Covid-19 Patient Count Prediction Using LSTM

Chetan Pandey,
Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Abstract

In Wuhan, China,the COVID-19 pandemic first emerged, in December 2019, and it has reached to over nearly 200 countries in a matter of weeks.Every country that was affected by the disease began taking the necessary precautions to avert it from spreading, give sick individuals the best medical care possible, and take preventative measures to restrict the spread. The fact that those with modest symptoms and others who don't experience any negative effects might distribute COVID-19 is the cause of its extensive distribution. Individuals' pre-existing medical conditions, like a cardiovascular problem, high blood pressure, diabetes, weakened immune system, or age-related frailty, are also linked to critical cases and fatality rates. The time series prediction approach was used by the forecasting model.A dataset repository, such as the UCI repository, serves as the input for the Pakistan Covid-19 dataset. The pre-processing procedure must then be carried out. The data separation must then be put into practise. We must divide the data into test and train groups in this phase. The feature selection approach must then be put into practise. The LSTM algorithm must then be used to forecast the covid-19 patient population in Pakistan. Last but not least, the research findings indicate that the MAPE (Mean absolute Percentage Error).

DOI Number: 10.48047/nq.2021.19.6.NQ21086 **NeuroQuantology2021;19(6): 194-199**

1. Introduction

Viruses are tiny creatures that can only reproduce inside an organism's live cells. One of the most prevalent virus groups responsible for a variety of respiratory illnesses in living things is the coronavirus family. One of the difficulties in recent years is predicting the spread of the coronavirus. High precision pandemic forecasting will aid many nations in creating a strategy to wage war against the virus' expansion. The modelling of real-world issues makes considerable use of machine learning techniques. In particular, machine learning techniques have been utilised extensively recently to forecast illnesses. Regression modelling and time series modelling are two conventional techniques for predicting an outbreak. We have selected both classic and contemporary machine learning methods.

While we have contrasted some of these ancient strategies with a new one, the Support Vector Machine methodology has typically been utilised for predictions. Starting December 2019, Wuhan (China) has seen a global epidemic of the sickness caused by the COVID-19 virus, which originally claimed 2873 lives alone in China and 104 lives outside of the nation. Up until February 2020, it raised the mortality rate by correspondingly 3.6% and 1.5%. While it started in China, it quickly spread over the world and had a high fatality rate, particularly in the US, Italy, the UK, and Spain. Over 80 000 cases were recorded in China as of March 1, 2020, while 7 200 people were spotted elsewhere in the world.

The coronavirus outbreak first appeared in China, but by March 15, 2020, the region's population had overtaken that of Europe and the USA. There are currently more deaths in a few areas than there are overall China. Positive carriers numbered more than three million globally by the end of April 2020, with more than 210 000 fatalities. COVID-Any one or a combination of more than one symptoms, such as a fever, a cough, diarrhoea, a sore throat, a rapid respiratory rate, low oxygen saturation, or difficulty breathing, are the starting point for 19 early signs. A condition called acute respiratory distress syndrome might develop as a result of this in the future. The virus has been

identified in millions of individuals worldwide since December 2019, killing thousands of people.
The World Health Organization and study have found that COVID-19 is mostly disseminated by respiratory beads, although it may also spread from hard surfaces. The main factor causing their prevalence is these asymptomatic instances. Due to its rising fatality rate, the corona virus has established deep roots and may be highly deadly. A rare times, the development of symptoms might lead to death from severe alveolar injury and respiratory failure.
Our project's primary goals are to detect covid-19 positive patients, use deep learning techniques like Long Short Term Memory (LSTM), and improve overall performance. Early signs of COVID-19 include any one or a combination of more than one of the following: fever, cough, diarrhoea, sore throat, rapid breathing, low oxygen saturation, or shortness of breath. Finding many cases of COVID patients at once is challenging; deep learning algorithms can help solve this challenge.

2. Literature Survey

Many illnesses in mammals and birds are brought on by a group of viruses known as coronaviruses. They lead to a variety of respiratory conditions in people. One of the difficulties in modern times is predicting the spread of the coronavirus. High precision pandemic forecasting will aid many nations in creating a strategy to wage war against the virus' expansion. This study analyses the COVID19 disease's propagation and makes projections regarding the pandemic's scope, survival rates, and death rates. Together with mathematical modelling approaches including Rough Set Support Vector Machine, Bayesian Ridge as well as Polynomial Regression, SIR model, also RNN, we have employed several well-known machine learning techniques.Using polynomial regression has the benefit of overcoming variable dependence that may not be linear. The advantage of LSTM is that it is a suitable classification, processing, and prediction method. The ability to be improved when additional data enters the database was the main benefit of utilising a Bayesian estimator-based method. Nevertheless, this method's primary drawback is that the choice of prior and posterior probability has an impact on it [1].
The fundamental reproduction number R0 at the national and state levels is estimated in this policy article using the epidemiological SIR model and statistical machine learning model. The study shows that Punjab's (R0 16) status is not good and needs immediate proactive adjustment. Madhya Pradesh (3.37), Maharashtra (3.25), and Tamil Nadu (3.09) all have R0 values that are higher than 3. As of 04 March 2020, the R0 of Andhra Pradesh (2.96), Delhi (2.82), and West Bengal (2.77) is higher than the R0 of India (2.75). At the early stages of the illness, Hubei/China and India both have R0 values of 2.75. With lockdown in place, India could anticipate just as many instances as China, if not more. By May 1, 2020, India should anticipate fewer than 66,224 instances if lockdown is successful.The main benefit of the SIR model is that it provides a number R0 that are capable of assessing as well as benchmark the actual state of affairs in various states and allocate resources to the hardest-hit ones [2].
Malaysia is now seeing a fast spread of the coronavirus COVID-19. By April 5, 2020, there were 3662 confirmed instances of infection nationwide, resulting in a state of lockdown. Considering the incorporation of the mortality cases, this study attempts to predict the epidemic peak via the Susceptible-Exposed-Infectious-Recovered (SEIR) paradigm, since the primary public worry is whether the current scenario will persist for the next several months. The Adaptive Neuro-Fuzzy Inference System paradigm was employed to offer short-term prediction of infected patients' numbers, while the Genetic Algorithm (GA) was utilised to predict the infection rate.If the government took action, the epidemic peak would be postponed by 30 and 46 days, respectively, if the infection rate were to be reduced by 25% over the course of two or three months. The Normalized Root Mean Square Error (NRMSE), Mean Absolute Percentage Error (MAPE), and coefficient of determination (R2) for the predicting results utilising the ANFIS model are all low at 0.041, 2.45%, and 0.9964, respectively. The findings also demonstrate a significant impact that an actionon postponing the peak of the epidemic and that a longer intervention time would result in a smaller pandemic at the peak. For the purpose of containing the COVID-19 outbreak, the study offers crucial information to public health professionals and the government. Short-term forecasting utilising parametric models could not have a high degree of accuracy [3].
The author looks at the issue of simulating the new corona virus's (Covid19) spread in a realistic or limited situation in India. It has been noted that isolation and lock-down are crucial measures to stop the disease from spreading. We quantitatively analyse the effect of these prophylactic steps on Covid19 spread, and we provide a novel mathematical model to forecast the number of new cases or all infected individuals in real-world scenarios. To build up medical facilities and go on with future plans of action, this prognosis is very necessary. For the Covid19 spread, a new model for limited situation is suggested. World Health

Organization (WHO) has designated Covid19 to be a pandemic. Thus, there has to be enough planning for medical facilities everywhere.The number of new afflicted people is growing quickly because there is no vaccination for Covid19 and the illness is spreading. A tree-based model is taken into consideration, in which some individuals are quarantined and a small number go unnoticed (hidden nodes) for a variety of reasons, such as failing to display symptoms or concealing prior travel, etc.; these hidden nodes then spread the illness throughout the neighbourhood. The accuracy of the developed model is demonstrated by a close agreement between the analytical results and the accessible results [4].

The coronavirus disease 19 (COVID-19) has been the subject of a large number of clinical studies that concentrated on certain laboratory variables; however, these reports have not been the subject of a systematic study that summarised these findings. — to examine and summarise the most recent research on the prognostic value of several biomarkers in COVID-19 patients. — PubMed, medRxiv, and bioRxiv are only a few of the databases that were used for the literature search. 72 publications in all, including 54 peer-reviewed and 18 unreviewed preprints, were reviewed. C-reactive protein (CRP), ferritin, serum amyloid A (SAA), plus procalcitonin have all been found as acute-phase reactants that are sensitive markers of COVID-19 disease, despite the fact that the markers are thought to be nonspecific.Significantly raised white blood cell count, pronounced lymphopenia, reduced CD3, CD4, or CD8 T-lymphocyte counts, high neutrophil count, thrombocytopenia, as well as significantly Higher levels of inflammatory biomarkers have been associated with severe sickness and a higher risk for acquiring sepsis with quick development. In critically ill patient groups, trends like progressive lymphocyte count decline, thrombocytopenia, elevated CRP, procalcitonin, increased liver enzymes, decreased renal function, also coagulation derangements were more frequent, and furthermore connected to a greater incidence of clinical complications. These groups also showed elevated neutrophil to lymphocyte ratios, which are inflammatory biomarkers and indicators of systemic inflammation [5].

3. PROPOSED SYSTEM

One of the most serious problems of this century, the Covid-19 pandemic has an impact on social standards and economic operations. On the basis of Pakistani Covid-19 data from March 2020 to May 2020, we constructed a recurrent neural network paradigmto predict for June 2020 the Covid-19 Proportion of Positive Patients. To determine the model's predictive accuracy on various LSTM units, batch sizes, and epochs, we calculated the mean absolute percentage error (MAPE). At the same time period, projected patients are also evaluated with a prediction methodology, and the findings show that the suggested model's predicted patient count is significantly more comparable to the real patient count.

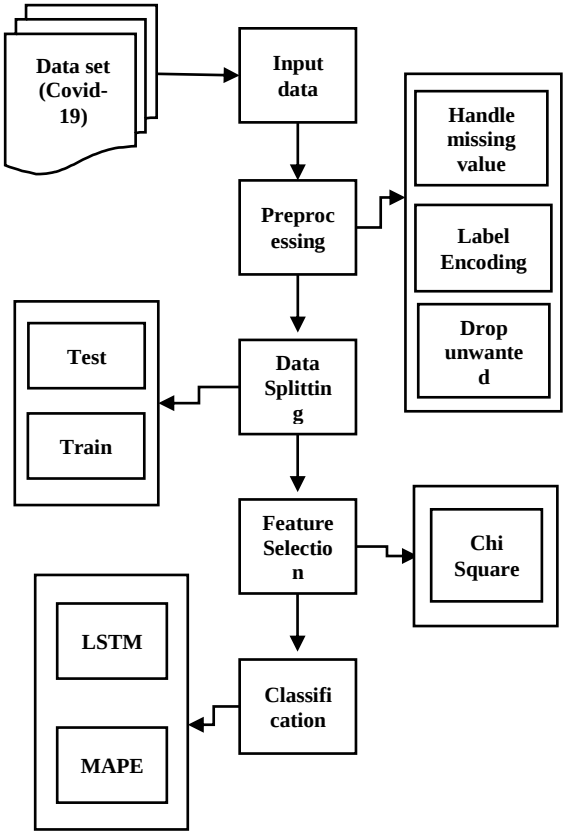

Fig 1: System Architecture

The Covid-19 dataset served as the system's source of input. The source of the input data was a dataset repository, such as the UCI repository. The data preparation procedure must then be implemented. At this phase, we must deal with the missing values to prevent incorrect prediction, encode the incoming data's label, and remove any unnecessary columns. After that, we must put data splitting into practise. We must divide the data into test and train groups in this phase. The feature selection approach must then be used.To find the best characteristics in this stage, we must apply the chi-square statistic. Finally, in order to get higher performance and because this is a time series dataset, we must employ a deep learning technique such as LSTM (Long Short Term Memory). The findings of the experiment indicate that performance measures like Mean Absolute Percentage Error (MAPE).

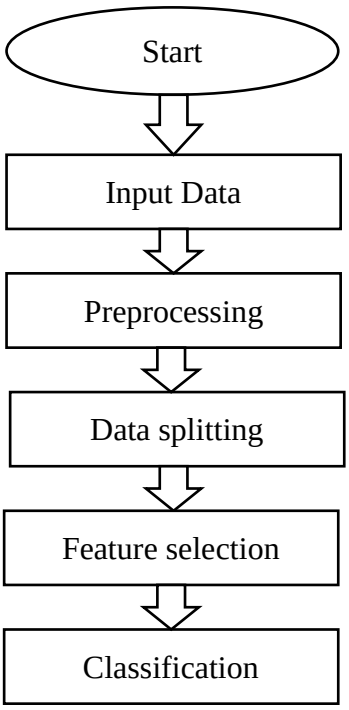

Fig 2: Flow Diagram

Following is a list of the several benefits of the suggested approach:

- With a large number of datasets, it is effective.
- When compared to the current system, the experimental outcome is excellent.
- Moreover, to put into practise feature selection in order to extract the best features possible from our input data.

The next section provides an explanation of the many phases that are involved in putting the suggested technique into practise:

1. Data selection

The source of the input data was a dataset repository. The Covid-19 dataset is employed in our procedure. The process of finding the covid-19 patients is called data selection. The dataset has seven columns for date, the number of incidents, the number of fatalities, and the number of survivors, the cases' travel histories, and the incidents' locations (province and city).

2. Data preprocessing

Pre-processing data involves deleting unnecessary information from the dataset. The dataset is transformed applying pre-processing data transformation procedures into a structure appropriate for machine learning. Cleaning the dataset in this stage also makes it more effective by eliminating faulty or unnecessary data that might impair the dataset's accuracy.

3. Data splitting

Data are required for training and testing purposes during the machine learning process. The Bot-IoT dataset that we used for this approach had 70% training data and 30% testing data. Dividing up readily available data into two parts, often for cross-validator needs, is known as data splitting. The majority of the information is intended for training. As well as a lesser amount is used for testing when data is divided into training and testing sets. This is crucial for assessing data mining models.

4. Feature selection

While creating a predictive model, the process of feature selection involves lowering the number of input variables. To create a simple security solution suitable for IoT systems, it is important to reduce the number of characteristics and only utilise those required to train and test the algorithms. Method for selecting features that reduces the feature's size and the computation required by the machine learning algorithm. We have employed chi square methods in this procedure. Chi Square: By choosing this feature, a predictive model will provide its best feature.

5. Classification

We must incorporate deep learning algorithms like LSTM into our workflow. An artificial recurrent neural network referred to as LSTM, or long short-term memory, is utilised in deep learning. LSTM features feedback connections in contrast to traditional feed forward neural networks. Along with single data points, it can examine whole data streams. LSTM networks are ideally adapted to categorising, processing, and then producing predictions based on time series data since there may be delays of varying lengths between major events in a time series.

6. Result Generation

On the basis of the overall categorization and forecast, the Final Result will be created. The effectiveness of this suggested strategy is assessed using some metrics, such as,

The mean absolute percentage error (MAPE) is a regularly used metric to assess how well models forecast the future. It is determined by:

MAPE = (1/n) * Σ(|actual – prediction| / |actual|) * 100

Where:

Σ – a symbol that means “sum”

n – sample size

actual – the actual data value

prediction – the predicted data value

4. RESULTS

In Wuhan, China,the COVID-19 pandemic first emerged, in December 2019, and It has reached to

over nearly 200 countries in a matter of weeks. The input for the forecasting model, the Pakistan Covid-19 dataset, was collected from a dataset repository like the UCI repository and used the time series prediction approach. The experimental findings demonstrated that the MAPE (Mean absolute Percentage Error) was attained. The pre-processing stage, data splitting, feature selection approach, and LSTM algorithm were all implemented. Preprocessing of the input data included managing missing values, encoding the label, and removing unnecessary columns once it was downloaded from the dataset source. The best features were then found using the feature selection approach after the data had been divided into test and train sets.In order to improve performance, a deep learning algorithm like LSTM was implemented. According to the trial findings, performance parameters such Mean Absolute Percentage Error (MAPE) were reached, as indicated by the pictures below.

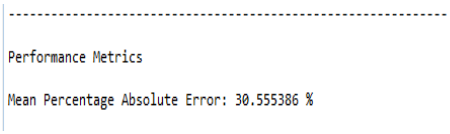

Fig 3: Performance Analysis

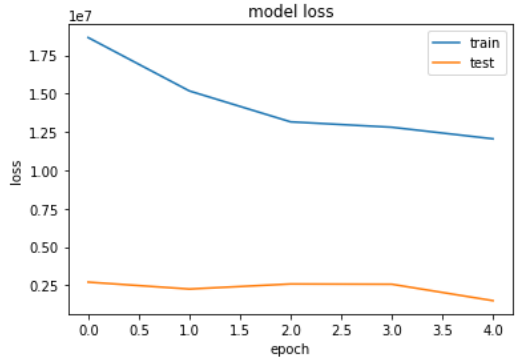

Fig 4: Validation

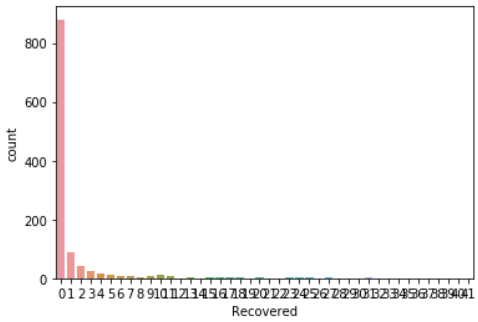

Fig 5: Visualization

5. CONCLUSION

We come to the conclusion that the Covid-19 dataset served as the input. Finally, in order to identify the best features, we employ feature selection techniques like chi-square. For improved performance, we applied LSTM (Long Short Term Memory), a deep learning technique. Also, because this is a time series dataset, we had to use this deep learning approach. The outcomes of the experiment demonstrate that performance indicators like Mean Absolute Percentage Error (MAPE).

6. FUTURE ENHANCEMENT

We want to train the model on death cases in the future and attempt to predict the mortality ratio and its relationships with critical cases. Using the revised dataset, the next step is to investigate the prediction approach and choose the most accurate deep learning predicting strategies. Future efforts will be heavily focused on real-time, live forecasting.

REFERENCE

[1] F. Wu et al., "A new coronavirus associated with human respiratory disease in China," Nature, vol. 579, no. 7798, pp. 265–269, 2020.
[2] C. Huang et al., "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," Lancet, vol. 395, pp. 497–506, May 2020.
[3] Coronavirus Disease 2019 (COVID-19) Situation Report, World Health Organization, Geneva, Switzerland, 2020.
[4] E. Dong, H. Du, and L. Gardner, "An interactive Web-based dashboard to track COVID-19 in real time," Lancet Infectious Diseases, vol. 20, no. 5, pp. 533–534, May 2020.
[5] Coronavirus Disease 2019 (COVID-19) Situation Report, World Health Organization, Geneva, Switzerland, 2020.
[6] World Health Organization. Accessed: May 31, 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus2019
[7] K. J. Clerkin et al., "COVID-19 and cardiovascular disease," Circulation, vol. 141, no. 20, pp. 1648–1655, 2020.
[8] B. Resnik. World Health Organization. Accessed: May 31, 2020.
[9] Report of the Who-China Joint Mission on Coronavirus Disease 2019 (COVID-19), World Health Organization, Geneva, Switzerland, 2020.
[10] P. K. Ozili and T. Arun, "Spillover of COVID-19: Impact on the global economy," Tech. Rep., Mar. 2020.
[11] M. Gupta et al., "COVID-19 and economy," Dermatol. Therapy vol. 33, no. 4, 2020, Art. no. e13329
[12] A. Tariq et al., "Real-time monitoring the transmission potential of COVID-19 in Singapore, March 2020," BMC Med., vol. 18, no. 1, pp. 1–14, Dec. 2020.

[13] S. Deb and M. Majumdar, "A time series method to analyze incidence pattern and estimate reproduction number of COVID19," 2020, arXiv:2003.10655.
[14] A. J. Kucharski et al., "Early dynamics of transmission and control of COVID-19: A mathematical modelling study," Lancet Infectious Diseases, vol. 20, no. 5, pp. 553–558, 2020.
[15] S. K. Dey, M. M. Rahman, U. R. Siddiqi, and A. Howlader, "Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach," J. Med. Virology, vol. 92, no. 6, pp. 632–638, Jun. 2020.