Performance Evaluation Of Machine Learning Techniques For COVID-19 Prediction From The Occurrence

Atul Tiwari[1], Manender Dutt[2], Prasath Alias Surendhar S[3], Sarika Panwar[4], J.Sundararajan[5], S.Karthikkumar[6], P.John Augustine[7]

[1]Department of Pathology, Government Medical College, Udaipur Road, Bojunda, Chittorgarh, Rajasthan, India

[2]University Institute of Computing, Chandigarh University, Mohali, Punjab, India

[3]Department of Biomedical Engineering, Aarupadai Veedu Institute of Technology (AVIT), Chennai, Tamil Nadu, India

[4]Department of Electronics and Telecommunication engineering, AISSMS'S Institute of Information Technology Pune India

[5]Department of Electronics and Communications Engineering, NPR College of Engineering and Technology

Dindugul, Tamil Nadu, India

[6]Department of Civil Engineering, Faculty of Engineering, Karpagam Academy of higher Education, Coimbatore, Tamil Nadu, India

[7]Department of Information Technology, Sri Eshwar college of Engineering, Kondampatti, Kinathukadavu, Coimbatore, Tamil Nadu, India

Corresponding mail id: atultiwari.in@gmail.com

ABSTRACT

The corona, often identified as SARS-CoV-2, has ravaged large sections of the globe, and the situation is bad. It is a kind of pandemic sickness that is distribution from individual to individual on a daily basis. It is critical to have pathway of the amount of patients that are afflicted as a result. Due to the present manner of electronic data collecting, it is difficult to evaluate and anticipate disease transmission both locally and worldwide. To get around this problem, machine learning methods might be employed to effectively map the illness and its evolution. Deep learning, a part of computer technology, is critical for accurately identifying persons with the illness by analysing chest X-ray pictures. Controlled artificial intelligence models with connected algorithms (such as LR, SVR, and Time series algorithms) for information analysis for regression and classification are useful for teaching the model to forecast the total amount of confirmed patients worldwide who will be susceptible to the illness in the coming days. Once the global dataset has been gathered, pre-processed, and retrieved, the number of verified occurrences up to a given date is acquired and used as the training set for the model in this proposed work. To anticipate the rise in occurrences over the following several days, the model is built using supervised machine learning methods. a schedule Holt's classical outperforms regression of linear and vector of support machine in the experimental situation when using the aforementioned methodologies.

Keyword: COVID 19, Holts model, machine learning, x-ray

1. INTRODUCTION

Corona Virus Illness (COVID-19) is caused by the viral illness SARS CoV-2. The primary instance of this illness was detected in December 2019 in, China. Since its discovery, the illness has spread around the globe, and the World Health Organization designated it a easily transmitted disease in March 2020 [1]. The disease has spread to over 100 nations in a very short period of time. The microorganism that causes COVID - 19 transmit when an infected being comes into close contact with others. When a virus-infected individual sneezes or coughs, little virus droplets are discharged into the air. Once an sick person originates into communication with contaminated outsides, the virus spreads [2].

Coughing, fever, weariness, weakness, loss of taste and smell, and, in rare cases, no symptoms at all are some of the signs of this condition. The sickness has the greatest impact on the lower and upper respiratory systems [3]. As a result, patients are at a significant risk of dying. Depending on their symptoms, a patient may be prescribed medication for this illness. Wear a mask, avoid ill individuals, avoid crowded places, wash your hands often, and take other measures. Various organisations have developed vaccines, and numerous nations have started significant immunisation campaigns. Several organisations are always working to develop new vaccinations. Despite continuous research to identify drugs that obstruct virus growth inside the human body, current treatment is indicative and unsuccessful. The drug may have negative side effects and cannot completely prevent the patient's death [4].

COVID-19 may be recognised by medical practitioners founded on indications or transportable history, and the virus may be identified through RT-PCR testing (RT-PCR). In addition to laboratory studies, chest CT scans may be utilised to notice COVID-19 in those are having a strong clinical suspicion of infection. Serological testing may identify antibodies generated by the body in comeback to an contamination as well as antibodies created during previous diseases [5].

Researchers have developed a variety of methods, including machine learning, to reliably detect the condition. Using text and visual data, machine learning aids in healthcare procedures and disease diagnostics. Monitoring COVID-19 development, in addition to diagnosis, is critical. A hereditary algorithm, a statistical development regression model, and a sigmoid model may all be used to forecast the number of infected patients in the coming days [6]. Furthermore, present algorithms only examine a small quantity of data and are geographically limited. They also anticipate future occurrences with less accuracy on a global scale. For enhancing prediction accuracy and analysing huge datasets, supervised machine learning methodologies—specifically, the time series forecasting methodology known as Holt's winter—are preferred above alternative techniques. To estimate future numerical values, statistical analysis may use supervised machine learning methods such as support vector regression (SVR) and linear regression (LR) [7]. a timetable A classifier-supervised machine learning methodology is a forecasting method that assists in predicting standards of a time series at numerous subsequent period points.

Using both regression and period series techniques, the proposed study trains the model to estimate the entire amount of positive occurrences globally in the next days [8]. The dataset is assembled, pre-processed to speed up processing, and then utilised to train and test the model. It provides the total amount of established, healthier, and fatal cases worldwide. As a result, the present study's major goal is to forecast a future increase in the amount of COVID-19 positive patients utilising the clinical text COVID-19 dataset (Fig. 1).

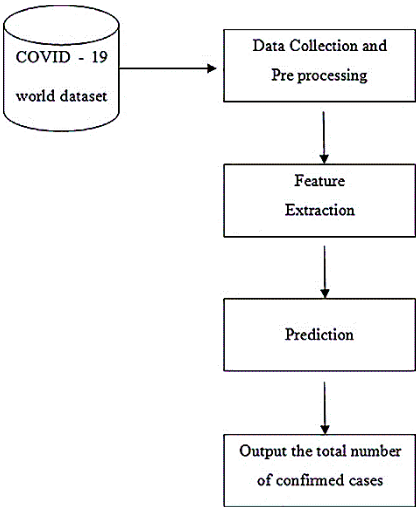

Fig. 1 Methodology of the research

The flowchart above shows how machine learning approaches are used to forecast the number of infected instances that will arrive. The three processes in the process of recognising forest smokes are preprocessing, feature extraction, and prediction. The dataset is initially converted into a CSV (Comma-Separated Values) file, which has rows and columns produced from many sources and containing various sorts of data [9]. The multi-variable dataset is then routed to pre-processing, where it is turned into a standard dataset that includes the time, state/province, nation, amount of confirmed, healthier, and death cases, among other details. Then, Holt's winter time series model is used to obtain LR, SVR, and SVR characteristics. Feature extraction is used to find the proper columns or data to feed the model. Future events that have been confirmed are collected after using the right techniques to predict the model.

2. Methodology

2.1 Collection of information

The possibility of a Corona Virus pandemic was classified as a health emergency by the World Health Organization. Covid-19 is a free resource provided by researchers and medical facilities. We made use of the university repository's covid-19 dataset from John Hopkin. Total number of infected person, confirmed, date, place, Last status, , deaths, and Improved are the eight qualities. The data was collected in 2020 among January and December. 17285,456 papers from December 6, 2020, to December 6, 2020, are included in the collection. The number of identified cases projected by the proposed method over the next several days is likened to the actual data from the covid-19 long-established cases to assess the model's accurateness.

2.2 Preliminary processing

The formless manuscript must be enhanced in order to achieve machine learning. The actions performed during this period varied. The text is cleaned up by removing any extraneous stuff. Punctuation and Lemmatization are employed to improve the findings. Stop words, ties, and icons are removed to improve sorting and division accurateness.

2.3 Extraction of feature

Many attributes from pre-processed clinical reports are retrieved and turned into probabilistic values using semantics. To extract critical characteristics, the Pandas package and NumPy are utilised. We picked critical indicators such as complete cases, data-wise cases, mortality, recovered cases, and so on to aid in categorising. The feature was assigned a corresponding weight, and the deep learning algorithms used the same inputs.

2.4 Forecasting

We made extremely precise predictions about the number of persons that will subsequently be universally conforming using the Holt-Winters Exponential Smoothing method. The SVM algorithm and linear regression both fared better than this approach.

2.5 Information used

The COVID-19 dataset includes data on the observation, the nation, the place, the time for that day, and the number of informed cases, saved cases, and demise cases. It was compiled from the John Hopkin's University repository.

Table 1. Sample data obtained for the research

S.No.	Date	Place	Place	Identified	Death rate	Recovered
1	14-08-2020	Yukon	Canada	59	6	46

2	15-08-2020	Yunnan	China	226	7	215
3	16-08-2020	Zabaykalsky krai	Russia	2088	344	18960
4	17-08-2020	Zacatecas	Mexico	18103	1435	5
5	18-08-2020	Zeeland	Netherland	6715	109	5

One of the most important aspects in assessing the proposed job is accuracy. Because of the projected rise in infected patients, the model's predicted data is correct when compared to actual data. The projected output of the model is compared against the actual data acquired from the institution using administered artificial intelligence methods.

Table 2 Total amount of cases

S.No.	Date	Total confirmed cases in the world
1	09-11-2020	6,71,99,374
2	10-11-2020	6,77,67,203
3	11-11-2020	6,84,14,187
4	12-11-2020	60,69,464
5	13-11-2020	7,11,64,745

A popular and simple Machine Learning strategy is linear regression. It serves as both a statistical method and a tool for predictive analysis. The linear regression technique, as its name suggests, shows that a in need of variable and one or more autonomous variables have a linear relationship. Since it shows a linear connection, linear regression is used to determine the modification in the reliant on variable's value as a purpose of the self-governing variable's value.

Using this technique, we classified the dataset into four categories: total identified cases, active cases, mortal cases, and disposed cases.

Table 3 Trial identified cases

S.No.	Dates	Regression model predicted
0	09-11-2020	4,48,86,035
1	10-11-2020	4,50,60,052
2	11-11-2020	4,52,34,069
3	12-11-2020	4,54,08,086

It is also recognized as the most often utilised administered learning approach for dealing with classification and regression problems. However, its primary use in deep learning is for problems. The basic purpose of SVM is to find the finest boundary and line for categorising subsequent data points so that classification can be done rapidly. A hyper plane is a boundary that characterizes the optimum choice.

We separated the information into the four categories described overhead to discover the confirmed cases in SVR. For example, we employed the Support Vector Machine approach on the last five dates. The results of the support vector regression and linear regression methods were compared. They varied greatly from one other and from real-time data collecting. The model is then trained using a novel approach known as the Time Series algorithm (Table 4).

Table 4 Confirmed cases from both methods

S.No.	**Dates**	**Regression Output**	**SVM**
0	09-11-2020	4,48,86,035	2,42,91,026
1	10-11-2020	4,50,60,052	2,45,36,849
2	11-11-2020	4,52,34,069	2,47,85,755
3	12-11-2020	4,54,08,086	2,50,37,772

A time series approach's purpose is to develop a time series model by comprehending the underlying notion of the time series data points or by making suggestions or forecasts. It includes sequential information opinions that are charted across a present period. A important model is used to predict assumptions based on past collected discoveries when predicting or modelling data utilising timeseries research. For example, the number of customers at a restaurant is forecasted based on the hour and previous customer appearances, such as when more customers would visit the establishment at a specific time.

Deep learning, which is famous, is one of the finest approaches for predicting normal linguistic dispensation for a huge dataset. Periodic model challenges, on the other hand, often lack interpreted datasets until data from several sources is gathered and reveals significant gaps in features, traits, qualities, temporal balances, and dimensionality. To examine periodic based, a sorting algorithm that can handle period-dependent patterns using replicas other than sights and sounds must be devised. Machine learning techniques aimed toward practical, commercial applications include prediction, assemblage, anomaly detection, and gathering discovery. Machine learning algorithms are employed in time-series analysis. The Holt Winter exponential smoothing technique outperformed the assist vector machine and regression approaches in terms of accuracy.

Holt-Winters forecasting is a technique for predicting and proud the behaviour of a set of variables across time. The most often used time series prediction techniques is Holt-Winters. Holt-Winters is a period behaviour method. The Holt-Winters method is a strategy for replicating the standard value, slope through period, and recurrent trend of a time series—all three aspects necessary for forecasting. Holt-Winters predicts conventional values for the present and future and uses the exponential smoothing technique to encrypt new historical values.

In the exponential smoothing strategy, the weighted average of all previous values is utilised to estimate the future value, with weights decreasing exponentially from the most recent to the oldest. When you use Exponential Smoothing, you believe that the most recent values of time series models are much more essential than their older counterparts. The exponential Smoothing method has two major drawbacks: it does not account for seasonal swings or patterns in your data. The Holt ES approach addresses the different weaknesses in the standard ES methodology. Holt ES can anticipate the outcomes of trending time series. Holt ES, on the other hand, battled with seasonal fluctuations in the sequencing of events. The Holt-Winters Exponential Smoothing approach correctly forecasted the number of futures globally conforming events.

3. Result and Discussion

The graph below depicts the model's algorithm-based prediction. In this study, Holt's Winter's proposed time series model provides more confirmed occurrences than current approaches like LR and SVR. As a result, the model's algorithm's predictions may or may not be correct. It must consequently be compared with real statistics on the number of afflicted persons. The model may be trained further by utilising new training sets created particularly for the real information (Fig. 2, Table 5).

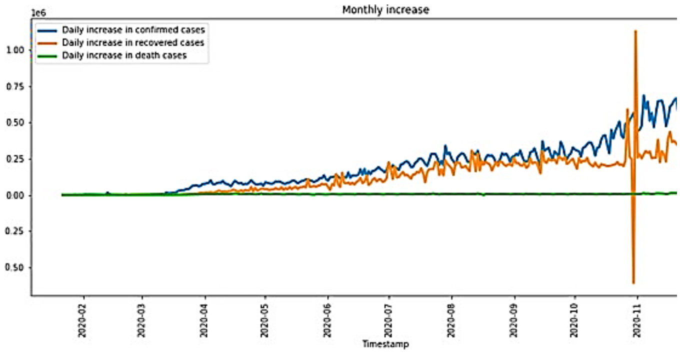

Fig. 2 COVID virus statistics plot

Table 5 Results of forecasted accuracy

S.No.	**Times**	**LR**	**SVR**	**Holts Winter**	**Confirmed cases**
1	09-11-2020	4,48,85,623	2,42,90,616	5,88,50,814	6,71,99,765
2	10-11-2020	4,50,59,640	2,45,36,439	5,92,25,220	6,77,67,594
3	11-11-2020	4,52,33,657	2,47,85,345	5,95,99,625	6,84,14,578

4	12-11-2020	4,54,07,674	25,03,37,362	5,99,74,031	6,90,69,855
5	13-11-2020	4,55,81,691	2,52,92,520	6,03,48,437	7,11,65,136

The findings are presented in the graph below, which compares the trained model's output to real data on the worldwide spread of COVID-19 infections on that specific day. The qualified model used the LR, SVR, and Holt's winter model. In comparison to the other two methodologies used to develop the model, the research found that Holt's linear model estimates future worldwide confirmed cases with an accuracy of roughly 87%. Furthermore, Holt's Winter's time series forecasting model exceeds the competition meanwhile it can forecast future data with high correctness using the given period series COVID - 19 information.

CONCLUSION

The proposed project's primary goal is to analyse worldwide COVID-19 data and predict future numbers of reported cases globally using supervised machine learning methods. On associated to the LR and SVR algorithms, this study established a period series forecasting Holt's winter model with higher accuracy in forecasting future data. Besides, we investigate the curve founded on the disease's monthly worldwide trend and utilise Python libraries to measure and visualise COVID-19's current global trend. Real-time deployment will also be included in future projects.

Internet based analytics might be used to screen the epidemic, forecast its path, and develop regulations and measures to prevent it from dispersal. The ongoing development of machine learning technologies may result in more exact projections of the amount of new persons identified by covid and the end date of the epidemic. We present a mechanism for organizing these models on web-internet for fast and safe computing. In a cloud-based setting, both public and private hospitals often provide positive patient data.

REFERENCES

[1] G. Taheri and M. Habibi, "Comprehensive analysis of pathways in Coronavirus 2019 (COVID-19) using an unsupervised machine learning method," *Applied Soft Computing*, vol. 128, p. 109510, 2022, doi: 10.1016/j.asoc.2022.109510.

[2] S. I. Busari and T. K. Samson, "Modelling and forecasting new cases of Covid-19 in Nigeria: Comparison of regression, ARIMA and machine learning models," *Scientific African*, vol. 18, p. e01404, 2022, doi: 10.1016/j.sciaf.2022.e01404.

[3] N. Leelawat *et al.*, "Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning," *Heliyon*, vol. 8, no. 10, p. e10894, 2022, doi: 10.1016/j.heliyon.2022.e10894.

[4] M. E. Elkin and X. Zhu, "A machine learning study of COVID-19 serology and molecular tests and predictions," *Smart Health*, vol. 26, no. October, p. 100331, 2022, doi: 10.1016/j.smhl.2022.100331.

[5] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, and S. H. Malik, "Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques," *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100120, 2022, doi: 10.1016/j.jjimei.2022.100120.

[6] Z. Wang, "Use of supervised machine learning to detect abuse of COVID-19 related domain names," *Computers and Electrical Engineering*, vol. 100, no. March, p. 107864, 2022, doi: 10.1016/j.compeleceng.2022.107864.

[7] Y. V. Kistenev, D. A. Vrazhnov, E. E. Shnaider, and H. Zuhayri, "Predictive models for COVID-19 detection using routine blood tests and machine learning," *Heliyon*, vol. 8, no. 10, p. e11185, 2022, doi: 10.1016/j.heliyon.2022.e11185.

[8] H. Passarelli-Araujo, H. Passarelli-Araujo, M. R. Urbano, and R. R. Pescim, "Machine learning and comorbidity network analysis for hospitalized patients with COVID-19 in a city in Southern Brazil," *Smart Health*, vol. 26, no. April, p. 100323, 2022, doi: 10.1016/j.smhl.2022.100323.

[9] M. T. Ahemad, M. A. Hameed, and R. Vankdothu, "COVID-19 detection and classification for machine learning methods using human genomic data," *Measurement: Sensors*, vol. 24, no. October, p. 100537, 2022, doi: 10.1016/j.measen.2022.100537.