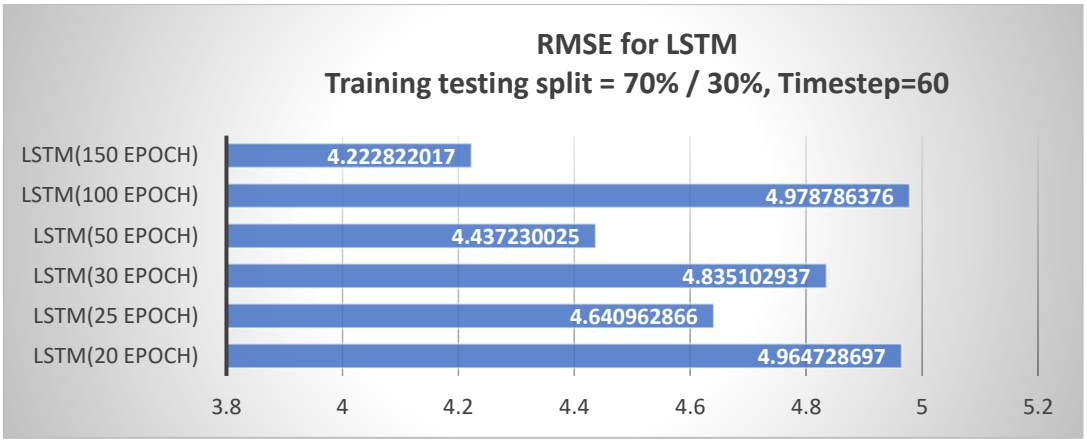

FIGURE 29 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=60,TRAINING SET = 70% AND TEST SET = 30% OF ADENOVIRUS IN BAY AREA DATASET)

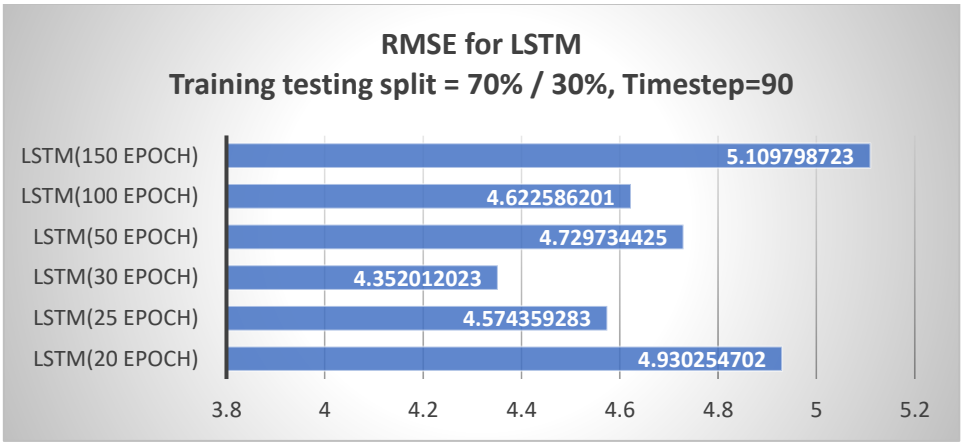

FIGURE 30 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=90,TRAINING SET = 70% AND TEST SET = 30% OF ADENOVIRUS IN BAY AREA DATASET)

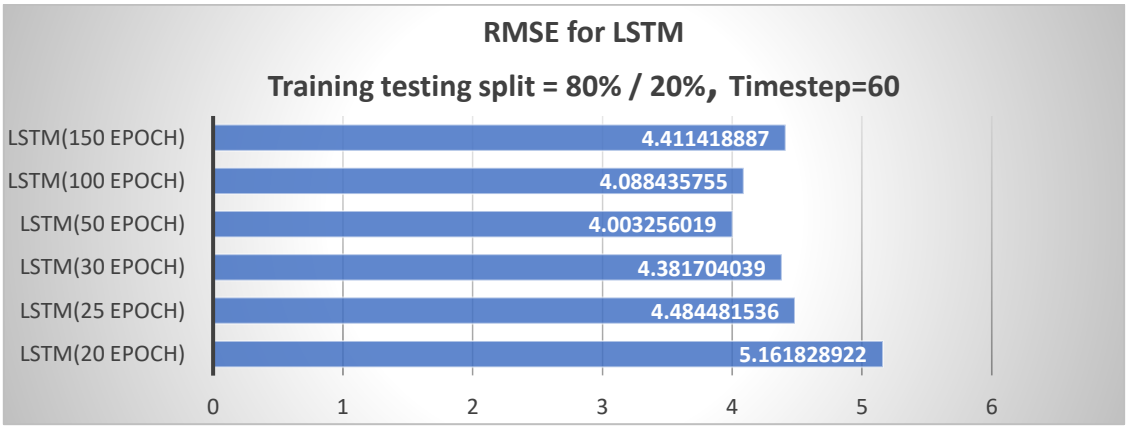

FIGURE 31 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=60,TRAINING SET = 80% AND TEST SET = 20% OF ADENOVIRUS IN BAY AREA DATASET)

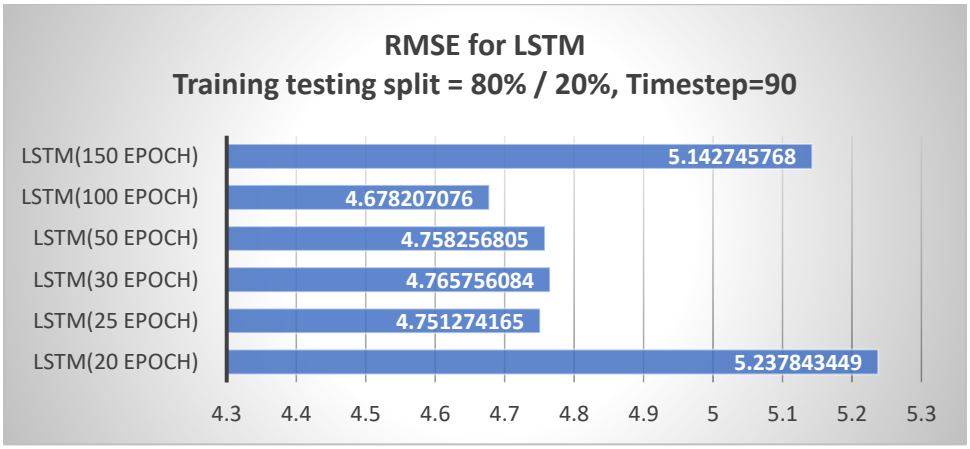

FIGURE 32 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=90,TRAINING SET = 80% AND TEST SET = 20% OF ADENOVIRUS IN BAY AREA DATASET)

From Figure 29,30,31,32, the lowest RMSE value = 4.003256019(epoch = 50, timestep=60,training set = 80% and test set = 20% of the dataset) and second lowest RMSE value = 4.222822017 (epoch = 150, timestep=60,training set = 70% and test set = 30% of the dataset)

LSTM(epoch = 50, timestep=60) produces RMSE = 4.003256019 where 80% of the dataset has been used for training purpose and 20% for testing purpose.

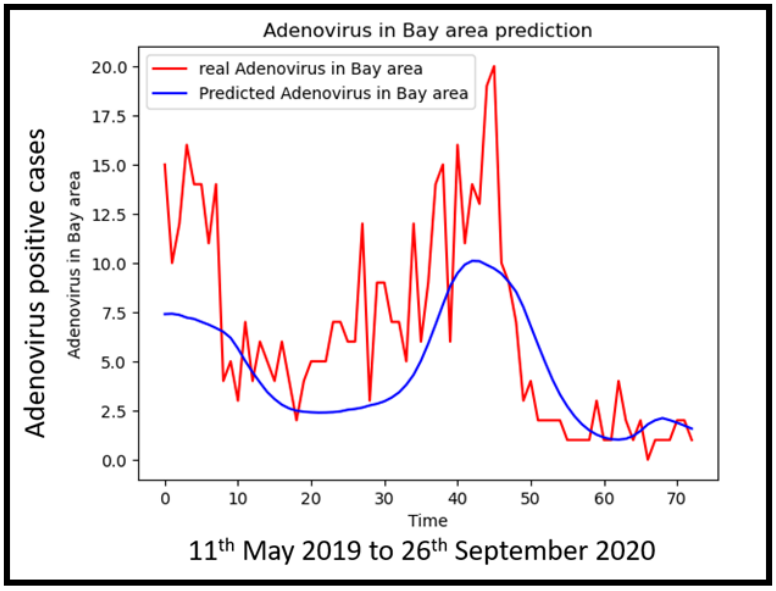

FIGURE 33 : POSITIVE CASES VS LSTM(EPOCH = 50, TIMESTEP=60) PREDICTIONS FOR ADENOVIRUS POSITIVE CASES IN BAY AREA(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Figure 33 represents the positive cases of Adenovirus in Bay Area from 11th May 2019 to 26th September 2020 and predicted positive cases by the LSTM(epoch = 50, timestep=60) prediction algorithm where the x-axis shows the date of the occurrence of positive cases and y – axis shows the number of positive cases. Here, 80% of the dataset has been used for training purpose and 20% for testing purpose.

LSTM(epoch = 150, timestep=60) produces RMSE = 4.222822017 where 70% of the dataset has been used for training purpose and 30% for testing purpose.

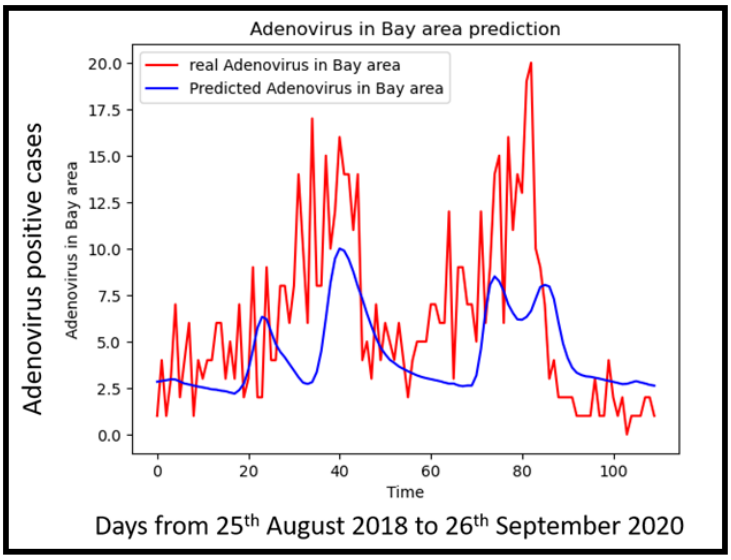

FIGURE 34 : POSITIVE CASES VS LSTM(EPOCH = 150, TIMESTEP=60) PREDICTIONS FOR ADENOVIRUS POSITIVE CASES IN BAY AREA(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Figure 34 represents the positive cases of Adenovirus in Bay Area from 25th August 2018 to 26th September 2020 and predicted positive cases by the LSTM(epoch = 150, timestep=60) prediction algorithm where the x-axis shows the date of the occurrence of positive cases and y – axis shows the number of positive cases. Here, 70% of the dataset has been used for training purpose and 30% for testing purpose.

4.5.4.2. Dataset – Hungary chickenpox

For dataset Hungary chickenpox, 4 different epochs are used i.e. 50, 100, 150, 200 and 3 different timesteps are used i.e. 30, 60, 90.

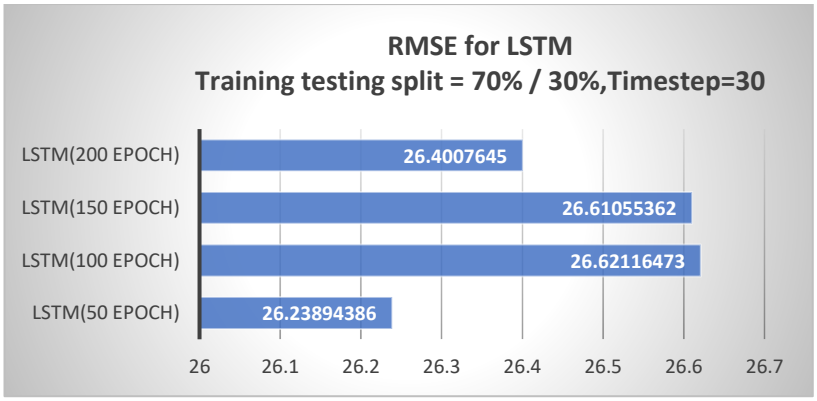

FIGURE 35 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=30,TRAINING SET = 70% AND TEST SET = 30% OF HUNGARY CHICKENPOX DATASET)

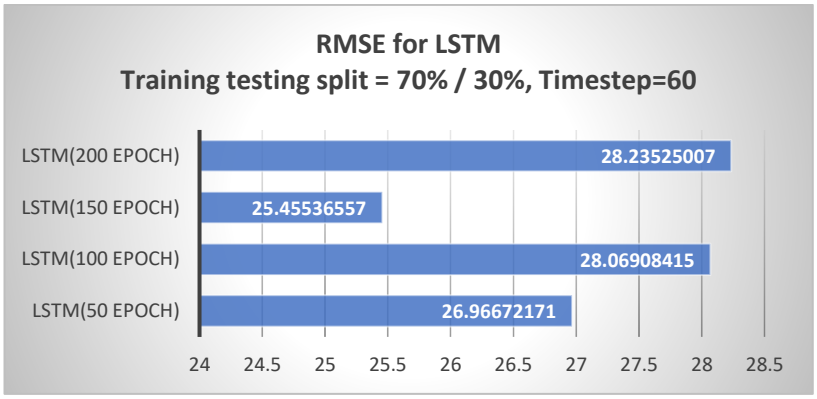

FIGURE 36 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=60,TRAINING SET = 70% AND TEST SET = 30% OF HUNGARY CHICKENPOX DATASET)

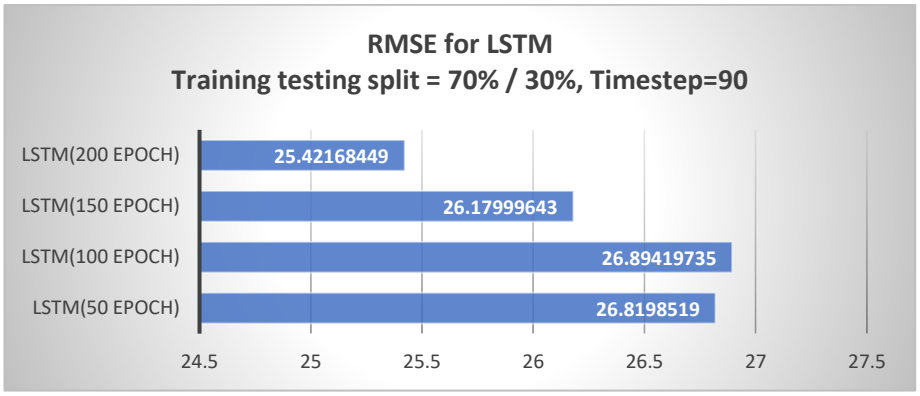

FIGURE 37 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=90,TRAINING SET = 70% AND TEST SET = 30% OF HUNGARY CHICKENPOX DATASET)

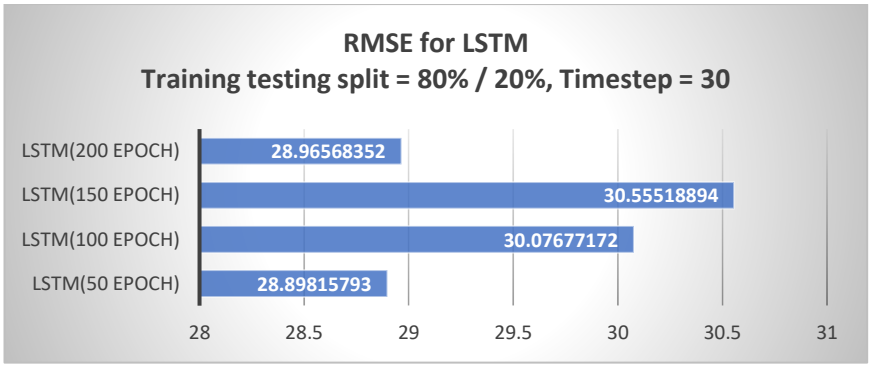

FIGURE 38 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=30,TRAINING SET = 80% AND TEST SET = 20% OF HUNGARY CHICKENPOX DATASET)

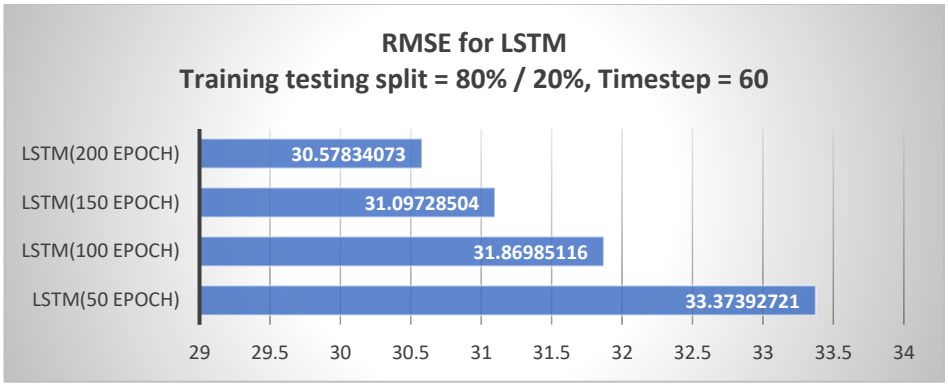

FIGURE 39 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=60,TRAINING SET = 80% AND TEST SET = 20% OF HUNGARY CHICKENPOX DATASET)

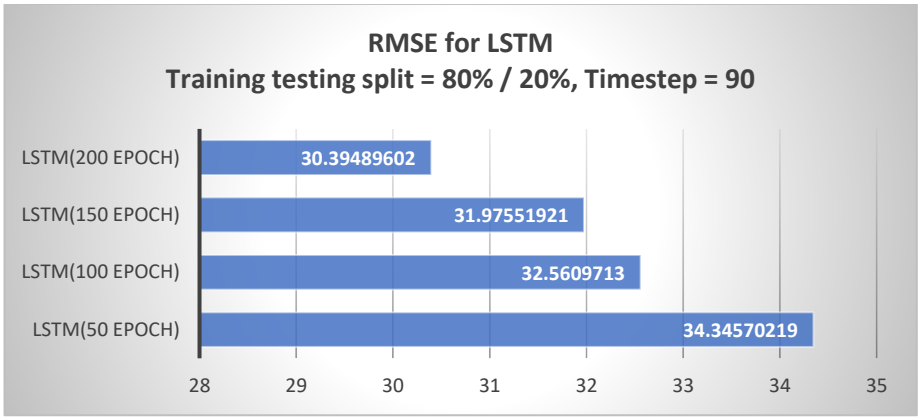

FIGURE 40 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=90,TRAINING SET = 80% AND TEST SET = 20% OF HUNGARY CHICKENPOX DATASET)

From figure 35 to figure 40, the lowest RMSE value when training set = 70% and test set = 30% of the dataset = 25.42168449(epoch = 200, timestep=90) and lowest RMSE value where training set = 80% and test set = 20% of the dataset = 28.89815793399269(epoch = 50, timestep=30)

LSTM(epoch = 200, timestep=90) produces RMSE = 25.42168449 where 70% of the dataset has been used for training purpose and 30% for testing purpose.

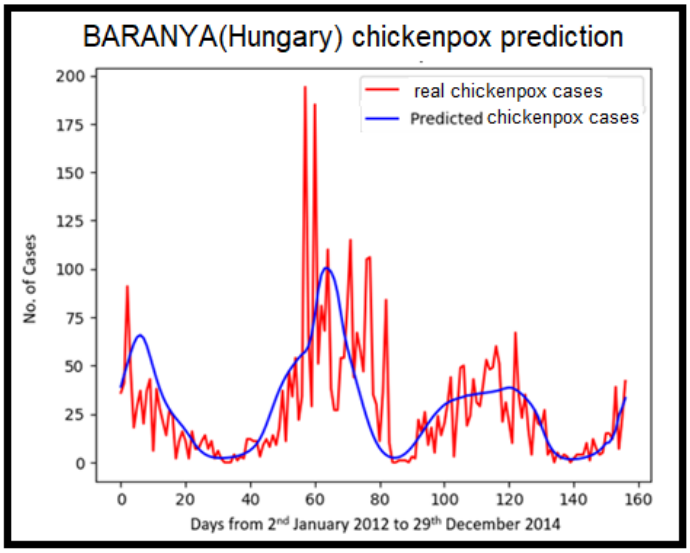

FIGURE 41 : ACTUAL CASES VS LSTM(EPOCH = 200, TIMESTEP=90) PREDICTIONS FOR HUNGARY CHICKENPOX DATASET(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Figure 41 represents the cases of Chickenpox in Hungary(BARANYA) from 2nd January 2012 to 29th December 2014 and predicted positive cases by the LSTM(epoch = 200, timestep=90) prediction algorithm where the x-axis shows the date of the occurrence of cases and y – axis shows the number of cases. Here, 70% of the dataset has been used for training purpose and 30% for testing purpose.

LSTM(epoch = 50, timestep=30) produces RMSE = 28.89815793399269 where 80% of the dataset has been used for training purpose and 20% for testing purpose.

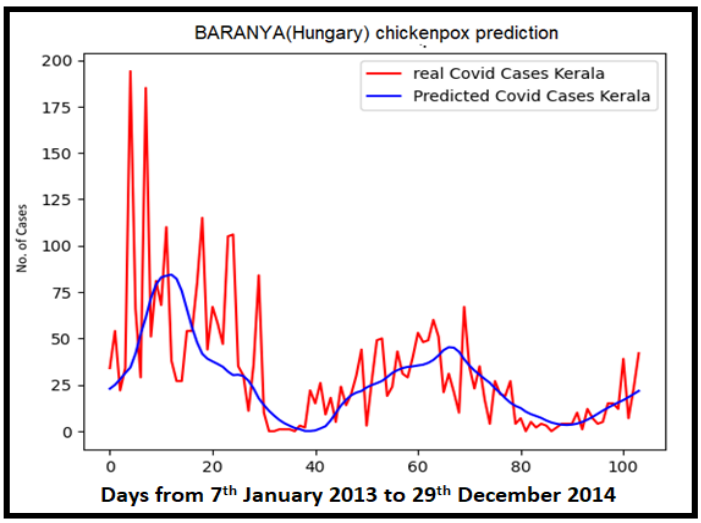

FIGURE 42 : ACTUAL CASES VS LSTM(EPOCH = 50, TIMESTEP=30) PREDICTIONS FOR HUNGARY CHICKENPOX DATASET(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Figure 42 represents the cases of Chickenpox in Hungary(BARANYA) from 7th January 2013 to 29th December 2014 and predicted positive cases by the LSTM(epoch = 50, timestep=30) prediction algorithm where the x-axis shows the date of the occurrence of cases and y – axis shows the number of cases. Here, 80% of the dataset has been used for training purpose and 20% for testing purpose.

4.5.4.3. Dataset – Covid kerala

For dataset Covid 19 confirmed cases Kerala, 5 different epochs are used i.e. 20, 50, 100, 150, 200 and 3 different timesteps are used i.e. 30, 60, 90.

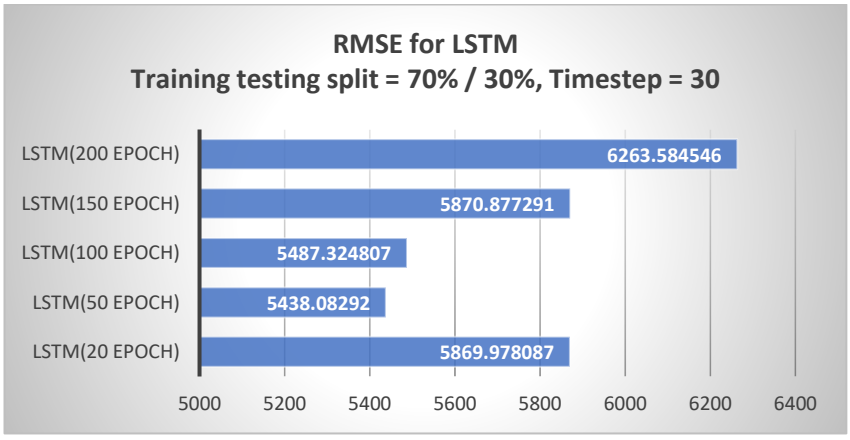

FIGURE 43 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=30,TRAINING SET = 70% AND TEST SET = 30% OF COVID 19 CONFIRMED CASES KERALA DATASET)

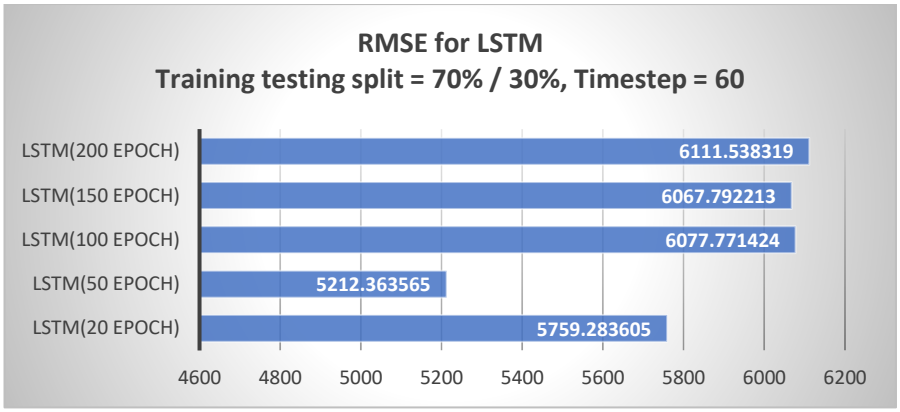

FIGURE 44 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=60,TRAINING SET = 70% AND TEST SET = 30% OF COVID 19 CONFIRMED CASES KERALA DATASET)

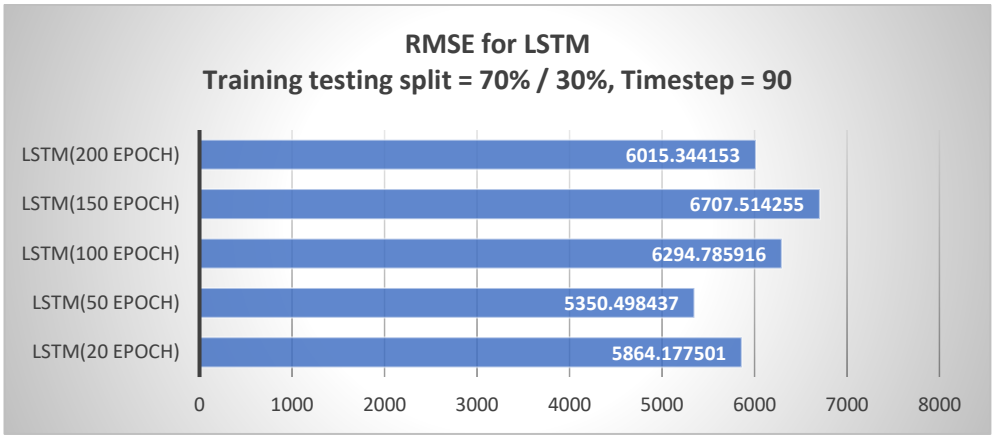

FIGURE 45 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=90,TRAINING SET = 70% AND TEST SET = 30% OF COVID 19 CONFIRMED CASES KERALA DATASET)

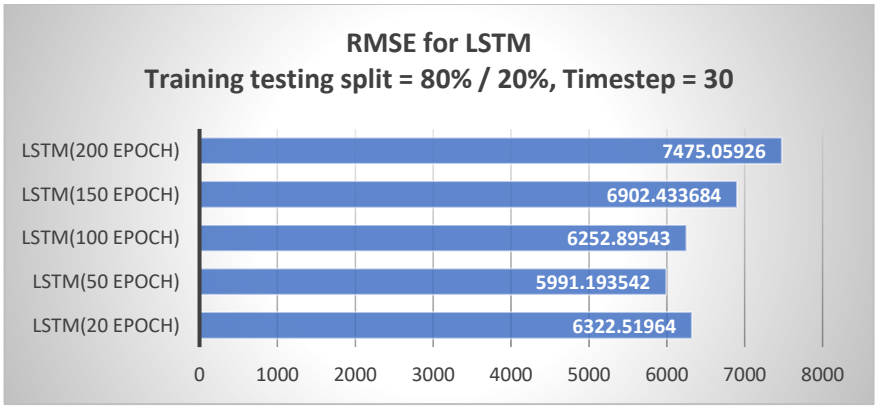

FIGURE 46 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=30,TRAINING SET = 80% AND TEST SET = 20% OF COVID 19 CONFIRMED CASES KERALA DATASET)

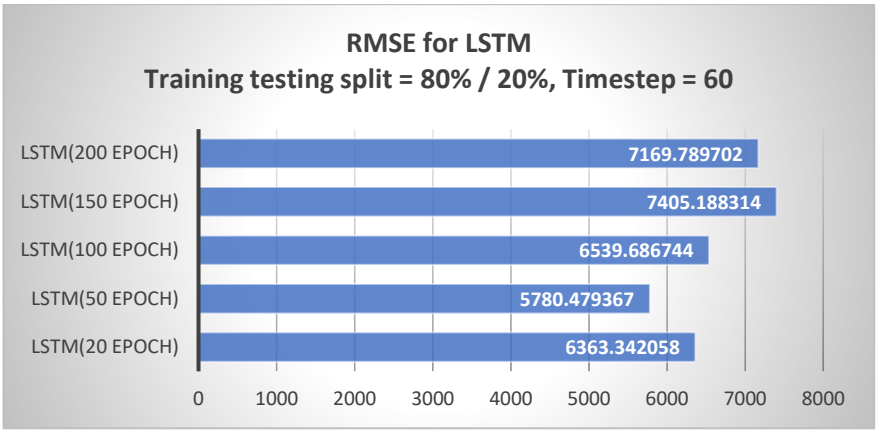

FIGURE 47 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=60,TRAINING SET = 80% AND TEST SET = 20% OF COVID 19 CONFIRMED CASES KERALA DATASET)

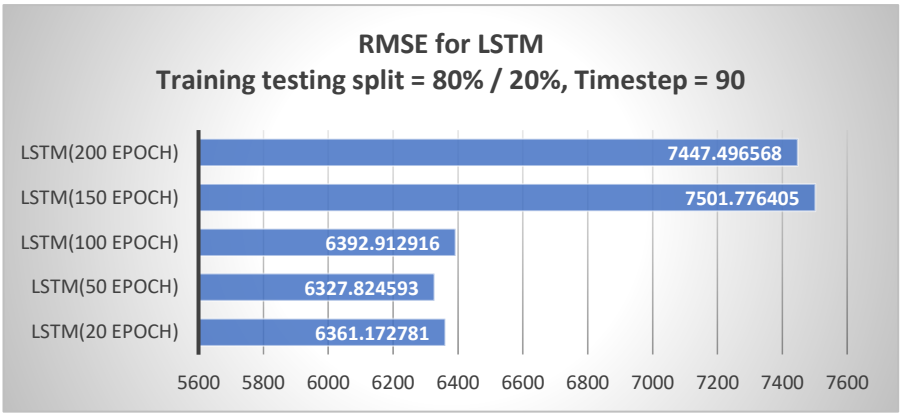

FIGURE 48 : DIFFERENT RMSE VALUES FOR DIFFERENT EPOCH IN LSTM MODEL(TIMESTEP=90,TRAINING SET = 80% AND TEST SET = 20% OF COVID 19 CONFIRMED CASES KERALA DATASET)

From Figure 43 to Figure 48, the lowest RMSE value where training set = 70% and test set = 30% of the dataset is 5212.363565 (epoch = 50, timestep=60) and lowest RMSE value where training set = 80% and test set = 20% of the dataset is 5780.4793673180775 (epoch = 50, timestep=60)

LSTM(epoch = 50, timestep=60) produces RMSE = 5212.363565 where 70% of the dataset has been used for training purpose and 30% for testing purpose.

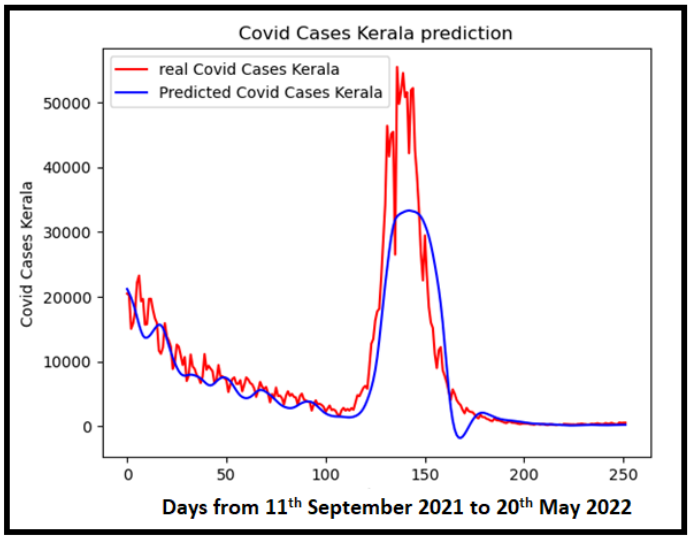

FIGURE 49 : ACTUAL CONFIRMED CASES VS LSTM(EPOCH = 50, TIMESTEP=60) PREDICTIONS FOR COVID 19 CONFIRMED CASES KERALA DATASET(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Figure 49 represents the confirmed cases of Covid 19 in Kerala from 11th September 2021 to 20th May 2022 and predicted confirmed cases by the LSTM(epoch = 50, timestep=60) prediction algorithm where the x-axis shows the date of the occurrence of confirmed cases and y – axis shows the number of confirmed cases. Here, 70% of the dataset has been used for training purpose and 30% for testing purpose.

LSTM(epoch = 50, timestep=60) produces RMSE = 5780.4793673180775 where 80% of the dataset has been used for training purpose and 20% for testing purpose.

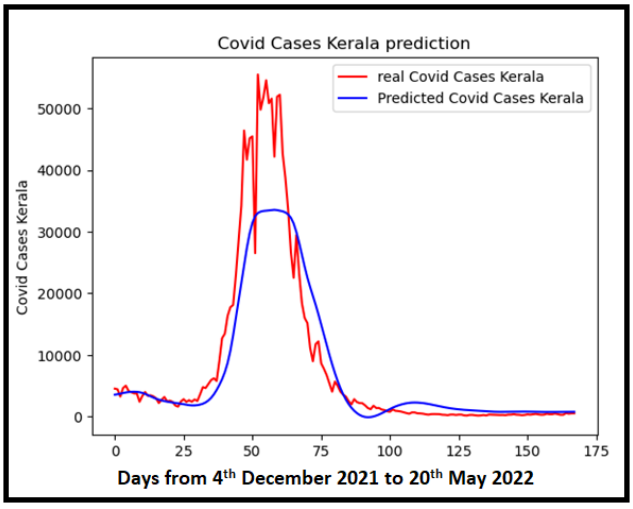

FIGURE 50 : ACTUAL CONFIRMED CASES VS LSTM(EPOCH = 50, TIMESTEP=60) PREDICTIONS FOR COVID 19 CONFIRMED CASES KERALA DATASET(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Figure 50 represents the confirmed cases of Covid 19 in Kerala from 4th December 2021 to 20th May 2022 and predicted confirmed cases by the LSTM(epoch = 50, timestep=60) prediction algorithm where the x-axis shows the date of the occurrence of confirmed cases and y – axis shows the number of confirmed cases. Here, 80% of the dataset has been used for training purpose and 20% for testing purpose.

4.6. Comparative Analysis

In the Result Analysis section, all the results of four models using Simple Linear Regression, Multiple Linear Regression, Random Forest, LSTM on 3 datasets namely, Adenovirus-Bay Area, Covid 19 Confirmed Cases-Kerala, hungary_chickenpox) are discussed. Now on a specific dataset, performances of models are discussed and according to lower RMSE value the best fit model and the worst fit model is declared.

4.6.1. Dataset - Adenovirus-Bay Area

Model Name	Training data length	Testing data length	RMSE
Linear regression	255(70%)	110(30%)	3.5263457845865958
MLR(prev. 3 weeks)	255(70%)	110(30%)	3.107213906665679
RF(prev. 3 weeks)	255(70%)	110(30%)	3.590153632677852
LSTM (no. of epoch=150, timestep=60)	255(70%)	110(30%)	4.2228220170991

TABLE 2 : RMSE SCORE OF DIFFERENT MODELS FOR ADENOVIRUS – BAY AREA DATASET(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

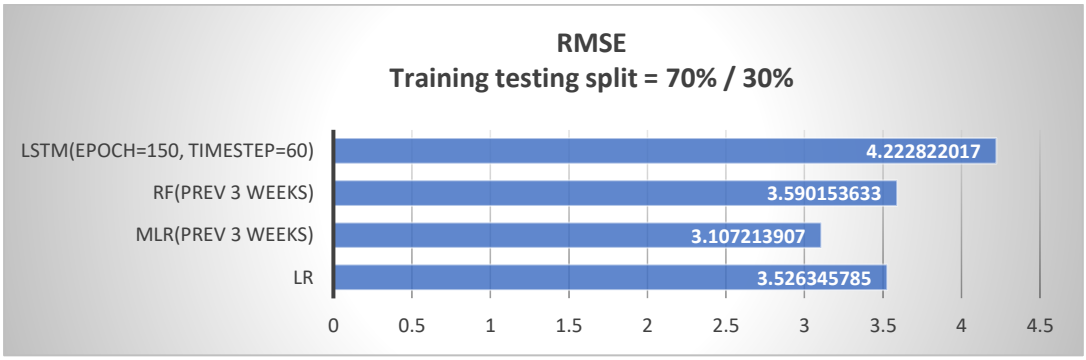

FIGURE 51 : RMSE SCORE OF DIFFERENT MODELS FOR ADENOVIRUS – BAY AREA DATASET(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET

Model Name	Training data length	Testing data length	RMSE
Linear regression	292(80%)	73(20%)	3.6123305769740273
MLR(prev. 3 weeks)	292(80%)	73(20%)	3.106168140083212
RF(prev. 3 weeks)	292(80%)	73(20%)	3.850452235342625
LSTM (no. of epoch=50, timestep=60)	292(80%)	73(20%)	4.003256018900395

TABLE 3 : RMSE SCORE OF DIFFERENT MODELS FOR ADENOVIRUS – BAY AREA DATASET(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

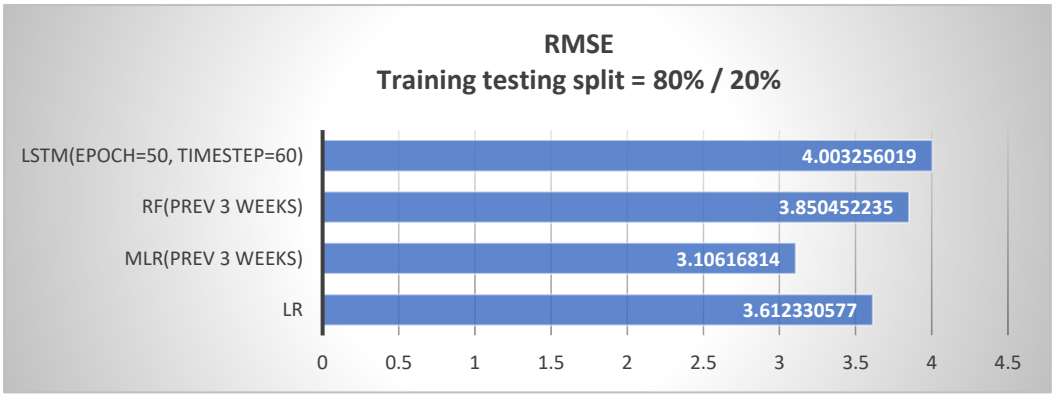

FIGURE 52 : RMSE SCORE OF DIFFERENT MODELS FOR ADENOVIRUS – BAY AREA DATASET(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Table 1 and 2 and figure 51 and 52 show that Multiple linear regression produced the least RMSE = 3.107213906665679(training set = 70% and test set = 30% of the dataset) and 3.106168140083212(training set = 80% and test set = 20% of the dataset) value than Simple Linear Regression, Random Forest and LSTM, which shows the better performance of Multiple linear regression. RMSE of Simple Linear regression is greater than Multiple Linear regression But lesser than Random forest and LSTM. RMSE of Random forest is greater than Multiple Linear regression and Simple linear regression but lesser than LSTM. RMSE of LSTM is greater than all other models(Simple linear regression, Multiple Linear regression, Random Forest) so the performance of LSTM on Adenovirus – Bay Area dataset is worst.

Figure 13 shows the graph(actual cases vs predicted cases) of Multiple linear regression on Adenovirus – Bay Area dataset, RMSE = 3.106168140083212(least RMSE value when training set = 80% and test set = 20% of the dataset)

Figure 14 shows the graph(actual cases vs predicted cases) of Multiple linear regression on Adenovirus – Bay Area dataset, RMSE = 3.107213906665679(least RMSE value when training set = 70% and test set = 30% of the dataset)

4.6.2. Hungary_chickenpox

Model Name	Training data length	Testing data length	RMSE
Linear regression	365(70%)	157(30%)	27.826431981713657
MLR(prev. 5 days)	365(70%)	157(30%)	25.561457096623982
RF(prev. 10 days)	365(70%)	157(30%)	24.79959850455771
LSTM (no. of epoch=200, timestep=90)	365(70%)	157(30%)	25.421684491872412

TABLE 4 : RMSE SCORE OF DIFFERENT MODELS FOR HUNGARY CHICKENPOX DATASET(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

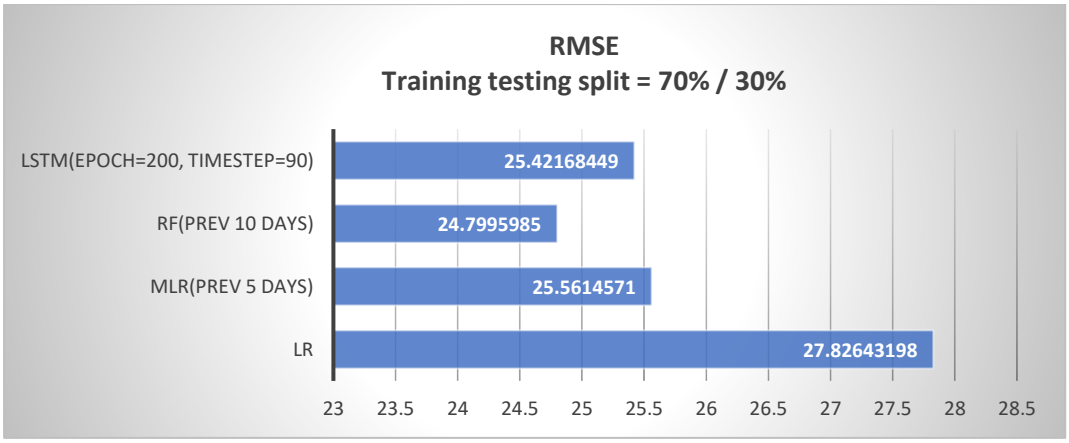

FIGURE 53 : RMSE SCORE OF DIFFERENT MODELS FOR HUNGARY CHICKENPOX DATASET(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Model Name	Training data length	Testing data length	RMSE
Linear regression	418(80%)	104(20%)	32.484645567112175
MLR(prev. 5 days)	418(80%)	104(20%)	29.747966116900173
RF(prev. 10 days)	418(80%)	104(20%)	29.19783272054491
LSTM (no. of epoch=50, timestep=30)	418(80%)	104(20%)	28.89815793399269

TABLE 5 : RMSE SCORE OF DIFFERENT MODELS FOR HUNGARY CHICKENPOX DATASET(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

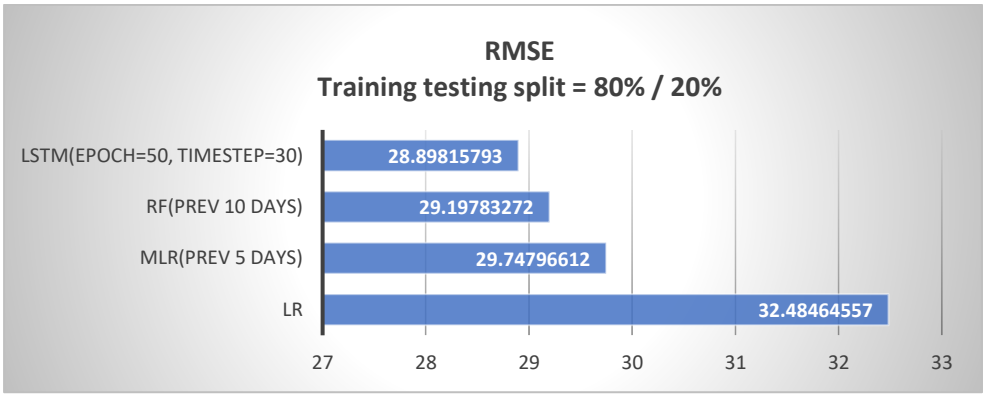

FIGURE 54 : RMSE SCORE OF DIFFERENT MODELS FOR HUNGARY CHICKENPOX DATASET(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Table 3 and 4 and figure 53 and 54 show that LSTM and Random forest shows better result than Multiple linear regression and Simple linear regression and LSTM and Random forest produces the least values of RMSE.

Figure 42 shows the graph(actual cases vs predicted cases) of LSTM on Hungary Chickenpox dataset, RMSE = 28.89815793399269(least RMSE value when training set = 80% and test set = 20% of the dataset)

Figure 26 shows the graph(actual cases vs predicted cases) of Random forest on Hungary Chickenpox dataset, RMSE = 24.79959850455771(least RMSE value when training set = 70% and test set = 30% of the dataset)

4.6.3. Covid 19 Confirmed Cases-Kerala

Model Name	Training data length	Testing data length	RMSE
Linear regression	589(70%)	252(30%)	3063.8380423113335
MLR(prev. 10 days)	589(70%)	252(30%)	2910.8113378352673
RF(prev. 5 days)	589(70%)	252(30%)	3601.5329058843618
LSTM (no. of epoch=50, timestep=60)	589(70%)	252(30%)	5212.363564678058

TABLE 6 : RMSE SCORE OF DIFFERENT MODELS FOR COVID 19 CONFIRMED CASES KERALA DATASET(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

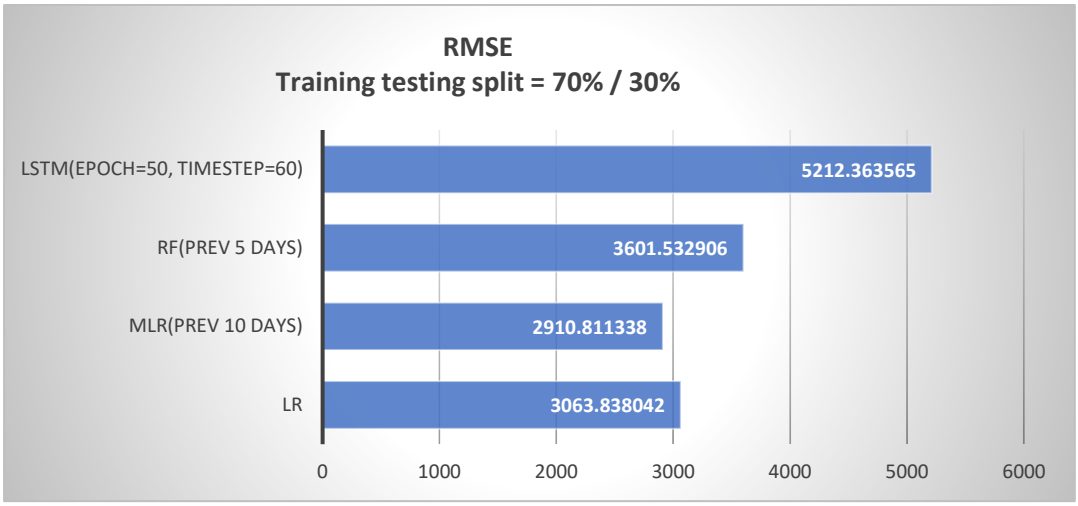

FIGURE 55 : RMSE SCORE OF DIFFERENT MODELS FOR COVID 19 CONFIRMED CASES KERALA DATASET(TRAINING SET = 70% AND TEST SET = 30% OF THE DATASET)

Model Name	Training data length	Testing data length	RMSE
Linear regression	673(80%)	168(20%)	3532.3447049312704
MLR(prev. 10 days)	673(80%)	168(20%)	3315.1235691891816
RF(prev. 5 days)	673(80%)	168(20%)	4202.757072430393
LSTM (no. of epoch=50, timestep=60)	673(80%)	168(20%)	5780.4793673180775

TABLE 7 : RMSE SCORE OF DIFFERENT MODELS FOR COVID 19 CONFIRMED CASES KERALA DATASET(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

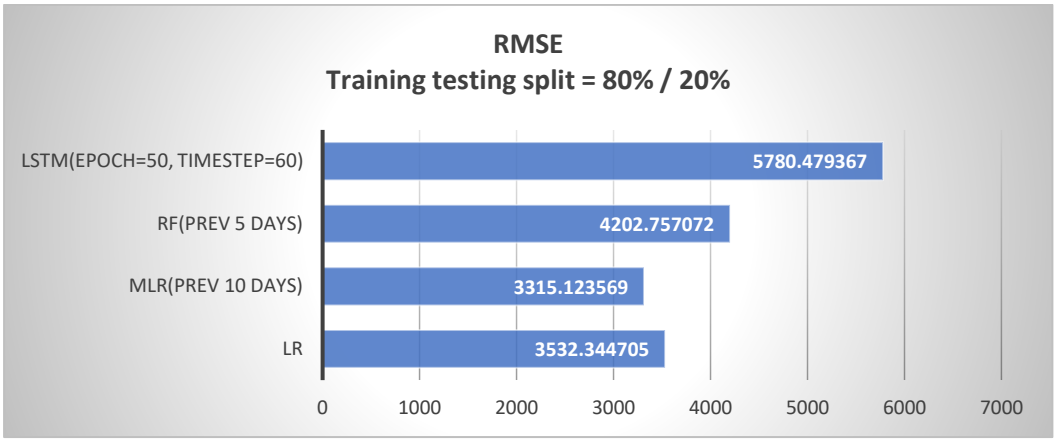

FIGURE 56 : RMSE SCORE OF DIFFERENT MODELS FOR COVID 19 CONFIRMED CASES KERALA DATASET(TRAINING SET = 80% AND TEST SET = 20% OF THE DATASET)

Table 5 and 6 and figure 55 and 56 show that Multiple linear regression produced the least RMSE = 2910.811338 (training set = 70% and test set = 30% of the dataset) and 3315.123569(training set = 80% and test set = 20% of the dataset) value than Simple Linear Regression, Random Forest and LSTM, which shows the better performance of Multiple linear regression. RMSE of Simple Linear regression is greater than Multiple Linear regression But lesser than Random forest and LSTM. RMSE of Random forest is greater than Multiple Linear regression and Simple linear regression but lesser than LSTM. RMSE of LSTM is greater than all other models(Simple linear regression, Multiple Linear regression, Random Forest) so the performance of LSTM on Covid 19 Confirmed Cases-Kerala dataset is worst.

Figure 21 shows the graph(actual cases vs predicted cases) of Multiple linear regression on Covid 19 Confirmed cases Kerala dataset, RMSE = 3315.1235691891816 (least RMSE value when training set = 80% and test set = 20% of the dataset)

Figure 22 shows the graph(actual cases vs predicted cases) of Multiple linear regression on Covid 19 Confirmed cases Kerala dataset, RMSE = 2910.8113378352673 (least RMSE value when training set = 70% and test set = 30% of the dataset)

Chapter 5

Conclusion and future scope

The project work deals with three different time series datasets having data related to three different disease outbreaks. Four different machine learning models are trained by these datasets. The key goal of this project is to find best machine learning which can forecast the future outbreak of these diseases.

There are some preprocessing is done to get the actual datasets to fit directly to the models

The preprocessed datasets are:

1. covid19-confirmed-cases-kerala
2. Hungary(BARANYA) chickenpox cases
3. Adenovirus cases in Bay Area

The considered machine learning models are,

1. Simple Linear Regression
2. Multiple Linear Regression
3. Random Forest
4. LSTM

After an exhaustive set of experiments, it has been observed that

1. covid19-confirmed-cases-kerala works better in Multiple linear regression generating RMSE between 2910.8113378352673 and 3315.1235691891816
2. Hungary (BARANYA) chickenpox cases dataset works better in LSTM and Random Forest generating RMSE between 24 to 30
3. Adenovirus cases in Bay Area dataset works better in Multiple linear regression generating RMSE approximately 3.1

Some key limitations are

1. Small size of dataset: The datasets which are used in this project are very short in length. That's why it creates a generalization problem.
2. Lack of features in the dataset: No specified dependent features are mentioned in the dataset to predict the infected cases.
3. Time and hardware requirement in LSTM : LSTM models can be time consuming as it requires high hardware resource and software optimization.

A few ideas which can be incorporated in future to improve the work are described here.

1. Length of the datasets : Large datasets can be extremely beneficial in improving the accuracy and reliability of forecasting models. More data can help the model identify patterns and relationships in the data, leading to more accurate predictions.

2. Regular updates: As new data becomes available, it's important to retrain the machine learning model regularly to keep it up to date and accurate. This helps ensure that the model continues to provide accurate forecasts as new samples are added.

3. Ensemble models: Ensemble methods combine multiple machine learning models to improve accuracy. This involves training multiple models on the same dataset and combining their predictions to create a more accurate and reliable forecast. Hence, different ensemble approaches can be implemented in future.

Bibliography

[1] C. Chatfield, Time-series forecasting, Chapman & Hall/CRC, 2001.

[2] "find-out-how-to-use-machine-learning-for-time-series-forecasting," [Online]. Available: https://nix-united.com/blog/find-out-how-to-use-machine-learning-for-time-series-forecasting/.

[3] "CORONAVIRUS IN THE WORLD SUNDAY APRIL 9, 2023: NEW CASES AND DEATHS IN 24 HOURS," 9 April 2013. [Online]. Available: https://www.sortiraparis.com/en/news/coronavirus/articles/212134-coronavirus-in-the-world-new-cases-and-deaths-in-24-hours.

[4] D. H. B. M. Jeremy Page, "In Hunt for Covid-19 Origin, Patient Zero Points to Second Wuhan Market," *The Wall Street Journal,* 26 February 2021.

[5] S. K. ,. S. A. ,. H. M. A. K. R. H. Md Asiful Islam, "Prevalence and characteristics of fever in adult and paediatric patients with coronavirus disease 2019 (COVID-19): A systematic review and meta-analysis of 17515 patients," *Plos One,* 6 April 2021.

[6] S. S. A. S. K. T. H. M. A. K. C. C. Md Asiful Islam, "Prevalence of Headache in Patients With Coronavirus Disease 2019 (COVID-19): A Systematic Review and Meta-Analysis of 14,275 Patients," *Frontiers in Neurology,* 27 November 2020.

[7] M. A. I. B. A. Jeyasakthy Saniasiaya, "Prevalence of Olfactory Dysfunction in Coronavirus Disease 2019 (COVID-19): A Meta-analysis of 27,492 Patients," *Laryngoscope,* April 2021.

[8] M. A. I. B. A. Jeyasakthy Saniasiaya, "Prevalence and Characteristics of Taste Disorders in Cases of COVID-19: A Meta-analysis of 29,349 Patients," *American Academy of Otolaryngology—Head and Neck Surgery Foundation,* 15 December 2020.

[9] "Coronavirus Pandemic (COVID-19)," 8 April 2023. [Online]. Available: https://ourworldindata.org/coronavirus.

[10] "India's first coronavirus case: Kerala student in Wuhan tested positive," *Business Standard,* 1 May 2023.

[11] "India's first coronavirus patient discharged after being cured," *Hindustan Times,* 20 February 2020.

[12] D. D. B. L. J. M. 3. L. T. K. H A Guess, "Population-based studies of varicella complications," *PubMed,* vol. 78, no. 4 Pt 2, pp. 723-7, 1986.

[13] H. F. Judith Breuer, "Chickenpox," *National Library of Medicine, PubMed Central,* 11 April 2011.

[14] C. N. Holmes, "Predictive value of a history of varicella".

[15] D. P. J. Z. L. M. K. V. C. J. A. S. L. B. W. R. C. Stephanie R Bialek, "Impact of a routine two-dose varicella vaccination program on varicella epidemiology," *National Library of Madicine, PubMed,* November 2013(Vol. 132, pp. 1134 - 1140).

[16] “Adenovirus,” 28 Movember 2022. [Online]. Available: https://www.cdc.gov/adenovirus/.

[17] “Prevention & Treatment,” 16 February 2023. [Online]. Available: https://www.cdc.gov/adenovirus/prevention-treatment.html.

[18] A. Z. M. M. Farah Shahid, “Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM,” *Chaos, Solitons and Fractals ,* 2020(Vol. 140).

[19] V. S. P. M. C. M. T.-G. Amit Kumar Gupta, “Prediction of COVID-19 pandemic measuring criteria using support vector machine, prophet and linear regression models in Indian scenario,” *Journal of Interdisciplinary Mathematics,* 2020(Vol. 24, pp. 89-108).

[20] A. T. A. R. T. Smita Rath, “Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model,” *Diabetes & Metabolic Syndrome: Clinical Research & Reviews,* 2020(Vol. 14, pp. 1467 - 1474).

[21] C. M. Yesilkanat, “Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm,” *Chaos, Solitons and Fractals,* 2020(Vol. 140).

[22] A. T. a. D. T. V. Wadie Skaf, “Towards Automatic Forecasting: Evaluation of Time-Series Forecasting Models for Chickenpox Cases Estimation in Hungary,” 2022.

[23] P. S. O. K. R. S. T. F. Benedek Rozemberczki, “Chickenpox Cases in Hungary: a Benchmark Dataset for Spatiotemporal Signal Processing with Graph Neural Networks,” 2021.

[24] L. Vinay Kumar Reddy Chimmula, “Time series forecasting of COVID-19 transmission in Canada using LSTM networks,” *Chaos, Solitons and Fractals,* 2020(Vol. 135).

[25] B. Mahesh, “Machine Learning Algorithms - A Review,” *International Journal of Science and Research (IJSR),* 2018.

[26] S. I. Bangdiwala, “Regression: simple linear,” *taylor and francis online,* 2018.

[27] N. G. Gülden Kaya Uyanık, “A Study on Multiple Linear Regression Analysis,” *Procedia - Social and Behavioral Sciences,* 2013.

[28] M. K. Naomi Altman, “Ensemble methods: bagging and random forests,” *Nature,* 2017.

[29] L. Breiman, “Random Forests,” *Kluwer Academic Publishers. Manufactured in The Netherlands.,* 2001.

[30] Y. B. G. H. Yann LeCun, “Deep learning,” *Nature,* 2015.

[31] A. P. J. Kamilya Smagulova, “A survey on LSTM memristive neural network architectures and applications,” *EDP Sciences, Springer-Verlag GmbH Germany, part of Springer Nature,* 2019.

[32] E. R. M. Ralf C. Staudemeyer, “Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks,” 2019.