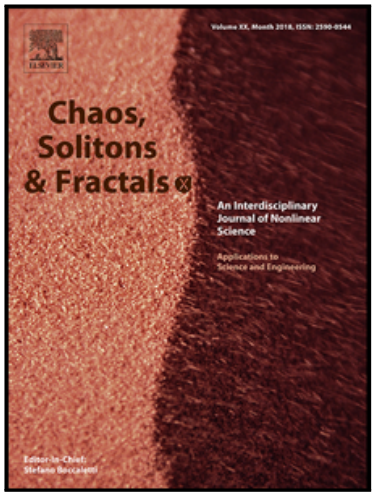
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal Pre-proof



Forecasting spread of COVID-19 using Google Trends: A hybrid GWO-Deep learning approach

Sikakollu Prasanth, Uttam Singh, Arun Kumar, Vinay Anand Tikkiwal, Peter H.J. Chong

PII: S0960-0779(20)30731-1
DOI: https://doi.org/10.1016/j.chaos.2020.110336
Reference: CHAOS 110336

To appear in: *Chaos, Solitons and Fractals*

Received date: 17 August 2020
Accepted date: 1 October 2020

Please cite this article as: Sikakollu Prasanth, Uttam Singh, Arun Kumar, Vinay Anand Tikkiwal, Peter H.J. Chong, Forecasting spread of COVID-19 using Google Trends: A hybrid GWO-Deep learning approach, *Chaos, Solitons and Fractals* (2020), doi: https://doi.org/10.1016/j.chaos.2020.110336

*Highlights (for review)

Forecasting spread of COVID-19 using Google Trends: A hybrid GWO-Deep learning approach

1. Utilized ECDC data + Google trend term frequency to forecast the spread of COVID-19 in different regions.
2. Used Spearmann correlation to select the effective COVID related search terms .
3. Proposed a novel technique based on meta-heuristic GWO algorithm to optimize hyperparameters for LSTM network.

Forecasting spread of COVID-19 using Google Trends: A hybrid GWO-Deep learning approach⋆

Sikakollu Prasanth[a], Uttam Singh[a], Arun Kumar[a,*], Vinay Anand Tikkiwal[b] and Peter H J Chong[c]

[a]National Institute of Technology, Rourkela, India, 769008
[b]Jaypee Institute of Information Technology, Noida, India - 201304
[c]Department of Electrical and Electronic Engineering, Auckland University of Technology, New Zealand - 1010

ARTICLE INFO

Keywords:
COVID-19
Forecasting
Long Short Term Memory (LSTM)
Deep Learning
Pandemic
Grey Wolf Optimization (GWO)
Google Trends
Optimization
Auto Regressive Integrated Moving Average (ARIMA).

ABSTRACT

The recent outbreak of COVID-19 has brought the entire world to a standstill. The rapid pace at which the virus has spread across the world is unprecedented. The sheer number of infected cases and fatalities in such a short period of time has overwhelmed medical facilities across the globe. The rapid pace of the spread of the novel coronavirus makes it imperative that its' spread be forecasted well in advance in order to plan for eventualities. An accurate early forecasting of the number of cases would certainly assist governments and various other organizations to strategize and prepare for the newly infected cases, well in advance. In this work, a novel method of forecasting the future cases of infection, based on the study of data mined from the internet search terms of people in the affected region, is proposed. The study utilizes relevant Google Trends of specific search terms related to COVID-19 pandemic along with European Centre for Disease prevention and Control (ECDC) data on COVID-19 spread, to forecast the future trends of daily new cases, cumulative cases and deaths for India, USA and UK. For this purpose, a hybrid GWO-LSTM model is developed, where the network parameters of Long Short Term Memory (LSTM) network are optimized using Grey Wolf Optimizer (GWO). The results of the proposed model are compared with the baseline models including Auto Regressive Integrated Moving Average (ARIMA), and it is observed that the proposed model achieves much better results in forecasting the future trends of the spread of infection. Using the proposed hybrid GWO-LSTM model incorporating online big data from Google Trends, a reduction in Mean Absolute Percentage Error (MAPE) values for forecasting results to the extent of about 98% have been observed. Further, reduction in MAPE by 74% for models incorporating Google Trends was observed, thus, confirming the efficacy of utilizing public sentiments in terms of search frequencies of relevant terms online, in forecasting pandemic numbers.

1. Introduction

The recent outbreak of coronavirus, popularly known as COVID-19, took place in Wuhan city of Hubei province in China [1]. Though the first case of coronavirus was reported in December, several countries started reporting the cases since late January. Since the details of this virus were not known initially, the spread has been very fast. COVID-19 has emerged as a global pandemic as declared by the World Health Organization (WHO). Initially, it started as an unknown case of pneumonia, the virus has spread to more than 150 countries in a short span of time. This virus has infected almost 16.5 million people [2] in the world.

The coronavirus infection has spread quickly all over the world, causing a huge number of deaths. It is an infectious disease that causes severe acute respiratory problems. The infected person shows mild symptoms of cold and fever initially, which worsens as time progresses. Body pain, nausea and high fever are also some of the characteristic symptoms of this infection [3][4]. It affects the people of all

⋆This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
*Corresponding author
prasanth.sikakollu12@gmail.com (S. Prasanth); shivamsinghnitr123@gmail.com (U. Singh); kumararun@nitrkl.ac.in (A. Kumar); vinay.anand@jiit.ac.in (V.A. Tikkiwal); peter.chong@aut.ac.nz (P.H.J. Chong)
ORCID(s):

age groups, the worst affected being the people of higher age groups [5] who are suffering from at least one health ailment or having a history of respiratory disease.

An exponential rise in the number of COVID-19 cases has led to a crisis of shortage of medical equipment and healthcare personnel in many countries. The number of ICU beds, ventilators, PPE kits for doctors and other health-care workers has seen a huge surge in demand. Ramping up the supply of medical equipment in adequate numbers still remains a challenge. The rapid rise in the COVID-19 infection, necessitates early forecasting of the spread in order to assist the governments and local authorities to plan for necessary measures including manpower and medical equipment deployment among others.

The impact and fast spread of Coronavirus have created a general fear among the people all over the world, and there has been a rise in the number of internet searches related to COVID-19. People learning about preventive measures and constantly searching the web for the updates. Referencing to this aspect of human behaviour, different techniques are explored to design the models that learn from the data mined from the search results of the people belonging to a particular region. The intensity of the impact of the virus can be related to the search trends. Google trends, obtained by processing a multiple types of Google search results, reflects the public attention towards a particular search keyword [6] (Li, Ma, Wang, & Zhang, 2015). Google trends represents the volume for a given search term, relative to the total number of searches on Google, on a scale of 0 to 100. Accordingly, Google trends have been widely utilized to be a sort of big data covering large-scale information [7]. Given these implications, this study utilises Google trends as a predictor for COVID-19 cases.

In this work, the future incidences of Coronavirus are forecasted using both European Centre for Disease prevention and Control (ECDC) data [8] and Google Trends (GT) data for three countries, namely USA, India and UK. The impact of people's internet browsing interest on the incidence of pandemic is studied using Google search trends data. Highly correlated Google search terms are selected using Spearman's rank correlation between ECDC COVID-19 trends and GT data.

Several time-series forecasting techniques have been proposed in the literature. Traditional statistical techniques such as Autoregressive Integrated Moving Average (ARIMA) have been commonly used to forecast time-series value of variables across disciplines. However, these methods suffer from poor accuracy as the influence of external factors is not well captured. Recently, deep learning based LSTM has garnered significant attention for time series forecasting of various trends. LSTM has been previously employed to forecast: weather [9], stock price movements [10][11], pandemics [12], solar irradiance [13] [14], atmospheric pollution levels [15]. They have also been employed to predict the answers to questions [16], predicting the next word [17] *etc*. The most important feature of an LSTM network is its capability to find the time-series dependencies. Since the trends of Coronavirus is a time-series data, this work utilizes LSTM-based forecasting model to forecast the future trends.

Hyperparameter-tuning is an important task while designing any neural network-based models. Usually, a huge amount of time is spent on finding the optimal set of hyperparameters manually. In this work, we automate the process of hyperparameter-tuning using a meta-heuristic search algorithm namely, Grey Wolf Optimization (GWO). GWO algorithm has shown success in various optimization tasks such as optimal feature set selection [18], node localization problem in wireless networks [19], Kernel ELM parameter tuning for bankruptcy prediction [20], etc. Also, GWO algorithm is proven to be superior [21] when compared to other meta-heuristic algorithms like GA, PSO, GSA, grid search, etc. GWO algorithm, being simple, robust, and flexible, is proposed for hyperparameter tuning of LSTM networks.

The rest of the paper is organized as follows - Section 2 presents the related work done on forecasting the outbreak and spreading of epidemics. Section 3 provides a description of the datasets used in this study.The proposed workflow and methods adopted are presented in Section 4. Section 5 illustrates the experimental designs and the results obtained. A detailed comparative analysis of the results is presented in Section 6, while, Section 7 concludes the work presented in the paper.

2. Related Work

The recent outbreak of COVID-19, which has spread at an unprecedented rate; has led to a lot of research being conducted to predict the spread. Domenico et al. [22] proposed the use of ARIMA on the Johns Hopkins data to predict the epidemiological trend of the incidences of COVID-19. Singh et al.[23] employed ARIMA to forecast COVID-19 related confirmed cases, deaths, and recoveries.The ARIMA model was validated using the AIC value; which were around 20, 14, and 16 for cumulative confirmed cases, deaths, and recoveries from COVID-19, respectively. Ceylan et al. [24] in their work has exploited an ARIMA-based framework for predicting the coronavirus trends. The study carried out for European countries like Spain, Italy and France. Kumar and Hembram [25] proposed a model which used Logistic equation, Weibull and, Hill equation to find infection rates and obtained the power index of top ten highly infected countries. A recent work on coronavirus [26] proposes the use of supervised machine learning models in their research to forecast the coronavirus trends. The researchers used four standard forecasting models, namely Linear Regression, Support Vector Machines and, Exponential Smoothing etc. to forecast the trends of recovery rate, deaths and daily new cases due to coronavirus. The best Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values obtained for new cases and recovery rate are 8867.43 and 15322.11 and 1827.85 and 2443.48 respectively. The research carried by Sahai et.al.,[27] has used the ARIMA model to forecast the spread of coronavirus in top five affected countries. They have used hannan and Rissenan algorithm to find the parameters of ARIMA model. Their study forecast that 1.39 , 2.47 and 4.31 million people will be affected in India, Brazil and United States of America (USA) respectively. Chintalapudi et.al.,[28] have used seasonal ARIMA to forecast the registered and recovered cases after sixty days of lockdown in Italy. An accuracy of 93.75% and 84.4% have been recorded for the registered and recovered cases respectively. A reduced space guassian process regression method has been used in [29] to forecast the spread of coronavirus in USA. They forecasted the peak in cases will occur on 14 th July for USA.

Vinay et al. [30] presented an LSTM framework for predicting the number of coronavirus cases in Canada. The authors have also compared the results of Canada with transmission rates of USA and Italy. Tomar and Gupta [31] used LSTM to forecast number of recovered cases, daily positive cases, deceased cases for India, thirty days in advance. The study also reported the effectiveness of preventive measures like social isolation and lockdown on the spread of COVID-19. Ibrahim et al. [32] proposed a variational LSTM autoencoder model to predict the global trends of coronavirus. The authors have not only used the historical data of the cases trends but also made use of some urban characteristics and government response to the virus; including, closing of workplaces and schools, cancellation of public events and, closure of public transport etc. These features, along with the COVID cases data, were used to forecast the future incidence of coronavirus. The RMSE values obtained for the prediction results were 12722.61, 2712.82 and 271.38, respectively for USA, UK and India.

In [33], authors present an infectious disease prediction model using different input variables, selected based on the OLS method. The authors compared the performance of three models, i.e., ARIMA, DNN and LSTM trained with optimal parameters to predict the future trends of three infectious diseases, Chickenpox, malaria and Scarlet fever. The results demonstrated that the neural network models gave more accurate prediction when compared to ARIMA. For Chickenpox, the DNN and LSTM models have reported an increase of 24% and 19% in average performance, respectively.

Information extracted from social media platforms such as Twitter, Google Trends, blogs etc. can prove to be useful sources of information for forecasting pandemics. Previously, Lampos et al.[34] have used thousands of tweets related to the flu disease and predicted a flu-score. This research carried out in the UK for the H1N1 virus. The work has shown a 95 % correlation with the data from the health agency. Similar work has been presented by Signorini et al. [35], for tracking the H1N1 virus spread in USA . The authors developed a SVM based model using the influenza data and data from tweets regarding the disease to predict the spread of the virus. The research provides a window of dates when the infection would obtain a peak in the number of infected cases. Anggraeni and Aristiani [36], stud-

ied the usage of GT data in forecasting the dengue fever in Indonesia. The research used the data from local hospitals on the number of cases of dengue fever and google search index to forecast the new cases using ARIMA. The model with ARIMA using Google Trends achieved better accuracy than the normal ARIMA with 3% decrease in MAPE value. Teng et al. [37] carried out experiments to dynamically forecast the spread of Zika virus epidemic using GT data. The authors in this work exploited ARIMA regressor with GT data as an external regressor to improve the prediction. Effenberger et al. [38] carried out a correlation-based study for the new cases data and 'Coronavirus' search term GT data. This work emphasized the increase in the search volume about the virus with the increase in the number of cases, which the authors inferred could be a result of public panicking in the face of the fast-spreading pandemic.

Previously conducted research work indicate that Google Trends may play an important factor in terms of forecasting a pandemic. However, it is important to identify more relevant search terms with respect to the forecasted variable, during the period of spread of a pandemic. In this research, the main objective is to use the Google Trends search frequencies of more significant terms related to COVID-19 for effectively forecasting the incidences of infection spread using a GWO optimized LSTM model. Implementation of metaheuristics for improved LSTM model through hyperparameter tuning in the pandemic domain has hardly been investigated. To the best of our knowledge, this is the first study to forecast the spread of COVID-19 through the use of optimized LSTM networks incorporating GT data.

3. Dataset Description

3.1. ECDC data for different Coronavirus trends

Since the outbreak of coronavirus and its spread to various countries, many different organizations including ECDC has maintained a count of the total number of infections, new daily infections, total deaths etc., due to coronavirus, to keep a track of the spread of the epidemic. This data is available country-wise and region-wise. For this study, the data between February 24, 2020 to May 20, 2020 for India, USA and United Kingdom (UK) has been taken into consideration [8]. This dataset contains Total Cumulative cases (TCC), New cases (NC), Total Cumulative Deaths (TCD) for the three countries.

3.2. Google Trends data

During the outbreak of COVID-19, people all over the world began searching for different terms related to the pandemic like COVID-19, coronavirus, symptoms, sanitizer, etc. The Google Trends (GT) data for a given search term represent the interest of the people for that particular search term on google search. It is represented using percentage values relative to the time period considered. GT data for search terms are good indicators for forecasting the future coronavirus trends. We select the countries that appear in the list of worsely affected countries due to the virus. USA is the worst affected country while UK has a large number of deaths. India has shown an exponential rise in cases in a span of few weeks. Hence these three countries have been selected for our research. The duration of the research starts from 24 th February as some of the countries have reported the infection lately. For example, the first confirmed case in India was reported on 30 th January 2020, as mentioned in ECDC dataset [8] and the next confirmed cases were reported after one week. A duration with zero cases or deaths would result in loss in accuracy of the model. Hence the duration was chosen to have finite values.

In this work, nine search terms related to coronavirus are considered that are mostly searched across the globe. These search terms were obtained by comparing the terms using the compare function in the google trends [7], [38]. This allows to mine the data which have significant search frequency. This research deals with the search terms. GT data [39] for the nine search terms are downloaded for three different countries, India, USA and UK for a duration of three months, i.e., from February 24, 2020 to May 20, 2020. The search terms are mentioned in Table 1.

4. Proposed Methodology

The problem of forecasting future trends is formulated as a three-step procedure. The three steps include relationship investigation, forecasting model and optimizing the hyperparameters. In step 1, the

Table 1
Google Trends search terms

S.No.	*Search Terms*
1.	Coronavirus symptoms
2.	Coronavirus
3.	Covid
4.	Handwash
5.	Healthcenter
6.	Mask
7.	Positive cases
8.	Sanitizer
9.	Coronavirus Vaccine

relationship between GT data and ECDC data is evaluated. Next, the top two search terms that are having the highest relationship are selected, and only those are considered for further experiments as input data. In step 2, the models are trained with ECDC data and GT data to forecast future coronavirus trends. In step 3, a nature-inspired metaheuristic algorithm, i.e., GWO algorithm is applied to find the optimal set of hyperparameters (window size, number of hidden layers and number of cells in each layer) to be considered for forecasting COVID-19 data. The complete proposed workflow is shown in Figure 1, and the three steps are explained below in detail.

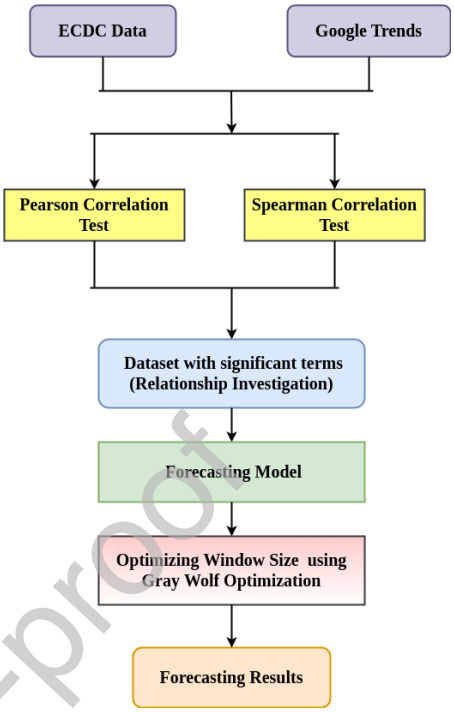


Figure 1: Proposed workflow for forecasting future COVID-19 trends

4.1. Relationship investigation and Feature selection

In this step, the relationship between GT data $(Y_1, Y_2, ..Y_n)$ and the ECDC data (TCC, NC, TCD) is computed using a correlation-based method. To use the impact of the internet searching in forecasting the trends of coronavirus, we make use of GT data of 90 days. Not all the search terms follow the same trend as that of coronavirus cases in a region. To find the optimal search terms, the correlation between the search term frequencies and the history of the cases are calculated. The terms with greater correlation values help for better forecasting. The coronavirus cases for TCC, NC, TCD have been collected for 90 days as shown for TCC = $\{x_1, x_2..., x_{90}\}$. For any search term Y_i, the data consist of normalized search frequencies $Y_i = \{y_1, y_2..., y_{90}\}$. To obtain the correlation between the GT data and the ECDC data, two correlation coefficients are used, namely Pearson correlation and Spearman Correlation. The Pearson correlation coefficient between two vectors, x and y, is calculated after subtracting the mean of the corresponding data series. It can be viewed as the dot product of two mean subtracted vectors. The equation for computing the Pearson correlation is given in Eq. 1.

$$\gamma = \frac{\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \cdot \sqrt{\sum_i (y_i - \bar{y})^2}} \quad (1)$$

The drawback of Pearson correlation is that it can identify only the linear relationship between the two variables. But most of the data in the real world do not follow a straight line equation. Spearman rank correlation can efficiently identify the non-linear relationship between the variables. Thus, the Spearman rank correlation is employed to derive the relation between the two variables.

Spearman Correlation organises the data by as-

signing ranks (rg) to them. For two variables to have a maximum correlation, it is very important that the difference between the ranks of each data point is very minimal. Spearman correlation is also defined as the Pearson correlation between ranks of each data point belonging to two different data series. The formula for calculating the Spearman correlation coefficient is shown in Eq. 2.

$$\rho = \frac{cov(rg_x, rg_y)}{\sigma_{rgx} \cdot \sigma_{rgy}} \tag{2}$$

Both Pearson and Spearman correlation coefficients value ranges from -1 to $+1$. For the creation of input dataset, the top two search terms with the highest correlation coefficient values are selected for further experiments.

The plots of the top two highly correlated Google Trend search terms and NC for the three countries are shown in Fig. 2. The graphs are plotted on daily basis with NC on the primary Y-axis and normalized search frequency of the Google Trend terms on secondary Y-axis. NC and search frequency can be visualized of having a non-linear correlation. Hence these plots support the idea of using a correlation which takes the non-linear relationship of both the trends.

4.2. Forecasting Methods

In this step, the forecasting models are designed using different inputs and forecasting techniques. This research work uses two time-series forecasting techniques, namely ARIMA and LSTM. The working of the two techniques is described below in detail.

4.2.1. ARIMA

ARIMA [40]-[41] is a class of regression models used for forecasting time-series data. It predicts the time-series values based on its past values, i.e., its own lags and forecast errors. An ARIMA model is characterised by three parameters, (p, d, q). 'p' stands for the order of the Auto-Regressive (AR) term. It describes the number of lags to be considered for forecasting. 'q' stands for the order of the Moving Average (MA) term. It describes the number of forecasting error lags needed for prediction. 'd' stands for the number of differencing needed to make the data stationary. To apply the ARIMA model, the input data should be stationary. Data is made stationary by subtracting the previous value from the current value. Depending on the data, multiple differencing is needed to make the data stationary. The predicted value is a function of the 'p' lag terms, 'q' forecasting error lags and the constant term. The working equation of the ARIMA model is given in Eq. 3.

$$\begin{aligned} Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \\ \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + ... + \phi_q \epsilon_{t-q} \end{aligned} \tag{3}$$

where Y_t represents the value of the time-series data at time t, $\beta_1, \beta_2, .., \beta_p$ represents the coefficients of the previous p time step terms, ϵ_t represents the forecasting error at time t, $\phi_1, \phi_2, .., \phi_q$ represents the corresponding coefficients of error terms, and α corresponds to the constant term.

4.2.2. LSTM network

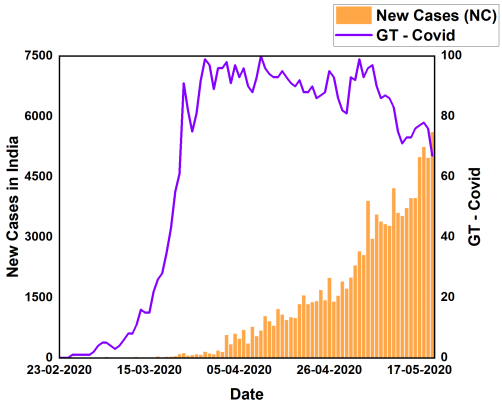
Long Short-Term Memory (LSTM) [42] [43] networks are a kind of recurrent neural networks that are used for forecasting time series data. LSTMs overcome the drawbacks of the vanilla RNN, i.e., the problems of exploding and diminishing gradients, by having memory. Therefore it is well suited to learn from important experiences that have very long time lags in between. The units of an LSTM are used as the building units for the layers of the network.

LSTM cells enable the network to remember their inputs over a long period of time. This is because LSTMâĂŹs contain their information in a memory, that is much like the memory of a computer because the LSTM can read, write and delete information from its memory. This scheme is implemented using three gates; namely, input, forget and output gates. The working equations of LSTM cell are shown in Eq. 4 - 9.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{4}$$

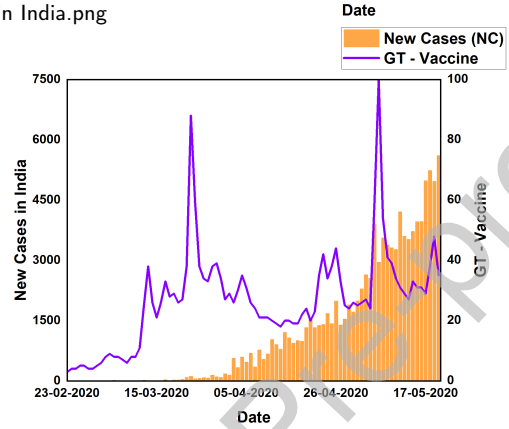
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{5}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{6}$$

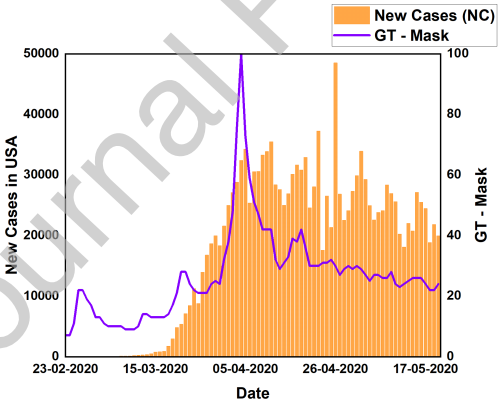


vs covidGT in India.png

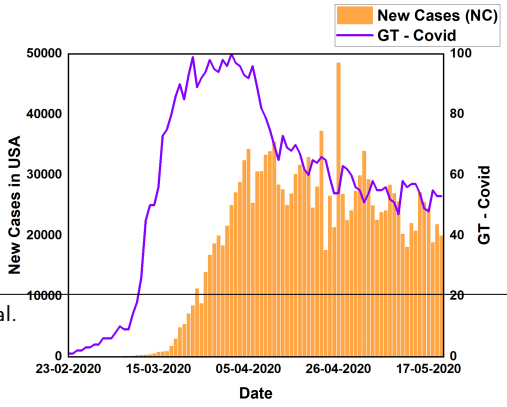
vs vaccineGT in India.png



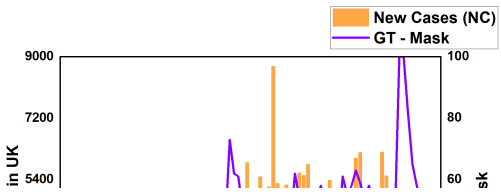
vs maskGT in USA.png



vs covidGT in USA.png



vs maskGT in UK.png



$$\tilde{c}_t = tanh(W_c[h_{t-1}, x_t] + b_c) \quad (7)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (8)$$

$$h_t = o_t \odot tanh(c_t) \quad (9)$$

where f_t is the output of forget gate for time t, h_{t-1} output of the hidden state vector form time t-1, i_t represents the input at state t, o_t represents the output from the state t, c_t represents the output from the cell unit, W_i , W_c , W_f and W_o are the weights associated with the input, cell unit, forget and output gates and b_i, b_c, b_f and b_o are the bias associated with input, cell unit, forget and output gates, σ represents the sigma activation function and $\odot$ represents the Hadamard operation on two matrices. The architecture of the LSTM cell with all the gates and variables is shown in the Fig.3

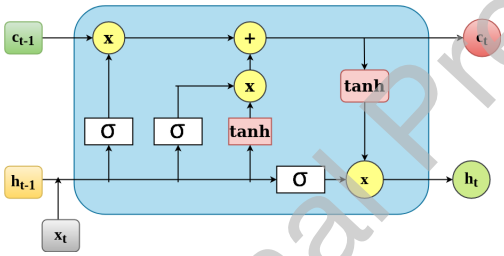


Figure 3: Architecture of an LSTM cell

4.3. Optimizing LSTM parameters using GWO

Grey Wolf Optimization [44] is a nature-inspired metaheuristic search algorithm that enables to find the optimal solution from the solution space in an efficient manner. This algorithm is used to find the optimal hyperparameters namely, window size, number of cells in layers and the number of hidden layers to be considered for LSTM network. This algorithm imitates the social behaviour and hunting mechanism of grey wolves to find the exact location of the prey. There are four kinds of grey wolves in a pack - alpha, beta, delta and omega. Alpha wolves are the most dominant wolves in the group, and these are few in number(usually one or two). They lead the hunt and are responsible for decision-making. Beta wolves assist the alpha wolves in decision making and other activities. These are more in number than alpha wolves but less than delta and omega wolves. Alpha and Beta wolves are the most-experienced wolves in the group. Delta wolves assist the alpha and beta wolves but dominate the omega wolves. Omega wolves, the least dominant category, are mostly baby-sitters. The three phases of the grey wolf hunting are: (i) search and approach the prey, (ii) encircle and make the prey not to move and (iii) attacking the prey.

To model the social behaviour of the grey wolves mathematically, the fittest candidate in the population(of size N) is considered as alpha(α), second and third fittest candidates are considered as beta(β) and delta(δ) respectively. The other candidate solutions are considered as omega(ω). The hunting procedure is guided by α, β *and* δ. To mathematically encircle the prey, the following equations are used.

$$\vec{L} = \mid \vec{K} \cdot \vec{X_P}(t) - \vec{X}(t) \mid \quad (10)$$

$$\vec{X}(t+1) = \vec{X_p}(t) - \vec{H} \cdot \vec{L} \quad (11)$$

where t indicates the current iteration number, $\vec{H}$ and $\vec{K}$ are the coefficient vectors, $\vec{X_P}$ denotes the position of the prey and $\vec{X}$ denotes the position of the grey wolf (candidate). The vectors, $\vec{H}$ and $\vec{K}$ are given by

$$\vec{H} = 2 \cdot \vec{h} \cdot \vec{r_1} - \vec{h} \quad (12)$$

$$\vec{K} = 2 \cdot \vec{r_2} \quad (13)$$

While hunting, it is assumed that the alpha, beta and gamma candidates have a better idea of the location of the prey and they guide the entire search operation towards the optimal solution. During each iteration, the positions of the candidates are updated based on the positions of the top three candidates. If the updated values are outside the solution space, i.e., if the window size is updated to a negative value,

those values are updated according to the Evolution scheme as given in [45]. The formulae for updating the positions of the wolves are given below.

$$\vec{L}_\alpha = |\vec{K}_1 \cdot \vec{X}_\alpha - \vec{X}| \quad (14)$$

$$\vec{L}_\beta = |\vec{K}_2 \cdot \vec{X}_\beta - \vec{X}| \quad (15)$$

$$\vec{L}_\delta = |\vec{K}_3 \cdot \vec{X}_\delta - \vec{X}| \quad (16)$$

$$\vec{X}_1 = \vec{X}_\alpha - \vec{H}_1 \cdot (\vec{L}_\alpha) \quad (17)$$

$$\vec{X}_2 = \vec{X}_\beta - \vec{H}_2 \cdot (\vec{L}_\beta) \quad (18)$$

$$\vec{X}_3 = \vec{X}_\delta - \vec{H}_3 \cdot (\vec{L}_\delta) \quad (19)$$

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \quad (20)$$

The values of r_1, r_2 are randomly chosen in the range (0,1). These allow the wolves to reach any position around the prey. h is chosen in the range [0,2], and H takes the values in the range $[-h, h]$. When $|H| < 1$, it allows the wolves to exploit the solution space, meaning that it gets close to the prey. When $|H| \geq 1$, it allows the wolves to explore the solution space, meaning that it moves away from the prey, enabling to explore the search space. H and K also allow the wolves to come out of the local minima or maxima. Finally, at the end of the last iteration, the fittest candidate, α is returned as the optimal solution.

5. Experimental Study

In this section, the correlation between ECDC data and GT data is investigated. Different experiments are conducted to compute Pearson and Spearman correlation coefficients between each of the search terms' GT and ECDC data (TCC, NC and TCD). Based upon the correlation values, the top 2 search terms having the highest correlation values have been used for forecasting the trends. Next, four different forecasting models are designed, based on different techniques and inputs. The effect of ECDC and GT data on forecasting COVID-19 trends can be studied using these models. The description and implementation details of these four models are described in Section 5.2.

5.1. Feature Selection

Google search trends often show what is about to come in future. But not all related search terms convey the future trend. Therefore, the selection of features (search terms) is essential for any future trend forecasting. The Pearson and Spearman correlation values of different search terms with TCC, NC and TCD are computed for the three countries. The correlation values are shown in Tables 2, 3, 4 for India, USA and UK respectively. The best features are obtained among the list of search terms by selecting the top two features having the highest correlation values.

From the correlation values, it is observed that for India, the search terms having highest correlation values with TCC, NC and TCD are "covid" and "coronavirus vaccine". Therefore, those two search terms are selected along with ECDC data as input to the forecasting models for India. Similarly, "mask" and "covid" search terms data are selected for the USA, and "mask" and "coronavirus vaccine" search terms data are selected for the UK as input features.

It is also observed that many search terms have a negative correlation with ECDC data. It means that these search terms are inversely associated with ECDC data. This inverse relationship is mainly due to the fact that as time progresses, the interest in searching these terms on the internet falls drastically. For e.g., the interest of the search term, 'sanitizer' decreased as the awareness and importance of applying hand sanitizer and sanitizing commonly used surfaces is known to people, and hence resulting in negative correlation coefficients.

5.2. Designing Experimental models

This section discusses the four different models used for forecasting the coronavirus trends. These four models use different combinations of models and inputs required for the forecast. The selected features, along with the ECDC data, are given as input to different models. These include ARIMA with ECDC data as input, LSTM with ECDC data as input, LSTM with both GT and ECDC data as input, and LSTM with both GT and ECDC data as input using optimized hyperparameters using GWO. The implementation details for each of these four combinations are described in subsections below.

Table 2
Correlation values between Google Trends data and ECDC data for India

S.No.	Search Term	TCC		NC		TCD	
		Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
1.	Sanitizer	−0.207	−0.041	−0.247	−0.055	−0.282	−0.136
2.	Coronavirus	−0.252	0.174	−0.206	0.166	−0.184	0.127
3.	Covid	0.374	0.550	0.448	0.539	0.498	0.539
4.	Health center	0.037	0.071	0.026	0.054	0.058	0.071
5.	Mask	−0.032	0.020	−0.070	0.018	−0.112	−0.056
6.	Handwash	−0.155	0.026	−0.175	0.017	−0.184	−0.052
7.	Coronavirus Symptoms	−0.077	0.292	−0.034	0.294	−0.018	0.266
8.	Positive Cases	−0.475	−0.488	−0.519	−0.489	−0.547	−0.562
9.	Coronavirus Vaccine	0.362	0.566	0.370	0.545	0.370	0.508

Table 3
Correlation values between Google Trends data and ECDC data for USA

S.No.	Search Term	TCC		NC		TCD	
		Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
1.	Sanitizer	−0.554	−0.463	−0.415	−0.071	−0.540	−0.463
2.	Coronavirus	−0.593	−0.443	−0.164	0.049	−0.616	−0.444
3.	Covid	0.136	0.181	0.653	0.605	0.058	0.181
4.	Health center	−0.347	−0.383	−0.420	−0.364	−0.308	−0.384
5.	Mask	0.234	0.541	0.683	0.851	0.143	0.540
6.	Handwash	−0.221	−0.176	−0.066	−0.060	−0.232	−0.176
7.	Coronavirus Symptoms	−0.390	−0.033	0.083	0.318	−0.427	−0.033
8.	Positive Cases	−0.632	−0.771	−0.588	−0.404	−0.597	−0.772
9.	Coronavirus Vaccine	−0.047	0.179	−0.045	0.260	−0.039	0.178

5.2.1. Forecasting COVID-19 trends using ARIMA (ECDC-A)

For the ARIMA model (**ECDC-A**), the previous coronavirus trends data is given as input. The

Table 4
Correlation values between Google Trends data and ECDC data for UK

S.No.	Search Term	TCC		NC		TCD	
		Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
1.	Sanitizer	−0.598	−0.652	−0.650	−0.455	−0.599	−0.666
2.	Coronavirus	−0.559	−0.358	−0.193	0.060	−0.568	−0.362
3.	Covid	0.213	0.263	0.50	0.386	0.193	0.262
4.	Health center	−0.347	−0.340	−0.309	−0.249	−0.350	−0.343
5.	Mask	0.583	0.698	0.502	0.604	0.574	0.697
6.	Handwash	−0.622	−0.551	−0.538	−0.323	−0.629	−0.563
7.	Coronavirus Symptoms	−0.214	0.017	−0.038	0.080	−0.231	0.015
8.	Positive Cases	−0.711	−0.832	−0.639	−0.513	−0.719	−0.837
9.	Coronavirus Vaccine	0.318	0.471	0.280	0.377	0.325	0.470

optimal combination of p, d, q is found out for each country and for each parameter forecasted using 'auto_arima' module of 'pmdarima' library in python. It tries to find the optimal combination suitable for the input data to obtain the best forecast. Next, the ARIMA model is trained, as mentioned in Section 4 using optimal parameters to obtain the predictions.

5.2.2. Forecasting COVID-19 trends using LSTM model (ECDC-L)

For the second model (**ECDC-L**), only coronavirus trends data (X) is given as input to the LSTM model, i.e., the input data has only one feature. 'Keras' library with 'Tensorflow' backend are used to implement the LSTM model in Python. LSTM model has two LSTM-cell layers, each layer having 128 cells. The window size of 14 days is used for this model, with 'Adam' optimizer and learning rate of 0.001. For updating the weights, 'MSLE' loss function is used. A dropout of 0.4 and recurrent dropout of 0.2 is also used to prevent overfitting of the model. If z is considered to be the forecasted output, then the mathematical representation of the input and output is given as Eq. 21.

$$z = f(X) \tag{21}$$

5.2.3. Forecasting COVID-19 trends with ECDC and GT data using LSTM (ECDC-GT-L)

In this model (**ECDC-GT-L**), both coronavirus trends data (X) and the top two google search trends (Y_1, Y_2) (discussed in section 5.1) data is fed as input to the LSTM model, i.e., the input data has three features. The rest of the hyperparameters used for LSTM model are as described in section 5.2.2 above. By running this model, the improvement in the performance of the model due to incorporating GT data is found out. The mathematical formulation of the inputs and output (z) of the forecasting model is given in Eq. 22.

$$z = f(X, Y_1, Y_2) \tag{22}$$

5.2.4. Forecasting COVID-19 trends with ECDC and GT data using optimized LSTM (ECDC-GT-GWO-L)

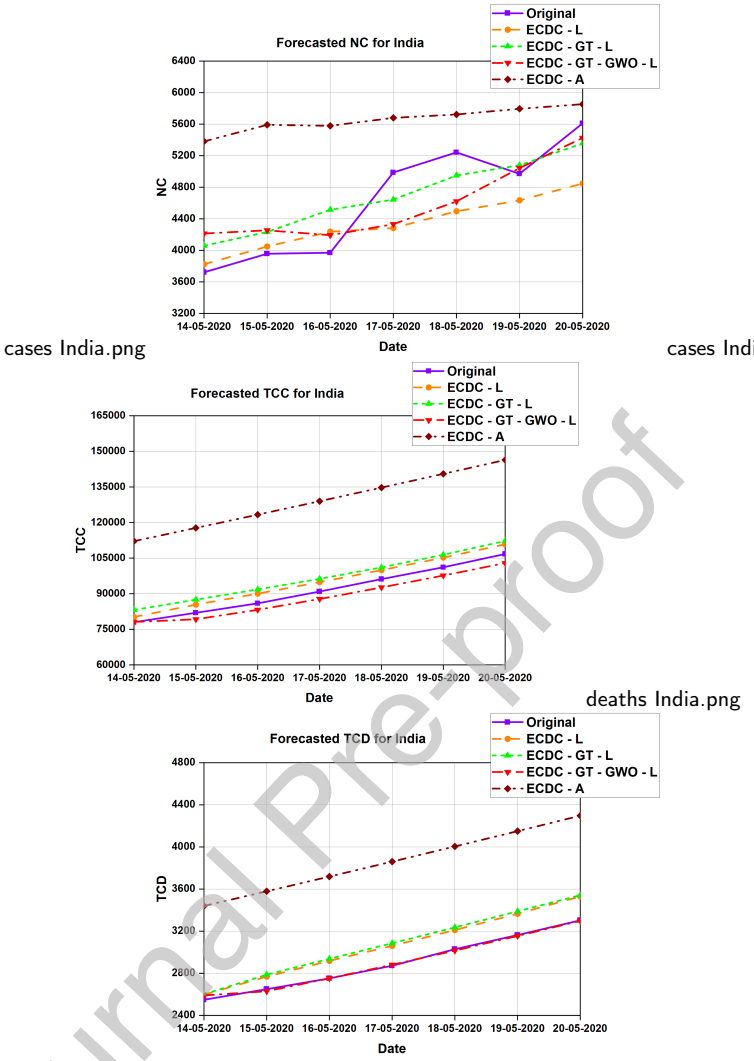
This is the proposed model (**ECDC-GT-GWO-L**) where the window size, number of units and number of hidden layers in the LSTM model are optimized using GWO algorithm. The input data remains the same, as the third model. Using different window sizes gave a lot of variation in the forecasting results. Therefore, the window size is optimized in the range 1 to 28 both inclusive, i.e., a maximum of four weeks to obtain the best forecasting. The number of iterations and population size in each iteration are chosen as 25 and 10, respectively. RMSE is used as the fitness function, and the fitness function is minimized to obtain the optimal set of hyperparameters. However, the other hyperparameters like the number of hidden layers and the number of LSTM cells in each layer didn't result in many variations in the metrics.

6. Results & Discussion

In this section, metrics used for the evaluation of the effectiveness of different techniques are described in detail. Experiments are conducted using different models described in Section 5.2 and the performance of these techniques are compared using nRMSE and MAPE metrics. The idea of taking the internet browsing data related to COVID-19 along with ECDC data is validated for different countries using plots and tables to forecast the cases.

6.1. Evaluation Metrics

The choice of evaluation metrics plays an important role in judging the performance of a model in a correct manner. Most popularly used metrics for evaluating time-series forecasting models are Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) [46]. In this paper, a variant of RMSE, i.e., normalised RMSE (nRMSE) is used along with MAPE to compare the performance of different models. nRMSE is more useful than RMSE as normalizing RMSE makes the comparison scale-free. Since the value of the RMSE is dependent upon the predicted values, the error in case of TCC will always be greater than NC and TCD because of the large TCC values. Normalizing the error values helps in better comparison of the model overall the trends. nRMSE is computed using the mean of the forecasted values. The two metrics are computed according to Eq. 23 and Eq. 24, respectively.



cases India.png cases India.png deaths India.png

Figure 4: Comparison of performance of different models for forecasting different parameters in India (a) Daily New cases (b) Total Cumulative cases (c) Total Deaths

$$nRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}}{\bar{y}} \quad (23)$$

$$MAPE = \frac{\sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{y_i}}{N} \times 100 \quad (24)$$

6.2. Discussion on Forecasted Results

A comparative study of LSTM models and regressors has been made to validate the effective-

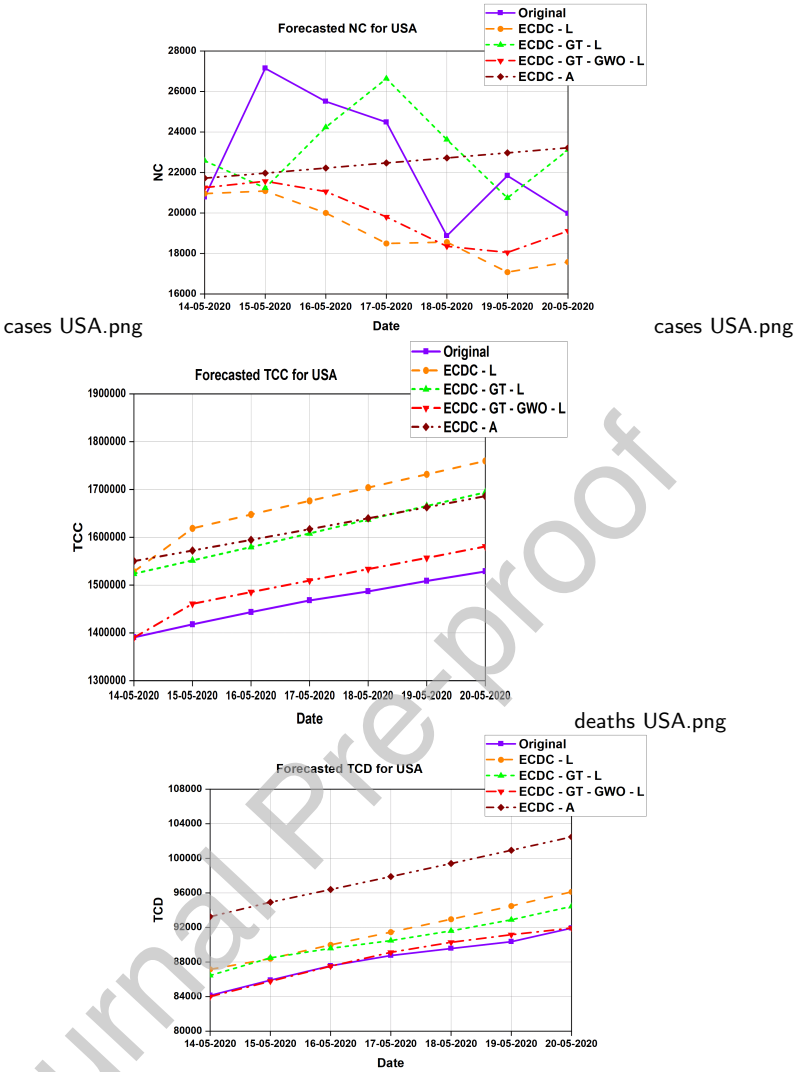
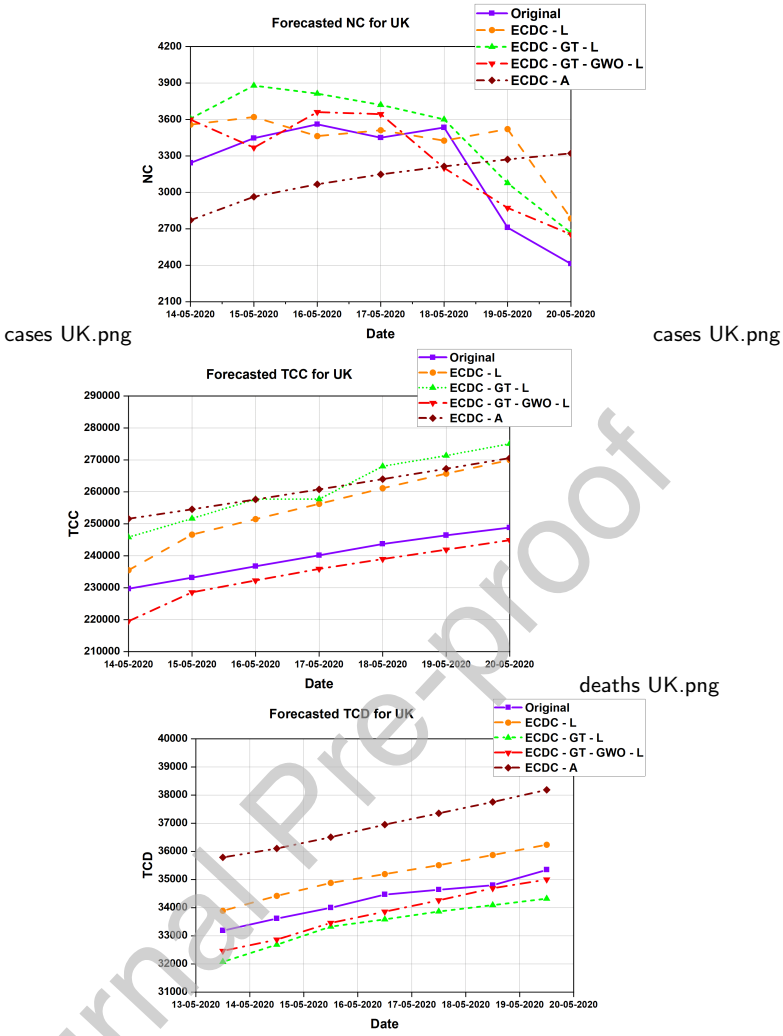


Figure 5: Comparison of performance of different models for forecasting different parameters in USA (a) Daily New cases (b) Total Cumulative cases (c) Total Deaths

ness of the proposed approach. The four experimental models are trained with eighty three days data i.e from 24^{th} February 2020 to 13^{th} May 2020,and tested with data points between 14^{th} May to 20^{th} May 2020. The forecasting performance of the models are compared, and the plots are shown in Figures 4, 5, 6 for India, USA and UK, respectively. The nRMSE and MAPE metrics are computed for each trend predicted for the three countries, and the values are shown in Tables 5, 6, 7 for the three countries, respectively.

For India, the ECDC-A gives least errors when the p, d, q values are 2, 2, 0 respectively. The input to ECDC-A model was the ECDC Coronavirus trend data. Using ECDC-A model, the values of nRMSE and MAPE obtained were 0.485 and 41.698



cases UK.png cases UK.png deaths UK.png

Figure 6: Comparison of performance of different models for forecasting different parameters in United Kingdom (a) Daily New cases (b) Total Cumulative cases (c) Total Deaths

for forecasted TCC, 0.274 and 22.591 for forecasted NC and 0.3098 and 27.535 for forecasted TCD.

For ECDC-L, the values of nRMSE and MAPE obtained were 0.035 and 14.161 for TCC, 0.14 and 20.923 for NC and 0.074 and 13.752 for the case of TCD. This model performed better when compared to ECDC-A model proving the effectiveness of LSTM-based models over the ARIMA regressors.

For the model, ECDC-GT-L, values of nRMSE and MAPE obtained were 0.0165 and 4.1540 when the input was TCC, 0.037 and 9.259 for NC and 0.027 and 6.885 for the case of TCD. This model achieved better results when compared to ECDC-L model indicating the effectiveness of GT data when used along with ECDC data in forecasting results.

For ECDC-GT-GWO-L model, values of nRMSE and MAPE obtained were 0.0136 and 3.452 in the

Table 5
Comparison of RMSE and MAPE values using different models and features for India

S.No.	Trend	Model	Inputs used	nRMSE	MAPE
1.	Total cumulative cases (TCC)	ECDC-A (2, 2, 0)	TCC	0.485	41.698
		ECDC - L	TCC	0.035	14.161
		ECDC - GT - L	TCC, GT - Covid, Vaccine	0.016	4.154
		ECDC - GT - GWO - L	TCC, GT - Covid, Vaccine	0.013	3.452
2.	Daily new cases (NC)	ECDC-A (2, 1, 0)	NC	0.274	22.591
		ECDC - L	NC	0.140	20.923
		ECDC - GT - L	NC, GT - Covid, Vaccine	0.037	9.259
		ECDC - GT- GWO - L	NC, GT - Covid, Vaccine	0.032	7.140
3.	Total cumulative deaths (TCD)	ECDC - A (0, 2, 1)	TCD	0.309	27.535
		ECDC - L	TCD	0.074	13.752
		ECDC - GT - L	TCD, GT - Covid, Vaccine	0.027	6.885
		ECDC - GT - GWO - L	TCD, GT - Covid, Vaccine	0.001	0.304

Table 6
Comparison of RMSE and MAPE values using different models and features for USA

S.No.	Trend	Model	Inputs used	nRMSE	MAPE
1.	Total cumulative cases (TCC)	ECDC - A (0, 2, 1)	TCC	0.109	10.46
		ECDC - L	TCC	0.135	12.914
		ECDC - GT - L	TCC, GT - Covid, Mask	0.011	3.831
		ECDC - GT - GWO - L	TCC, GT - Covid, Mask	0.012	3.132
2.	Daily new cases (NC)	ECDC - A (1, 1, 0)	NC	0.169	15.571
		ECDC - L	NC	0.157	13.262
		ECDC- GT- L	NC, GT - Covid, Mask	0.138	12.637
		ECDC - GT - GWO - L	NC, GT - Covid, Mask	0.132	11.78
3.	Total cumulative deaths (TCD)	ECDC - A (2, 2, 3)	TCD	0.099	9.517
		ECDC - L	TCD	0.250	14.55
		ECDC - GT - L	TCC, GT - Covid, Mask	0.014	3.746
		ECDC - GT - GWO - L	TCD, GT - Covid, Mask	0.009	2.565

case of forecasted TCC, 0.032 and 7.140 for NC and 0.001 and 0.304 for the case of TCD. Similarly for the other two countries, the proposed model surpasses other models. For USA, MAPE values obtained by the proposed model are 3.13, 11.78 and 2.565 for TCC, NC, TCD respectively. For UK, MAPE values obtained using the proposed model are 1.696, 6.946 and 1.443 for TCC, NC, TCD respectively.

The percentage improvements in MAPE values for the proposed model when compared to other experimental models are shown in Tables 8, 9, 10. The

Table 7
Comparison of RMSE and MAPE values using different models and features for UK

S.No.	Trend	Model	Inputs used	nRMSE	MAPE
1.	Total cumulative cases (TCC)	ECDC - A (2, 2, 0)	TCC	0.088	8.808
		ECDC - L	TCC	0.089	8.993
		ECDC - GT -L	TCC, GT - Covid, Vaccine	0.027	7.136
		ECDC - GT - GWO - L	TCC, GT - Covid, Vaccine	0.006	1.695
2.	Daily new cases (NC)	ECDC - A (2, 1, 0)	NC	0.168	16.931
		ECDC - L	NC	0.195	19.632
		ECDC - GT - L	NC, GT - Covid, Vaccine	0.0363	9.236
		ECDC - GT - GWO - L	NC, GT - Covid, Vaccine	0.027	6.945
3.	Total cumulative deaths (TCD)	ECDC - A (0, 2, 1)	TCD	0.077	7.749
		ECDC - L	TCD	0.058	3.12
		ECDC - GT - L	TCD, GT - Covid, Vaccine	0.025	2.484
		ECDC - GT - GWO - L	TCD, GT - Covid, Vaccine	0.009	1.442

proposed model achieved better results for all the three countries when compared to other models, indicating the effectiveness of using optimized LSTM framework using GWO.

To show the effectiveness of the use of GT data for forecasting the trends of the infection when compared to the traditional LSTM , the percentage improvement of the ECDC - GT - L when compared to ECDC - L is shown in Table. 11. From the values obtained it is clear that using GT data improves the forecasting results.

It is also observed that the error value in the case of NC is higher when compared to TCC and TCD. This is due to the fact that the TCC and TCD, being monotonic in nature, produces a very effective forecast. The same trend can be observed for the other two countries also.

Therefore, the proposed model, i.e., using ECDC data and GT as inputs to the LSTM model that uses optimal window size derived from GWO, and the features selected using Spearman's correlation coefficient prove to be the better-performing model, when compared to its other variants.

Table 8
Percentage improvement in MAPE values for proposed model (ECDC-GT-GWO-L) when compared to other three models for India

S.No.	*Model*	*TCC(%)*	*NC(%)*	*TCD(%)*
1.	ECDC-A	91.71	68.39	98.89
2.	ECDC - L	75.61	65.87	97.78
3.	ECDC - GT - L	16.8	22.89	95.57

Table 9
Percentage improvement in MAPE values for proposed model (ECDC-GT-GWO-L) when compared to other three models for USA

S.No.	*Model*	*TCC(%)*	*NC(%)*	*TCD(%)*
1.	ECDC-A	70.05	24.34	73.04
2.	ECDC - L	75.74	11.17	82.37
3.	ECDC - GT - L	18.24	6.78	31.53

7. Conclusion and Future Work

In this paper, a novel workflow for forecasting future trends of COVID-19 spread using the historical ECDC data and Google Trends has been proposed. The influence of the pandemic is incident on

Table 10
Percentage improvement in MAPE values for proposed model (ECDC-GT-GWO-L) when compared to other three models for UK

S.No.	Model	TCC(%)	NC(%)	TCD(%)
1.	ECDC-A	80.07	58.97	81.38
2.	ECDC - L	81.14	64.62	53.76
3.	ECDC - GT - L	76.23	24.80	41.93

Table 11
Percentage improvement in MAPE values for ECDC-GT-L when compared to ECDC-L for India, USA, UK

S.No.	Country	TCC(%)	NC(%)	TCD(%)
1.	India	70.66	55.74	49.93
2.	USA	70.33	4.71	74.25
3.	UK	20.64	52.95	53.78

the google search history of people across different countries. The well-known, Spearman's correlation has been used to calculate the most relevant search terms. GWO algorithm has been used to select the optimal hyper-parameters for LSTM architecture to achieve higher accuracy in forecasting total cumulative cases, deaths and new cases of infection. The analysis is carried out for some of the worst affected countries including India, USA and UK.

The results obtained establish that mining the search trends of the public in a particular region can be used to forecast the future number of cases of a disease or infection. The different experiments conducted have produced effective results for all the three countries. Also, the relevant search terms keeps on changing as a pandemic progresses. Thus, it becomes imperative that the GT terms are continuously evaluated and updated for forecasting the spread. The results can be improved further when the search terms of higher correlation values are used. Hence, the proposed workflow can be adopted significantly in future for the prediction of the spread of pandemics.

References

[1] Muhammad Adnan Shereen, Suliman Khan, Abeer Kazmi, Nadia Bashir, Rabeea Siddique. "COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses", *Journal of Advanced Research*, 2020, vol. 24, pp. 91-98.

[2] Coronavirus Updates, 28 July2020, Available. https://www.ECDC.int/emergencies/diseases/novel-coronavirus-2019

[3] Dayi Jin, Jiang-Shan Lian, Jian-Hua Hu, et al, "Epidemiological, clinical and virological characteristics of 74 cases of coronavirus-infected disease 2019 (COVID-19) with gastrointestinal symptoms", *BMJ Journals*, 2020, vol. 69, pp. 1002-1009.

[4] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, Zhenshun Cheng, Ting Yu, Jiaan Xia, Yuan Wei, Wenjuan Wu, Xuelei Xie, Wen Yin, Hui Li, Min Liu, Yan Xiao, Hong Gao, Li Guo, Jungang Xie, Guangfa Wang, Rongmeng Jiang, Zhancheng Gao, Qi Jin, Jianwei WangâĂă, Bin Cao, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China", *The Lancet*, 2020, vol. 395, pp. 497-506.

[5] Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu, Lulu Guan, Yuan Wei, Hui Li, Xudong Wu, Jiuyang Xu, Shengjin Tu, Yi Zhang, Hua Chen, Bin Cao, "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study", *The Lancet*, 2020, vol. 395, pp. 1054âĂŞ62.

[6] Xin Li, Jian Ma, Shouyang Wang and Xun Zhang. " How does Google Search affect traders position and crude oil prices? ".*Economic Modelling*, 2015 , vol. 49 , pp. 162-171

[7] Lean Yu, Yaqing Zhao, Ling Tang, Zebin Yang, "Online big data-driven oil consumption forecasting with Google trends", *International Journal of Forecasting*, 2019, vol. 35, pp 213-223.

[8] Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina, Joe Hasell, "Coronavirus Pandemic (COVID-19)", 2020, Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus' as of 21 May, 2020 [Online Resource].

[9] Dires Negas Fente, Dheeraj Kumar Singh, " Weather Forecasting Using Artificial Neural Network", in Proc. *Second International Conference on Inventive Communication and Computational Technologies*, 2018. pp. 1757-1761

[10] Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar,"Stock Closing Price Prediction using Machine Learning Techniques", in *Procedia Computer Science*, vol. 167,2020, pp. 599-606.

[11] Chun Yang Lai, Rung Ching Chen, Rezzy Eko Charaka, "Prediction Stock Price Based on Different Index Factors Using LSTM", in Proc.*2019 International Conference on Machine Learning and Cybernetics*, 2019, pp. 1-6.

[12] Siva R Venna, Amirhosseun Tavanaet, Raju N. Gottumukkala, Vijay V. Raghavan, Anthony S. Maidi and STtephen Nichols. "A Novel Data-Driven Model for Real-TimeInfluenza Forecastin". *IEEE Access*, 2018, vol. 7, pp. 7691 - 7701

[13] Yunjun Yu, Junfei Cao and Jianyong Zhu. "An LSTM Short-Term Solar Irradiance Forecasting Under Complicated Weather Conditions". 2019, vol. 7, pp. 145651 -

145666

[14] Deeksha Chandola, Harsh Gupta, Vinay Anand Tikkiwal, Manoj Kumar Bohra, " Multi-step ahead forecasting of global solar radiation for arid zones using deep learning", *Procedia Computer Science*, vol. 167, 2020, pp. 626-635.

[15] Baowei Wang, Weiwen Kong, Hui Guan and Neil. N Xiong. " Air Quality forecasting based on gated recurrent long short term memory model in Internet of Things". *IEEE Access*, 2019, vol. 7, pp. 69524 - 69534

[16] Uttam Singh, Shewta Kedas, S. Prasanth, Arun Kumar, Vijay Bhaskar Semwal and Vinay Anand Tikkiwal. "Design of A Recurrent Neural Network Model for Machine Reading Comprehension",*Procedia Computer Science* 2020, vol. 167, pp. 1791 - 1800

[17] A. F. Ganai and F. Khursheed, "Predicting next Word using RNN and LSTM cells: Stastical Language Modeling," in Proc. *2019 Fifth International Conference on Image Information Processing (ICIIP)*, Shimla, India, 2019, pp. 469-474

[18] Pei Hu, Jeng-Shyang Pan, Shu-Chuan Chu, "Improved Binary Grey Wolf Optimizer and Its application for feature selection", *Knowledge-Based Systems*, 2020, vol. 195, pp. 105746 - 105758

[19] R. Rajakumar, J. Amudhavel, P. Dhavachelvan, T. Vengattaraman. (2017) "GWO-LPWSN: Grey Wolf Optimization Algorithm for Node Localization Problem in Wireless Sensor Networks", *Journal of Computer Networks and Communications*, Volume 2017, Article ID 7348141.

[20] Mingjing Wang, Huiling Chen, Huaizhong Li, Zhennao Cai, Xuehua Zhao, Changfei Tong, Jun Li, Xin Xu. (2017) "Grey wolf optimization evolving kernel extreme learning machine: Application to bankruptcy prediction", *Engineering Applications of Artificial Intelligence*, vol. 63, pp. 54-68.

[21] Xianhai Song, Li Tang, Sutao Zhao, Xueqiang Zhang, Lei Li, Jianquan Huang, Wei Cai, "Grey Wolf Optimizer for parameter estimation in surface waves", *Soil Dynamics and Earthquake Engineering*, 2017, vol. 75, pp. 147-157.

[22] Domenico Benvenutoa, Marta Giovanettib, Lazzaro Vassalloc, Silvia Angelettid and Massimo Ciccozzib." Application of the ARIMA model on the COVID-2019 epidemic dataset". *Data Brief*, 2020, vol. 29, pp. 105340 - 105343

[23] Ram Kumar Singh, Meenu Rani, Akshay Sreekanth Bhagvathula, Ranjith Shah, Alfonso J R Morales, Himangshu Kalita, Chintan Nanda, Shashi Sharma, Yagya Dutt Sharma, A Rabban , Jamal Rahmani and Pavan Kumar. "Prediction of the COVID-19 Pandemic for the Top 15 Affected Countries: Advanced Autoregressive Integrated Moving Average (ARIMA) Model". *JMIR Public Health Surveill*, 2020, vol. 6.

[24] Zeynep Ceylan, "Estimation of COVID-19 prevelance in Italy, Spain and France", *Science Of the Total Environment*, 2020, vol. 79.

[25] Jagadish Kumar and K.P.S.S Hembram. " Epidemiological survey of novel coronavirus (COVID - 19)", https://arxiv.org/abs/2003.11376.

[26] Furqan Rustum, Aijaz A. Reshi, Arif Mehmood, Saleem Ullah, Byung-Won On, Waqar Aslam, and Gyu Sang Choi, " COVID-19 Future Forecasting Using Supervised Machine Learning Models", *IEEE Access*, 2020, vol. 8, pp 101489 - 101499

[27] Alok Kumar Sahai, Namita Rath, Vishal Sood and Manvendra Pratap Singh. "ARIMA modelling & forecasting of COVID-19 in top five affected countries", *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 2020, vol.14 , pp. 1419-1427

[28] Nalini Chintalapudi, Gopi Battineni and Francesco Amenta. "COVID-19 virus outbreak forecasting ofregistered and recovered cases after sixtyday lockdown in Italy: A data driven modelapproach",*Journal of Microbiology, Immunology and Infection*, 2020, vol. 53, pp. 396 - 403

[29] Ricardo Manuel Arias VelÃąsquez and Jennifer Vanessa MejÃŋa Lara."Forecast and evaluation of COVID-19 spreading in USA with re duce d-space Gaussian process regression ",*Chaos, Solitons and Fractals, 2020, vol. 136 ,pp. 109924 -109932*

[30] Vinay Kumar Reddy Chimmula and Lei Zhang, "Time Series forecasting of COVID-19 transmission in Canada Using LSTM Networks", Chaos Solitons and Fractals, 2020 . vol. 135, pp. 109864-109869

[31] Anuradha Tomar and Neeraj Gupta. "Prediction for the spread of COVID-19 in India and effectiveness of preventive measures" Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. 2020, vol. 728, pp. 138762 - 138767

[32] Mohammed R. Ibrahim, James Haworth, Aldo Lipani, Nilufer Aslam, Tao Cheng and Nicola Christie, "Variational LSTM Autoencoder to forecast the spread of coronavirus across the globe", med$R_x$iv, 2020, vol. 12, Issue. 1.

[33] Sangwon Chae, Sungjun Kwon, Donghyun Lee, "Predicting Infectious Disease Using Deep Learning and Big Data", International Journal of Environmental Research and Public Health, 2018, vol. 15, Issue. 8, pp. 1596.

[34] Vasileios Lampos, Nello Cristianini, "Tracking the flu pandemic by monitoring the Social Web." in Proc. Second International Workshop on Cognitive Information Processing, 2020, pp. 411-414.

[35] Alessio Signorini, Alberto Maria Segre, Philip M. Polgreen, "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic", Plos One, 2011, vol. 6, Issue 5, pp. 1-10.

[36] Wiwik Anggraeni and Laras Aristiani. " Using Google Trends data in forecasting number of dengue fever cases with ARIMAX method case study". in Proc. 2016 International Conference on Information & Communication Technology and Systems , 2016, pp. 114-118.

[37] Yue Teng, Dehua Bi, Guigang Xiu, Yuan Jin, Yong Huang, Baihan Lil, Xiaoping An, Dan Feng and Yigang Tong, "Dynamic Forecasting Of Zika Epidemics using Google Trends", Plos One, 2017, 12(1).

[38] Maria Effenberger, Andreas Kronbichler, Jae Il Shin, Gert Mayer, Herbert Tilg and Paul Percoe, "Association of the COVID-19 pandemic with Internet Search Volumes : A Google Trends Analysis", International Journal Of Infectious Diseases, 2020, vol. 95, pp. 192-197.

[39] Google Trends data retrieved from

'https://www.google.com/trends' [Online Resource].

[40] G. Peter Zhang, "Time series forecasting using a hybrid ARIMA and neural network model", Neurocomputing, 2003, Vol. 50, pp. 159-175.

[41] A.J Conejo, M.A Plazas, R. Espinola, A.B. Molina, "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models", IEEE transactions on power systems, 2005, vol. 52, pp. 1035-1042.

[42] Sepp Hochreiter, JÃijrgen Schmidhuber, "Long Short-Term Memory", Neural Computation, 1997, vol. 9, Issue. 8, pp. 1735-1780.

[43] Hasim Sak, Andrew Senior, Francoise Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling", INTERSPEECH-2014, 2014, pp. 338-342.

[44] Seyedali Mirjalili, Seyed Mohammad Mirjalili, Andrew Lewis, "Grey Wolf Optimizer", Advances in Engineering Software, 2014, Vol. 69, pp. 46-61.

[45] Indrajit N. Trivedi, Amir H. Gandomi, Pradeep Jangir, Narottam Jangir, "Study of Different boundary constraint handling Schemes in Interior Search Algorithm", in International Conference on Artificial Intelligence and Evolutionary Computations in Engineering Systems, 2016, vol 517.

[46] P.M. Swamidass, "MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)", Encyclopedia of Production and Manufacturing Management, 2020, pp. 462-462.

Dear Editor,

We confirm that this is our first attempt of submitting this manuscript. This manuscript introduces our original work. None of the parts of this manuscript is currently under consideration or published in any other journals.

Kind Regards,

Dr. Arun Kumar
Assistant Professor,
Room No: CS-217,
Department of Computer Science & Engineering,
National Institute of Technology,
Rourkela, Odisha - India 769008
Web: http://www.nitrkl.ac.in/CS/~kumararun/
Phone: +91-661-2462373 (O), +91-9971867785 (M)