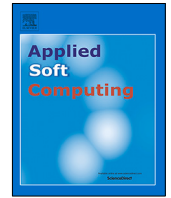




Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.



# Real-time measurement of the uncertain epidemiological appearances of COVID-19 infections



Meenu Gupta<sup>a</sup>, Rachna Jain<sup>b</sup>, Soham Taneja<sup>b</sup>, Gopal Chaudhary<sup>b</sup>, Manju Khari<sup>c</sup>, Elena Verdú<sup>d,\*</sup>

<sup>a</sup> Chandigarh University, Punjab, India

<sup>b</sup> Bharati Vidyapeeth's College of Engineering, Delhi, India

<sup>c</sup> Netaji Subhas University of Technology, East Campus, Delhi, India

<sup>d</sup> Universidad Internacional de La Rioja, Logroño, Spain

## ARTICLE INFO

### Article history:

Received 14 October 2020

Received in revised form 4 December 2020

Accepted 13 December 2020

Available online 25 December 2020

### Keywords:

COVID-19

Epidemiology

Outbreak

Deep learning

LSTM

## ABSTRACT

Virus diseases are a continued threat to human health in both community and healthcare settings. The current virus disease COVID-19 outbreak raises an unparalleled public health issue for the world at large. Wuhan is the city in China from where this virus came first and, after some time the whole world was affected by this severe disease. It is a challenge for every country's people and higher authorities to fight with this battle due to the insufficient number of resources. On-going assessment of the epidemiological features and future impacts of the COVID-19 disease is required to stay up-to-date of any changes to its spread dynamics and foresee needed resources and consequences in different aspects as social or economic ones. This paper proposes a prediction model of confirmed and death cases of COVID-19. The model is based on a deep learning algorithm with two long short-term memory (LSTM) layers. We consider the available infection cases of COVID-19 in India from January 22, 2020, till October 9, 2020, and parameterize the model. The proposed model is an inference to obtain predicted coronavirus cases and deaths for the next 30 days, taking the data of the previous 260 days of duration of the pandemic. The proposed deep learning model has been compared with other popular prediction methods (Support Vector Machine, Decision Tree and Random Forest) showing a lower normalized RMSE. This work also compares COVID-19 with other previous diseases (SARS, MERS, h1n1, Ebola, and 2019-nCoV). Based on the mortality rate and virus spread, this study concludes that the novel coronavirus (COVID-19) is more dangerous than other diseases.

© 2020 Elsevier B.V. All rights reserved.

## 1. Introduction

To control the environment where humans are living is a challenging task. However, every living being has its specified boundaries (especially for human beings) where violation of the law of nature is prohibited. In the process of controlling the environment, humans constructed powerful instruments, which allow them to control the earth, air, and sea. Although they also succeed in violating fundamental laws of nature, that lead to many disasters. For example, due to personal benefits, human beings promote the creation of guns, bombs, etc.

Although nature has given many of the things like fruit, vegetables, etc., that may allow a human to survive, human beings eat sea living-being, some coming from seafood markets that are

reservoirs of infectious diseases. Moreover, the human being has not stopped to eat sea living beings, but they do some sexual contact with them, causing new diseases. HIV is the example of that. HIV is a deadly virus disease that came into this world because of the genesis of Chimpanzees to humans [1]. Many more diseases came into existence in the earth due to the carelessness of the behavior of human beings. A few of the deadly virus diseases, as shown in Fig. 1 are discussed below.

### 1.1. Swine flu

Flu, also named as influenza, originated from birds and other animals such as pigs. Swine flu came in 1988 from the U.S., where it was found in the pig. This disease spread from animals (i.e., birds) to humans. Usually, a human does not get swine flu, but human infection can and does happen. In 2009, the swine flu strain called h1n1 infected people [2,3].

\* Corresponding author.

E-mail addresses: [meenu.e9406@cumail.in](mailto:meenu.e9406@cumail.in) (M. Gupta), [rachna.jain@bharativedyapeeth.edu](mailto:rachna.jain@bharativedyapeeth.edu) (R. Jain), [sohamtaneja12@gmail.com](mailto:sohamtaneja12@gmail.com) (S. Taneja), [gopal.bvcoe@bharativedyapeeth.edu](mailto:gopal.bvcoe@bharativedyapeeth.edu) (G. Chaudhary), [manjukhari@aiactr.ac.in](mailto:manjukhari@aiactr.ac.in) (M. Khari), [elena.verdu@unir.net](mailto:elena.verdu@unir.net) (E. Verdú).

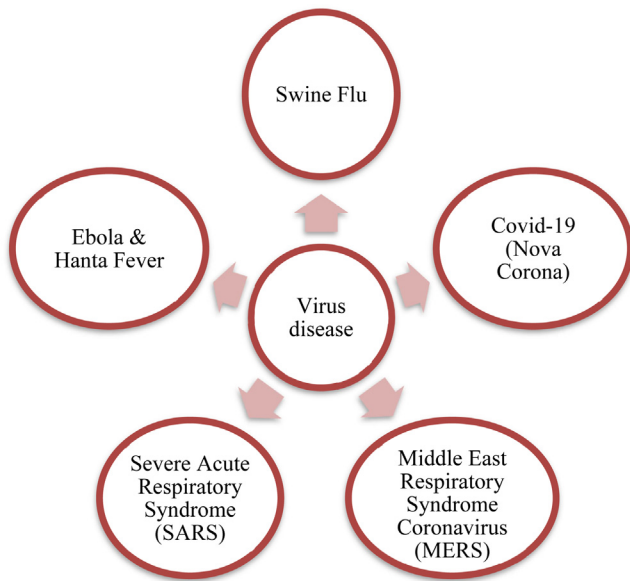


Fig. 1. Classification of deadly virus diseases.

### 1.2. Ebola & hanta fever

Ebola is a deadly virus disease which comes from infected animals such as fruit bat [1]. Because of this virus disease, lots of human beings lost their lives. The family of these viruses includes one of the members of a virus called hantavirus, which mostly comes from rodents and could be a cause of diverse disease syndromes in people worldwide. Infection with any hantavirus can produce a hantavirus disease in people. Hantavirus-infected deer mice (*Peromyscus maniculatus*) can excrete the virus in their saliva, urine, and droppings [4].

### 1.3. Severe Acute Respiratory Syndrome (SARS)

SARS is a deadly virus disease caused by the coronavirus SARS-CoV, which also comes from an animal reservoir, especially bats, and it spreads from animals (i.e., civet cats) to human beings. This virus came from China in the year 2002. In the year 2003, the SARS-CoV spread worldwide within a few months, and due to this virus, many people lost their lives. This virus transmitted through the air where someone had the disease [5,6].

### 1.4. Middle East Respiratory Syndrome (MERS)

MERS is a family of novel coronavirus which was identified in Saudi Arabia in the year 2012 [7]. This disease is not transmitted from person to person easily. It only spreads through the carelessness of medical staff [8].

### 1.5. COVID-19 (novel coronavirus)

The COVID-19 is a deadly virus disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV2). This is a positive-sense single-stranded RNA virus, and it was first identified in a China city, Wuhan, in the year 2019. This pandemic has drastically changed human life which has been affected badly. The lifestyle and thought process of every individual has changed with the current situation. This is considered as a situation with unpredictability and uncertainty. This deadly virus disease already infected many people around the world, and thousands of people lost their lives [9]. As far, human beings have developed

many vaccines for the above said diseases, but for COVID-19 the researchers are still finding vaccines to save human lives. COVID-19 outbreak raises an unparalleled public health issue for the world at large. It is a challenge for every country's people and higher authorities to fight in this battle due to the insufficient number of resources such as doctors, hospitals, securities, nurses, etc. On-going assessment of the epidemiological features and future impacts of the COVID-19 disease is required to stay up-to-date of any changes to its spread dynamics, and foresee needed health resources as well as consequences in different aspects as social or economic ones.

### 1.6. Contribution of the paper

The main aim of this paper is to create a model that predicts the total number of infected and death cases due to this virus disease and to compare it with past epidemics. The model is built using a deep learning algorithm. There are other works predicting COVID-19 cases but they are limited by the inaccurate data available during the first months, when few tests for diagnostics were done. We consider the available infection cases of COVID-19 in India from January 22, 2020, till October 9, 2020, and parameterize the model. The proposed model is an inference to obtain predicted coronavirus cases and deaths for the next 30 days, taking the duration of the pandemic to a total of 260 days. Besides, this paper considers five different virus diseases (SARS, MERS, h1n1, Ebola, and 2019-nCoV) and performs a comparative analysis. Based on the mortality rate and high capacity to spread, the novel coronavirus (COVID-19) is more dangerous than other diseases. Also, the affected countries are far greater than those affected by Ebola, h1n1, SARS, and MERS. It is expected to rise further if the current trend continues. However, the mortality rate for Ebola was much higher (39.5%) against COVID-19 (1.55%) but the number of deaths is lower. This is primarily due to the lesser number of Ebola cases.

The contributions can be summarized as these three points:

- This paper considers five different virus diseases (SARS, MERS, h1n1, Ebola, and 2019-nCoV) to perform a comparative analysis and uses a COVID-19 prediction model built with a deep learning algorithm to predict the number of infected cases and deaths due to COVID-19.
- The proposed model is an inference to obtain predicted coronavirus cases and deaths for the next 30 days, taking the duration of the pandemic to a total of 260 days. This model concludes, based on the mortality rate, that the novel coronavirus (COVID-19) is more dangerous than other diseases.
- This paper focuses on a deep learning model based on LSTM networks to forecast future COVID cases.

Related work has been discussed in Section 2. The dataset used, the numerical model formulation and the experimental setup is explained in Section 3. Then, numerical results based on the data and model are discussed with comparative analysis in Section 4. Concluding remarks are provided in Section 5.

## 2. Related work

According to the World Health Organization (WHO) [10], COVID-19 is a beta coronavirus that affects the lower respiratory area. It is revealed as pneumonia in humans [11]. Despite severe global restraint and isolation efforts, the incidence of COVID-19 rises rapidly worldwide [12].

Two decades ago, looking for novel agents that caused respiratory infections, different novel respiratory viruses were discovered [13]. Coronaviruses have large genomes and a high frequency of recombination. SARS-CoV-2 is the novel coronavirus

causing COVID-19, that has required the launching of new investigations. Given the severity of this disease and urgency of the situation, many studies have been developed by researchers in different areas. For example, a study [14] discussed the genome sequence of nCoV-2019, which is publicly shared for phylogenetic analysis. In this study authors find out the estimating origin of virus in the human body and impact of last coronavirus. A serological method ELISA (Enzyme-Linked Immunosorbent Assays) with recombinant antigens inferred from the spike protein of SARS-CoV-2 is discussed in another research [15]. The negative control sample, which represents pre-COVID-19, has been considered to develop these assays. The proposed model discussed in this work was invented for detecting the viral cause and diagnosis. COVID-19 patient's data sample is considered in this work for analysis. Further, they examine samples for screening and identification of COVID-19 using human serum (or plasma) some days post symptoms. They remark this examination does not require the handling of infection virus; even it can be adjusted to detect different antibody types and is adaptable to scaling. Moreover, these type of studies helps to identify previous exposure to the virus in persons, to identify donors for plasma therapy and discover correlates of protection. Due to low availability of the RNA extraction kits that are required to detect this disease, authors in another work [16] discuss the investigation of two commercial (RT-qPCR) kits and they test whether they are compatible in detecting SARS-Cov-2 virus disease from nasopharyngeal swab samples. As a result, they find that one of the kits tested was fully compatible with direct SARS-Cov-2 disease detection and diagnosis. Due to this difficulty to carry out the required tests to control the pandemic via the above mentioned kits, different researchers have proposed alternative methods to support diagnosis, for example, those based on deep learning techniques that process X-ray images [17–19].

To discover the characteristics of the spread pattern of this disease, another research [20], which was undertaken in the first months of the appearance of the disease, considers the data of six persons of the same family, collecting their details about the clinical, laboratory, radiological, epidemiological, and microbiological findings. Five of these persons suffered an existence of unpredictable pneumonia after a visit to Wuhan, while another family member had not traveled to Wuhan. In their findings, they concluded that out of six people, five were infected from COVID-19. One person who did not visit Wuhan also became infected after a few days of contact with other family members. They concluded that COVID-19 is a deadly disease that spreads from person-to-person in family groups, hospitals, clinics, or any other geographical region. Another report [21] discusses the first case of 2019-nCov in the United States (US) and also about the identification, diagnosis, clinical course, and management of the case, including the evolution from first symptoms to a stage of pneumonia after 9 days of illness. This study also highlights the importance of close coordination between the clinic and public health authorities at the local, state, and federal level to overcome this spread of diseases. Further in another study [22], authors discuss the outbreak of COVID-19 in King County, Washington. From a first case identified, the researchers found that on March 18, 2020, 167 cases were confirmed positive from this disease. They concluded that proactive steps are required to prevent this disease. In another study the authors compare the clinical and immunologic characteristics of moderate versus severe COVID-19 impacts [23]. This study also characterizes the cytokine storm in severe COVID-19 and provides insights into the design of the vaccine.

From these first confirmed cases, as the disease spreads, more data is available worldwide, and different studies arise exploiting the data with different aims, as predicting its expansion or mortality rate. For example, one work [24] suggests an orthogonal

approach with a limited number of parameters taken from data from different countries available at the John Hopkins database, a worldwide reference database. The author discusses the estimation of lethality and undetected infection associated with SARS-cov2. He concludes that the apparent death rate of this outbreak can be extrapolated to infinite time after a few days from the starting of this disease. Another author worked on the impact of Coronavirus in Brazil, South America [25]. The author applies a susceptible–Infection–Quarantines–Recover model for collected data of patients from the health department of Brazil. In his analysis, he calculated the ratio of the confirmed cases and unidentified cases, and estimated the epidemic doubling time that is 2.72 days. According to another study [26], the basic reproduction number for COVID-19 can be estimated from data of some countries (China, Iran, South Korea, and Italy) using the SEIR type model with accurate results. The SEIR model is also used to simulate the spread of the epidemic under different intervention scenarios in different countries of Africa [27].

In [28], the authors discussed the impact of coronavirus-19 all over the world. After analyzing the rules followed by COVID-19 infected countries, especially China, applying a microsimulation model created to support pandemic influenza, they concluded that social distancing, home isolation, closure of industries, and institutions could help to suppress this transmission and rapidly reduce the incidence cases. Further, in [29], the authors considered two cases of virus diseases such as moderate (i.e., SARS and HIV) and highly transmissible viruses (i.e., Smallpox and Pandemic influenza). They consider two health measures, which are isolating symptomatic individuals and quarantine their contacts, and use a mathematical model on the relative timing of infectiousness and the appearance of symptoms. As a result, they concluded that SARS and smallpox were easier to control by using simple health measures. But HIV and Pandemic influenza are infections that spread by person to person and have a high-risk factor. Machine learning methods have been used to assess the countries performance against COVID-19 and analyze influential parameters [30], finding that results of measures adopted are not affected by parameters as smoking rates or rate of diabetes patients.

A report estimates 4000 cases of COVID-19 in Wuhan city by January 18, 2020 [31]. This study discusses the delay time (i.e., ten days) of infection and the detection of this disease. For their analysis they assume that international travel is not related to the risk of exposure or to the infectious status. Another work [32] uses travel-based connectivity to estimate the potential risk and geographical range of COVID-19 spread within China and all over the world from January 20 to April 20, 2020. This study concludes that risk factors are high where the populations are traveling more. In another study [33], the authors worked on reported cases of COVID-19 infection within China, combining Bayesian inference, mobility data, and Networked Dynamic Meta population to understand acute epidemiological features associated with SARS-CoV2. They concluded that 86% of infection cases were undocumented before January 23, 2020, before travel restriction. Due to travel, these undocumented cases increased rapidly. This study indicates the restraints due to SARS-Cov-2 disease and their challenges. Later on, these authors discussed the outbreak of COVID-19 within China for documented and undocumented infection [34].

Other researchers discuss [35] the impacts of influenza in the United States and propose a forecast system predicting the spatial transmission of influenza in the United States. To generate patterns of spatial transmission, they develop a metapopulation model that uses people mobility data. They achieve to predict local outbreaks up to six weeks in advance at state level. Later on, in another work [36] these authors used the metapopulation model



in the United States to analyze the spread and growth of virus disease COVID-19. They considered the dataset from February 21 to March 13, 2020, and evaluated epidemiological constraints, including the fraction of undocumented infections and their contagiousness. They project the outbreak for 180 days after March 13, 2020, and evaluate the outcomes of social distancing and travel restrictions in the condition of the outbreak.

Other authors worked on COVID-19 infected patient's cases in Hubei, mainland China collected from national and provincial health commissions on February 8, 2020, and on cases outside of mainland China collected from government or ministry of health websites [37]. They also considered dataset from media reports for 37 countries, as well as Hong Kong and Macau, until February 25, 2020. They calculate the fatality ratio of persons in different age group. Similarly, in another work [38], the authors consider the COVID-19 infected and death cases of Wuhan city from December 31 to January 28, 2020. The authors applied Markov Chain Monte Carlo methods for prediction and analysis of this acute disease. In another work [39], the authors considered the infected cases of coronavirus in Wuhan city and calculated the fatality risk of death and infected cases. Further other authors have studied about the occurrence of COVID-19 infected peoples in Wuhan, China [40]. This evaluation is done based on the infected reported cases over time by health authorities. The challenging task was to judge the level of severity of the collected information.

One difficulty in studies on new diseases is the insufficient amount of data, thus a methodology gathering data mining methods to augment data can be used [41] that has shown low prediction error in an experiment based on coronavirus outbreak when used together with a polynomial neural network with corrective feedback. Finally, other researchers [42] used a statistical method (i.e., purely data-driven) to evaluate the Case Fatality Rate (CFR) in the starting phase of the occurrence of COVID-19. Daily basis confirmed, and death cases due to COVID-19, were collected from January 10 to February 3, 2020, and segmented into three different clusters (i.e., City Wuhan, other cities of Hubei province, and other provinces of mainland China). Next, the authors applied a simple linear regression model to calculate the CFR from each cluster. As a result, they concluded that the CFR of COVID-19 is lower than the previous coronavirus epidemics caused by SARS-CoV and MERS-CoV.

Regarding other previous virus diseases, a study [43] reported the outbreak of Ebola in five countries of West Africa (i.e., Liberia, Senegal, Nigeria, Sierra Leone, and Guinea). The authors of this study considered the dataset of 3343 confirmed cases of Ebola and 667 probable cases. They used dynamic modeling and Bayesian inference to generate weekly forecasts of the outbreaks in the different countries, finding differences in accuracy per country.

Different researchers' views have been discussed in this section for the deathly disease. We propose to use advance predicting techniques, as those based on deep learning, to create a prediction model of infected and death cases of COVID-19. Other different methods can be found in literature with similar objectives, some above mentioned. For example, a mathematical model named SIPHERD has also been used to predict the total number of confirmed, active and death cases in India [44]. Also focusing on the country level, other authors propose the ARIMA prediction model to forecast the expected daily number of COVID-19 cases in Saudi Arabia for the next four weeks, from the collected data of almost two months in the country [45]. Using a modified Susceptible Exposed Infectious Recovered (SEIR) model, the spread of the epidemic under three intervention scenarios (suppression, mitigation, mildness) is simulated and predicted in South Africa, Egypt, Algeria, Nigeria, Senegal and Kenya, concluding in a series of epidemic controlling methods [27]. All these works focus on

country level. At local level, by using a simple Multiple Linear Regression Model, from daily time series of confirmed cases and calls received in a call center, the number of daily confirmed cases locally are forecasted, to help decision makers that need to organize the resources at local level [46]. The accuracy of this and benchmark models decreases as forecast horizon increases from a few days to 3 weeks. Focusing on other works using deep learning models, a work [47] uses also a LSTM based technique to predict confirmed cases of COVID-19 in India, with data from 30th January to 4th April 2020, then suffering from the limited data available in the first days as limitation. Other above mentioned studies as [27,45,47] also are based on data of the first months of the pandemic. The data considered in our study extend from January until October, thus implicitly includes the effects of different restrictions along the different months of the duration of the pandemic. Besides this work compares the potential impact of this novel coronavirus disease with other moderate and severe virus diseases, to get a better result in predicting and analyzing the impact of these diseases on human beings worldwide.

### 3. Materials and methods

#### 3.1. Datasets

We consider the dataset of five different virus diseases (i.e., Covid-19, Ebola, MERS, SARS, Swine flu (h1n1)). The Covid-19 dataset consists of daily statistics for coronavirus cases worldwide. Present data is taken from the web [48] from January 22, 2020 till October 9, 2020. In this, country-wise figure tally is provided for confirmed cases and deaths, along with recovered cases. With the help of this dataset, the number of affected cases and deaths were extracted to train our model for future predictions. Next, the second dataset considers data on Ebola outbreak in 2014 published on the web [49]. In this, a country-wise tally done for suspected and confirmed Ebola cases and deaths is provided. This dataset was deemed essential for our research to compare the mortality rates of the present pandemic. The next dataset of MERS is collected from the web [50]. This dataset considers several confirmed cases of MERS in the affected countries. The estimated deaths and mortality data were obtained from the CDC (Centres for Disease Control) [51] and WHO (World Health Organization) websites [52]. The fourth dataset of SARS is collected from the web (SARS: CDC [53]; SARS: WHO [54]). The mortality rate was inferred from the death statistics provided. Finally, the dataset of Swine flu (i.e., h1n1) is collected from the web (Pandemic: WHO; H1N1: CDC) [55]. In this case, results are approximate as provided on the internet.

#### 3.2. Model formulation

For model formulation, data has been collected from Github for coronavirus disease [48], and it was cleaned. A list was made to store the daily statistics for confirmed cases and deaths. We implemented a deep-learning-based time series prediction model to predict future cases.

The proposed model is an inference to obtain predicted coronavirus cases and deaths for the next days, taking the data of the pandemic of previous days. Fig. 2 shows the deep learning method for our proposed model. The "?" in Fig. 2 indicates data holders, which are variable and are set by the Deep Learning Framework (TensorFlow) when we provide the input data while training the model. Following, the mathematical equations used in the proposed model are shown:

$$\text{input gate: } i(t) = \text{sigmoid}(w_i[a(t-1), x(t)] + b_i) \quad (1)$$

$$\text{forget gate: } f(t) = \text{sigmoid}(w_f[a(t-1), x(t)] + b_f) \quad (2)$$

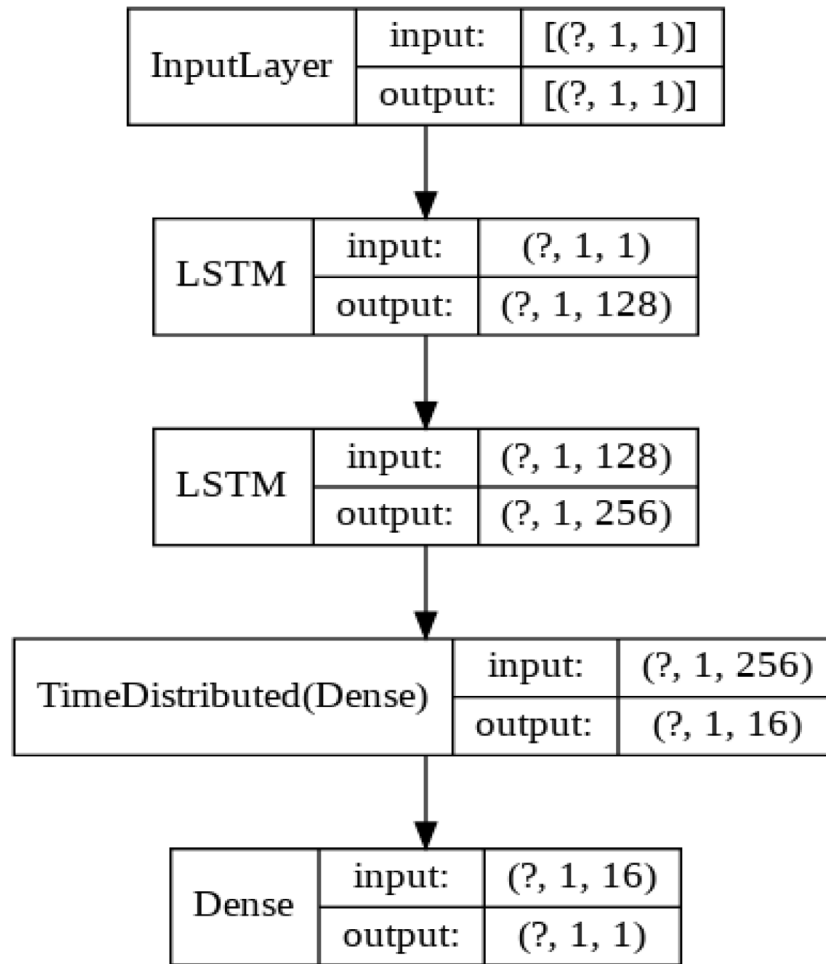


Fig. 2. Proposed model formulation.

$$\text{output gate: } o(t) = \text{sigmoid}(w_o + b_o) \quad (3)$$

$$\text{candidate: } \hat{c}(t) = \tanh(w_c [a(t-1), x(t)] + b_c) \quad (4)$$

$$\text{cellstate: } c(t) = i(t) * \hat{c}(t) + f(t) * c(t) \quad (5)$$

$$\text{output: } a(t) = o(t) * \tanh(c(t)) \quad (6)$$

$$\text{mean squared error} = \left(\frac{1}{2} * m\right) * \sum (y - \hat{y})^2 \quad (7)$$

$$m = W * A(t) + b \quad (8)$$

where,  $w_i$ ,  $w_f$ ,  $w_o$ , and  $w_c$  represent the weights of input gate, forget gate, output gate and candidate cell, respectively;  $b_i$ ,  $b_f$ ,  $b_o$ , and  $b_c$  represent the bias of input gate, forget gate, output gate and candidate cell, respectively;  $a(t-1)$  represents the previous layer output and  $x(t)$  represents the current input.

Eq. (1) shows that memory state receives the previous hidden state along with present input and it determines the usefulness of the present state. Eq. (2) is used to determine whether the previous memory state has to be propagated further or discarded. According to Eq. (3), the memory unit determines the next hidden state data. Eq. (4) is used to calculate the feasibility of the current input in determining the cell state. Using Eq. (5), we receive the input gate and forget gate data and use them to calculate the new cell state. Eq. (6) provides the final output prediction given by the LSTM cell. Further, Eq. (7) is used as an error matrix to calculate the deviation of our predictions with the actual results and use them to fine-tune our model.

Finally, the dense layer has been generated, as shown in Eq. (8). We infer the final results using a dense layer (denoted as

$A(t)$ ), which receives an output from the time distributed dense layer. In this proposed model, we first consider the impacts of COVID-19 and further compare it with another deadly disease.

### 3.3. Experimental setup

The proposed model is implemented using TensorFlow to infer predicted coronavirus cases and deaths for the next 30 days, taking the data of the pandemic of previous 260 days (from January 22, 2020 till October 9) in India obtained from [48].

Fig. 3 shows the flowchart of the proposed model. It starts with the initial data where we consider the confirmed and death cases of coronavirus disease of dates January 22, 2020 and January 23, 2020. This data further is converted into a time sequence of length 1. An LSTM layer with size 128 is used to learn the sequences provided, as shown in Fig. 3. Time sequences are provided as input to this layer. Another LSTM layer with size 256 is used to fetch sequences from the 1st LSTM layer and stacked on top of it. A time distributed dense layer with size 16 is used to produce neurons further to train the sequence on our data. Finally, a dense layer is used to fetch the predictions. Two separate models were trained, one for prediction of confirmed cases, other for prediction of the number of deaths. The ratio of the predicted number of deaths over the total number of cases was calculated to infer the expected mortality rate for the near future.

The evaluation of the model is done by comparing the obtained results with the real data. To further show the robustness

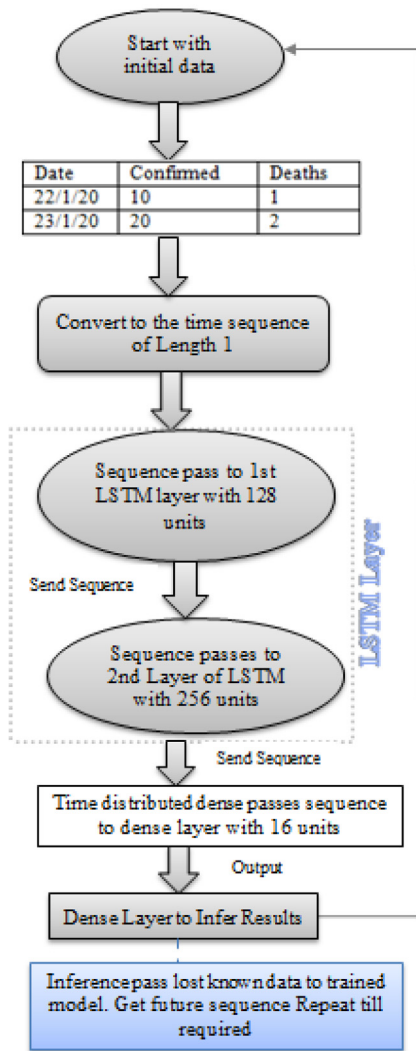


Fig. 3. Flowchart for the proposed model formulation.

of the proposed model, Root Mean Squared Error (RMSE) values are calculated for predicted confirmed cases and number of deaths of COVID-19.

RMSE is a measure of error between actual and predicted values, and it is a measure of how those residuals are spread out. The formula for RMSE is given in Eq. (9).

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2} \quad (9)$$

where  $N$  is the total number of observations which is 260 in our case,  $Y_i$  is the actual value, and  $\hat{Y}_i$  is defined as a predicted value. RMSE is finally normalized by mean normalization.

The RMSE will be used to compare the proposed deep learning model with other popular learning methods, namely Support Vector Machine [18], Decision Tree [56], and Random Forest [57].

#### 4. Results

##### 4.1. Prediction of confirmed cases and death cases of COVID-19 and comparison with previous diseases

An objective of this study is to predict virus-infected cases and deaths cases for the next 30 days, taking the duration of the

pandemic to a total of 260 days. This work also aims to compare the present disease impact with other epidemic diseases.

Table 1 contains the previous epidemic data collected from the WHO and CDC sites and actual data of coronavirus collected from [48]. The data of coronavirus in Table 1 corresponds to India while the other data is global. A comparison with the same is made using our predictions for the on-going coronavirus epidemic. Our model predicts that in the upcoming 30 days, the confirmed cases will shoot up to 6,493,402 cases in India. Also, the deaths are expected to rise to near about 100,563 steeply. Our model predicts the confirmed cases as well as the deaths similar to the current trend; hence our model can be used to calculate future predictions.

The mortality rate is calculated using Eq. (10) with these predicted values:

$$Mortality\ Rate = \left( \frac{predicted\ deaths}{predicted\ confirmed\ cases} \right) * 100 \quad (10)$$

The predicted mortality rate is 1.54 while actual mortality rate stands at 1.55.

The above data indicates that coronavirus is expected to have the worst death toll among the past epidemics and pandemics, although its mortality rate is lower than that of Ebola, SARS and MERS. The Ebola mortality rate was 39% due to the concentration of cases in the African subcontinent and less exposure to the rest of the world.

Fig. 4 gives a comparison of the real values of the number of current cases with the prediction of our model. Blue dots represent the actual values and orange line is for predicted values of the cases. Fig. 5 gives a comparison of the real values of the number of current deaths with the prediction of our model. Blue dots represent the actual number of deaths, whereas orange line represents our predicted values of deaths. The prediction graphs from our model depict that the model is performing close to the current scenario.

##### 4.2. Spread in most infected countries of COVID-19 and previous diseases

Fig. 6 shows the box plot representation of the current number of cases in the Top 5 worst-affected countries, and Fig. 7 shows the box plot representation of the current number of deaths in Top 5 worst-affected countries. It is evident from these plots that the coronavirus is much more widespread than the 2014 Ebola outbreak. The Ebola outbreak was limited to the US and African subcontinent. Hence, the nCov-19 possesses a threat to the world at an extremely high magnitude. The worst affected country seems to be the US. Despite China being the origin of coronavirus, we can see that China is no longer the worst affected country.

In Fig. 8, we have plotted the spread of infection for the first 50 days for the diseases COVID-19, h1n1, MERS, and SARS. From this graph, we infer that the rate of spread of COVID-19 is much more disastrous than any other past pandemic. The maximum spread in the past was caused by h1n1 swine flu, which was still less than the current conditions.

In Fig. 9, we have plotted the total Ebola cases that occurred worldwide. The maximum number of cases seems to be limited to the African subcontinent, and some in Europe, with a minor spread towards the USA. We can see that the Ebola Virus was contained and not spread in many countries, as compared to the last coronavirus outbreak.

In Fig. 10 world map, we have plotted the current cases globally. The map indicates a concentration of cases in the USA, which has become the new hotspot of the virus. The previous hotspots in Europe – Italy, Spain, etc., still seem to have a higher number of

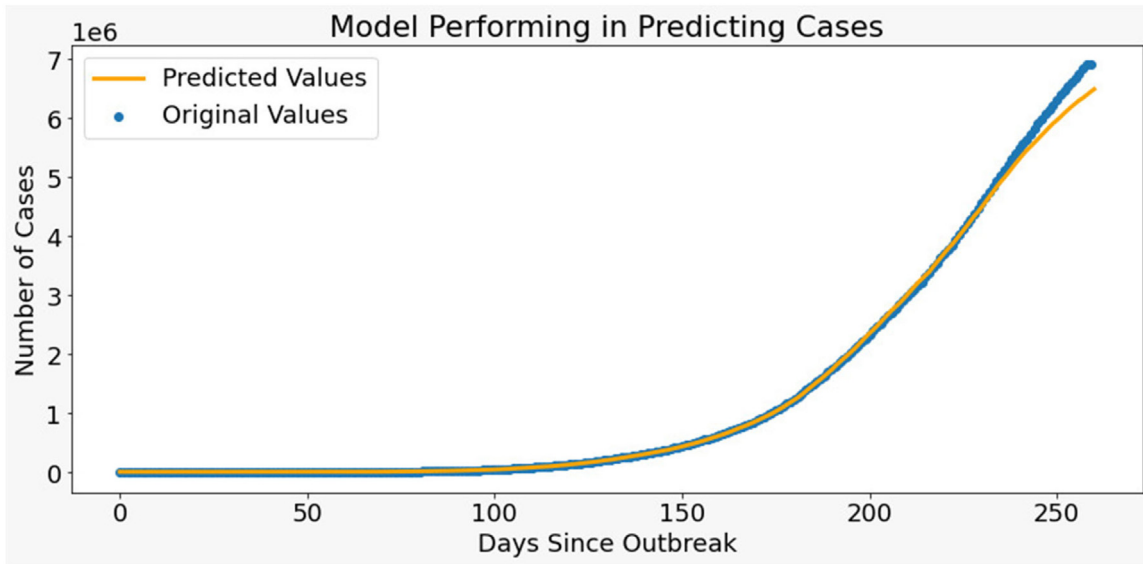


Fig. 4. Model Performance for predicting confirmed cases of Covid-19.

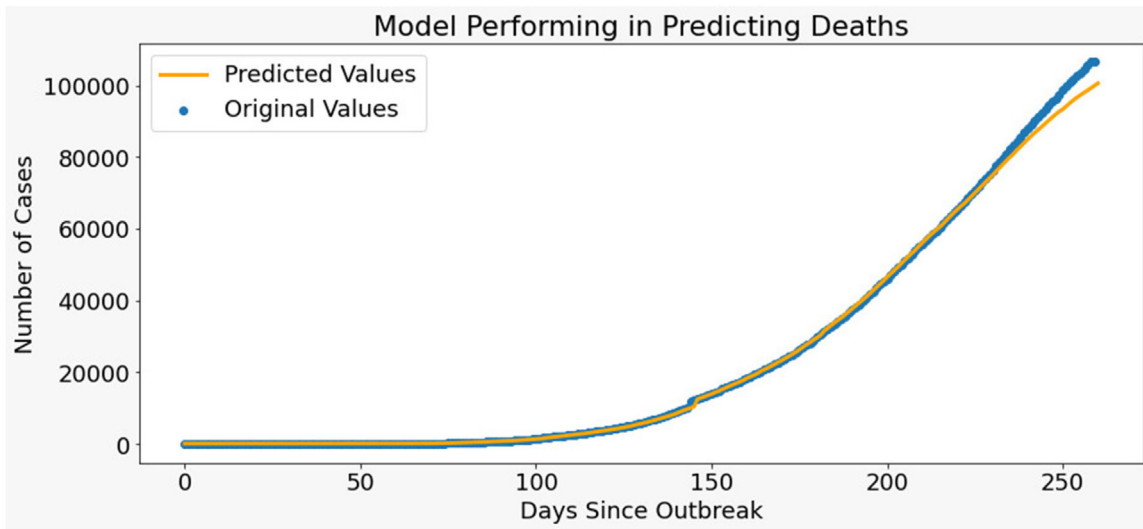


Fig. 5. Model performance for predicting the number of deaths of Covid-19.

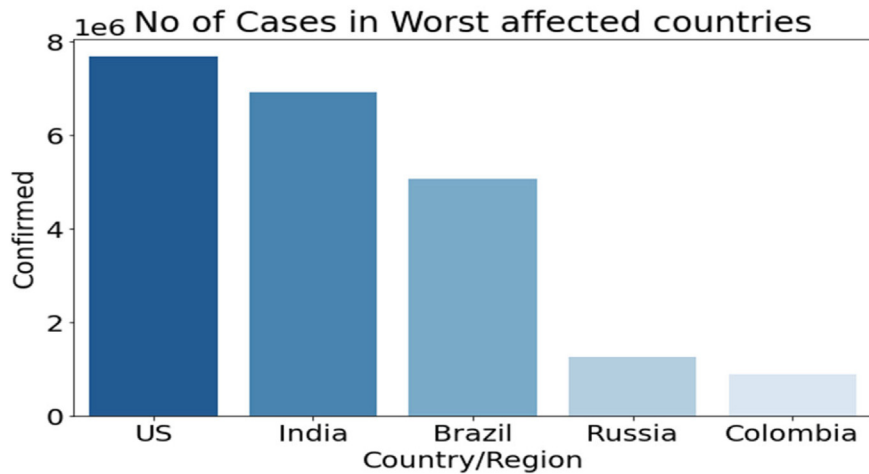


Fig. 6. Number of cases in Worst affected countries.



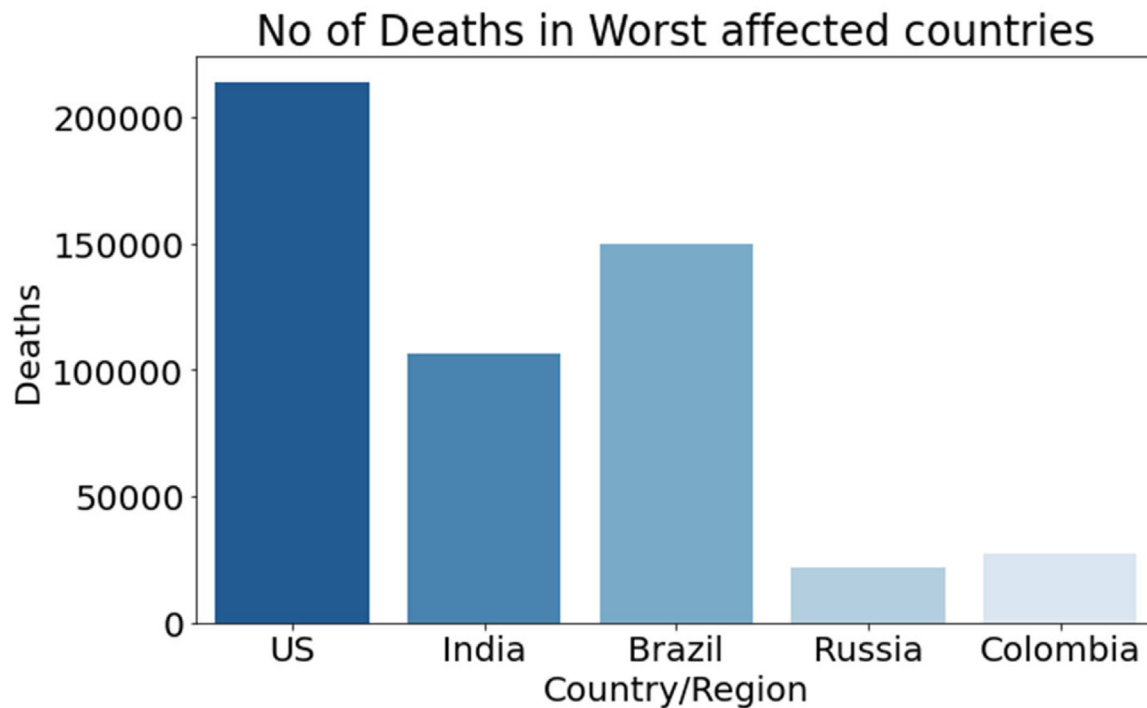


Fig. 7. Death Cases in Worst affected countries.

Table 1

Actual dataset collected from the WHO and CDC.

Outbreak	Start year	End year	Actual confirmed cases	Actual deaths	Mortality
CoronaVirus (India)	2019	Ongoing	6,906,151	106,490	1.55
Ebola	2014	2016	28,652	11,325	39.5
SARS	2003	2004	8096	774	9.5
MERS	2012	2019	2494	858	34
h1n1	2009	2010	60,800,000	151,700	0.25

cases in contrast with the rest of the world. Moreover, we infer from this map that coronavirus has spread in all the countries across the globe. Hence, we can conclude from these maps that the COVID-19 outbreak is more deadly than the Ebola outbreak.

### 4.3. Discussion

Firstly, a comparison of COVID-19 with previous epidemic diseases from data collected from the WHO and CDC site has been made. Then, as shown in Figs. 4 and 5, the proposed model predicts the confirmed cases as well as the deaths similar to the current trend; hence our model can be used to calculate future predictions. From COVID-19, h1n1, MERS, and SARS data, it is inferred that the rate of spread of COVID-19 is much more disastrous than any other past pandemic. The maximum spread in the past was caused by h1n1 swine flu, which was still less than the current conditions. In fact, against the 60.8M h1n1 cases, the number of cases for COVID-19 is 64.5M on December 3, 2020. The performance of the model depends on the data received from the online sources. One limitation is that the rate of testing varies, with less number of tests conducted in initial days. Hence, the confirmed cases were also less initially. The model is highly robust with the change of testing values in the later stages. For the model to work perfectly, the data must be accurate.

For comparison, RMSE values of other three algorithms, namely Support Vector Machine [18], Decision Tree [56], and Random Forest [57], are also calculated for the prediction of total confirmed cases and the number of deaths of COVID-19. For the SVM model, radial basis function has been used for prediction. The

Table 2

RMSE values of different algorithms for prediction of total confirmed cases and the number of deaths of COVID-19.

Algorithm used	Confirmed cases	Number of deaths
Support Vector Machine	0.2801	0.4323
Decision Tree	0.1108	0.1645
Random Forest	0.1101	0.1223
Proposed Method	0.0766	0.0533

normalized RMSE values of these methods and the proposed one are shown in Table 2.

Random Forest Algorithm shows a little bit better performance than the Decision Tree in the validation as normalized RMSE values of 0.1108 and 0.1223 are obtained for the prediction of total confirmed cases and the number of deaths of COVID-19, respectively. These results are far better than those of SVM. From the validation results, it is concluded that the Random Forest and Decision Tree algorithms both hold good results in the prediction. Support vector machine shows the lowest accuracy. Among all, the proposed model outperforms others in both the cases with normalized RMSE values of 0.0766 and 0.0533 for the prediction of total confirmed cases and the number of deaths of COVID-19, respectively. This shows the higher robustness of the proposed method in prediction of both the cases.

In this paper we have proposed a prediction method based on deep learning techniques due to their good performance in time series prediction problems and due to not having a time or complexity requirement. Other different methods can be found

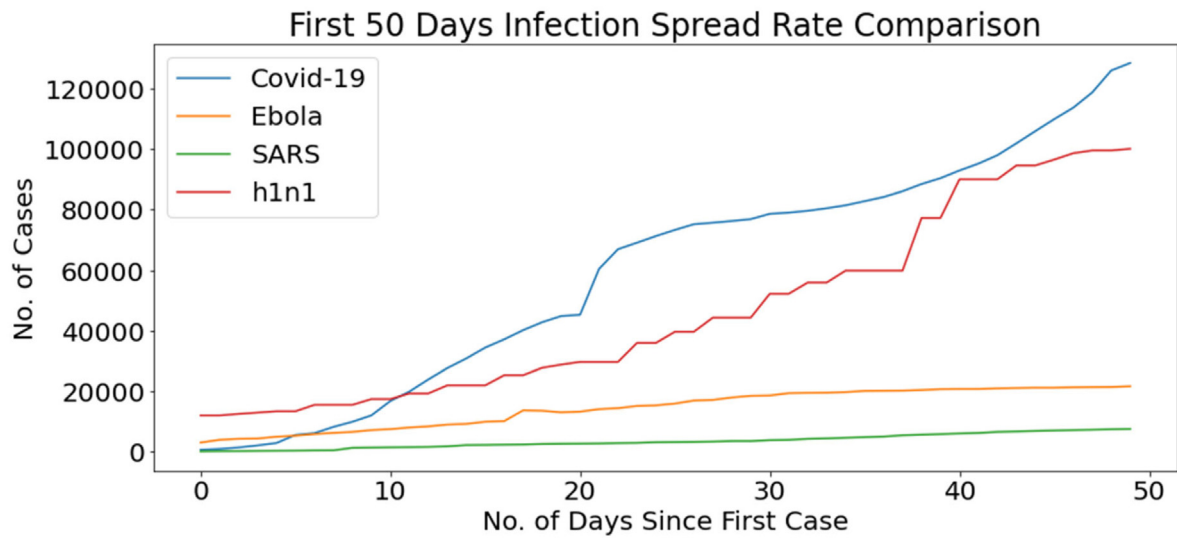


Fig. 8. First 50 days infection spread rate comparative analysis of Covid-19 vs. Ebola and SARS and h1n1.

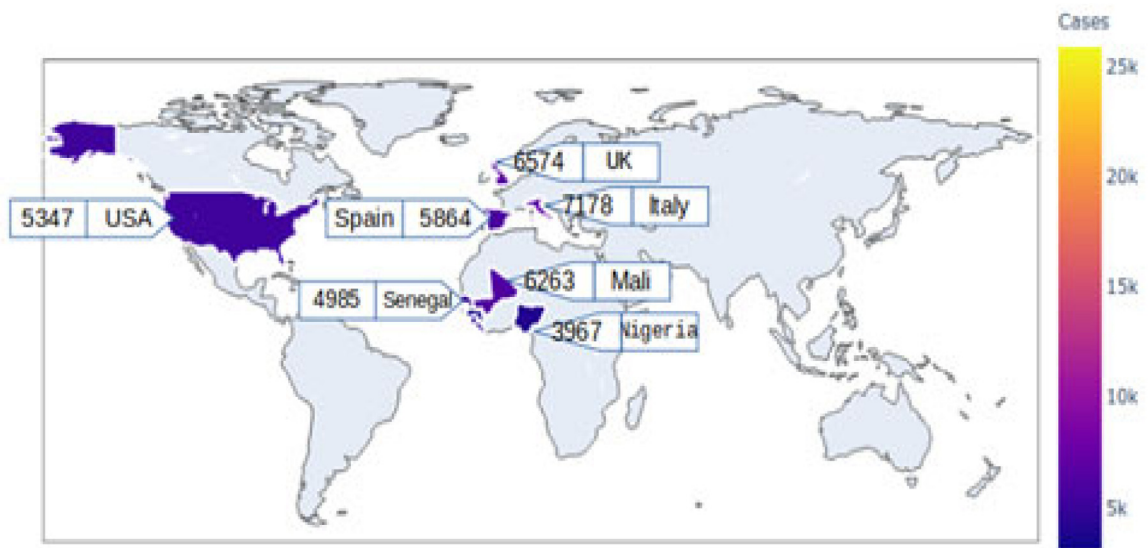


Fig. 9. Ebola spread on the world map.

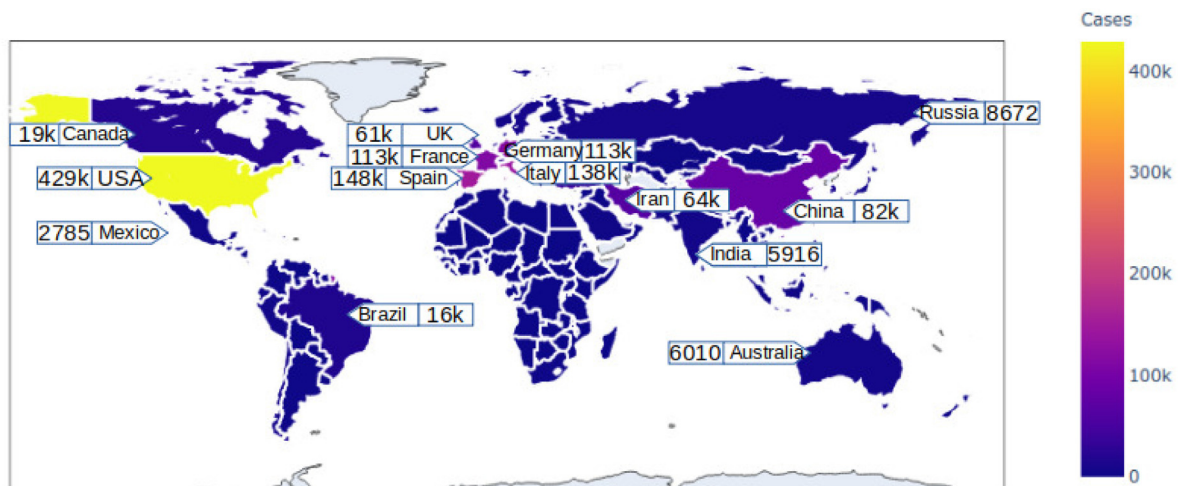


Fig. 10. Corona spread on the world map.

in literature with the same aim but using data obtained during the first months of the disease [27,45,47]. According to a research conducted at local level [46], the accuracy of the proposed model and benchmark models decreases as forecast horizon increases from a few days to 3 weeks. Our model shows accuracy even after 3 weeks working with data at country level. As shown in the experiments, our model performs better than other popular methods for prediction. Focusing on other works using deep learning models, the work [47] that uses also a LSTM based technique to predict confirmed cases of COVID-19 in India, with data from 30th January to 4th April 2020, suffers from the limited data available in the first days as limitation. The data considered in our study extend until October, thus implicitly includes the effects of different restrictions along the different months of the duration of the pandemic. Given that the coronavirus spread is higher than that of other diseases, the number of deaths are expected higher, although its mortality rate is not so high, as compared with other diseases. These results are in line with those obtained by other researchers that concluded that the CFR of COVID-19 is slower than previous coronavirus epidemics [42].

## 5. Conclusions

In this paper, we have compared the present coronavirus outbreaks with the previous epidemics according to the number of infected cases and total number of deaths. We built a deep learning model to learn the current trend of COVID-19 for the prediction of future cases. The proposed deep learning model has been compared with other popular prediction methods showing a lower RMSE. In our predictions, we found out that the numbers of cases are expected to increase rapidly in the next 30 days span. We compared this data (i.e., Covid-19) with the Ebola, MERS, SARS, and h1n1 data. COVID-19 has a huge spread in the initial days itself. Also, the affected countries are far greater than those affected by Ebola, h1n1, SARS, and MERS. Against the 60.8M h1n1 cases worldwide shown in Table 1, the number of cases for COVID-19 is 6.9M only for India. It is expected to rise further if the current trend continues. However, the mortality rate for Ebola was much higher (39.5%) against COVID-19 (1.55%) but the number of deaths for Ebola is lower. This is primarily due to the lesser number of Ebola cases. The present model has some limitations. The given model depends on the data received from the source, which relies on testing done in India. The rate of testing varies with a smaller number of tests conducted in initial days. Hence, the confirmed cases were also less initially. For the model to work perfectly, the data must be accurate. In the future, the model can be updated as the data grows with passing days. Also the proposed method can be applied to other countries to evaluate its performance with different datasets.

## CRedit authorship contribution statement

**Meenu Gupta:** Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Visualization. **Rachna Jain:** Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Visualization. **Soham Taneja:** Conceptualization, Methodology, Formal analysis, Writing - original draft. **Gopal Chaudhary:** Conceptualization, Methodology, Writing - original draft. **Manju Khari:** Conceptualization, Methodology, Validation, Writing - review & editing, Supervision. **Elena Verdú:** Conceptualization, Methodology, Writing - review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Funding:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Compliance with Ethical Standards:

This research does not involve any human or animal participation. All authors have checked and agreed to the submission.

## References

- [1] M.A. Khan, A. Atangana, Modeling the dynamics of novel coronavirus (2019-nCov) with fractional derivative, *Alex. Eng. J.* (2020).
- [2] H1N1 Flu Clinical and Public Health Guidance, Centres for disease control and prevention, 2020, <https://www.cdc.gov/h1n1flu/guidance/>. (Accessed 5 October 2020).
- [3] O.P. Rajoura, R. Roy, P. Agarwal, A.T. Kannan, A study of the swine flu (H1N1) epidemic among health care providers of a medical college hospital of Delhi, *Indian J. Community Med.* 36 (3) (2011) 187.
- [4] C.B. Jonsson, L.T.M. Figueiredo, O. Vapalahti, A global perspective on hantavirus ecology, epidemiology, and disease, *Clin. Microbiol. Rev.* 23 (2) (2010) 412–441.
- [5] V.C. Cheng, S.K. Lau, P.C. Woo, K.Y. Yuen, Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection, *Clin. Microbiol. Rev.* 20 (4) (2007) 660–694.
- [6] J.S.M. Peiris, Y. Guan, K.Y. Yuen, Severe acute respiratory syndrome, *Nature Med.* 10 (12) (2004) S88–S97.
- [7] P.A. Gastanaduy, Update: severe respiratory illness associated with middle east respiratory syndrome coronavirus (MERS-CoV)—worldwide, 2012–2013, *MMWR. Morb. Mortal. Wkly. Rep.* 62 (23) (2013) 480.
- [8] A. Mubarak, W. Alturaiki, M.G. Hemida, Middle east respiratory syndrome coronavirus (MERS-CoV): infection, immunological response, and vaccine development, *J. Immunol. Res.* 2019 (2019).
- [9] S. Ryu, B.C. Chun, K.S. of Epidemiology, An interim review of the epidemiological characteristics of 2019 novel coronavirus, *Epidemiol. Health* 42 (2020).
- [10] World Health Organization (WHO), Coronavirus disease (COVID-19) pandemic, 2020, <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>. (Accessed 14 October 2020).
- [11] National Health Commission of the People's Republic of China, Update on the novel coronavirus pneumonia outbreak, 2020, [http://www.nhc.gov.cn/xcs/yqtb/list\\_gzbd.shtml](http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml) (Accessed 6 April 2020).
- [12] C. Sohrabi, Z. Alsafi, N. O'Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis, R. Agha, World health organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19), *Int. J. Surg.* (2020).
- [13] P.C. Woo, S.K. Lau, C.M. Chu, K.H. Chan, H.W. Tsoi, Y. Huang, B.H. Wong, R.W. Poon, J.J. Cai, W.K. Luk, L.L. Poon, Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia, *J. Virol.* 79 (2) (2005) 884–895.
- [14] A. Rambaut, Phylogenetic analysis | 129 genomes | 6 march 2020, 2020, <http://virological.org/t/phylogenetic-analysis-129-genomes-24-feb-2020/356/>. (Accessed 6 April 2020).
- [15] F. Amanat, D. Stadlbauer, S. Strohmeier, T.H. Nguyen, V. Chromikova, M. McMahon, K. Jiang, et al., A serological assay to detect SARS-CoV-2 seroconversion in humans, *Nature Med.* (2020) 1–4.
- [16] C. Beltran-Pavez, C.L. Marquez, G. Munoz, F. Valiente-Echeverria, A. Gagero, R. Soto-Rifo, G.P. Barriga, SARS-CoV-2 detection from nasopharyngeal swab samples without RNA extraction, *bioRxiv* (2020).
- [17] F.A. Saiz, I. Barandiaran, COVID-19 detection in chest X-ray images using a deep learning approach, *Int. J. Interact. Multimed. Artif. Intell.* 6 (2) (2020) 11–14.
- [18] M. Nour, Z. Cömert, K. Polat, A novel medical diagnosis model for COVID-19 infection detection based on deep features and Bayesian optimization, *Appl. Soft Comput.* (2020) 106580, <http://dx.doi.org/10.1016/j.asoc.2020.106580>.
- [19] M. Gonçalves, D. Agarwal, I. Torre. Díez, Automated medical diagnosis of COVID-19 through efficientnet convolutional neural network, *Appl. Soft Comput.* 96 (2020) 106691.
- [20] J.F.W. Chan, S. Yuan, K.H. Kok, K.K.W. To, H. Chu, J. Yang, F. Xing, J. Liu, C.C.Y. Yip, R.W.S. Poon, H.W. Tsoi, A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster, *Lancet* 395 (10223) (2020) 514–523.

- [21] M.L. Holshue, C. DeBolt, S. Lindquist, K.H. Lofy, J. Wiesman, H. Bruce, C. Spitters, K. Ericson, S. Wilkerson, A. Tural, G. Diaz, First case of 2019 novel coronavirus in the United States, *New Engl. J. Med.* (2020).
- [22] T.M. McMichael, D.W. Currie, S. Clark, S. Pogosjans, M. Kay, N.G. Schwartz, J. Lewis, A. Baer, V. Kawakami, M.D. Lukoff, J. Ferro, Epidemiology of Covid-19 in a long-term care facility in King County, Washington, *New Engl. J. Med.* (2020).
- [23] S.F. Pedersen, Y.C. Ho, SARS-CoV-2: A storm is raging, *J. Clin. Invest.* (2020).
- [24] U. Bastolla, How lethal is the novel coronavirus, and how many undetected cases there are? The importance of being tested, 2020.
- [25] N. Crokidakis, Data analysis and modeling of the evolution of COVID-19 in Brazil, 2020, arXiv preprint arXiv:2003.12150.
- [26] M. Dur-e Ahmad, M. Imran, Transmission dynamics model of coronavirus COVID-19 for the outbreak in most affected countries of the world, *Int. J. Interact. Multimed. Artif. Intell.* 6 (2) (2020) 4, <http://dx.doi.org/10.9781/ijimai.2020.04.001>.
- [27] Z. Zhao, X. Li, F. Liu, G. Zhu, C. Ma, L. Wang, Prediction of the COVID-19 spread in African countries and implications for prevention and control: A case study in South Africa, Egypt, Algeria, Nigeria, Senegal and Kenya, *Sci. Total Environ.* 729 (2020) 138959, <http://dx.doi.org/10.1016/j.scitotenv.2020.138959>.
- [28] N.M. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, G. Cuomo-Dannenburg, Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand, 2020, 2020, <https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-9-impact-of-npis-on-covid-19/> (accessed 14 October 2020).
- [29] C. Fraser, S. Riley, R.M. Anderson, N.M. Ferguson, Factors that make an infectious disease outbreak controllable, *Proc. Natl. Acad. Sci.* 101 (16) (2004) 6146–6151.
- [30] A. Nezir, G. Yurdakul, Assessing countries' performances against COVID-19 via WSIDEA and machine learning algorithms, *Appl. Soft Comput.* 97 (2020) 106792.
- [31] N. Imai, I. Dorigatti, A. Cori, C. Donnelly, S. Riley, N.M. Ferguson, Report 2: Estimating the Potential Total Number of Novel Coronavirus Cases in Wuhan City, China, Imperial College London, 2020.
- [32] S. Lai, I.I. Bogoch, N. Ruktanonchai, A.G. Watts, Y. Li, J. Yu, X. Lv, W. Yang, Y. Hongjie, K. Khan, Z. Li, Assessing spread risk of wuhan novel coronavirus within and beyond China, january-2020: a travel network-based modelling study, 2020.
- [33] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, J. Shaman, Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2), *Science* 368 (6490) (2020) 489–493.
- [34] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, J. Shaman, Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (COVID-19), 2020, medRxiv. <https://doi.org/10.1101/2020.02.14.20023127>.
- [35] S. Pei, S. Kandula, W. Yang, J. Shaman, Forecasting the spatial transmission of influenza in the United States, *Proc. Natl. Acad. Sci.* 115 (11) (2018) 2752–2757.
- [36] S. Pei, J. Shaman, Initial simulation of SARS-CoV2 spread and intervention effects in the continental US. medrxiv, 2020.
- [37] R. Verity, L.C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P.G. Walker, H. Fu, A. Dighe, Estimates of the severity of coronavirus disease 2019: a model-based analysis, *Lancet Infect. Dis.* (2020).
- [38] J.T. Wu, K. Leung, G.M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study, *Lancet* 395 (10225) (2020) 689–697.
- [39] J.T. Wu, K. Leung, M. Bushman, N. Kishore, R. Niehus, P.M. de Salazar, B.J. Cowling, M. Lipsitch, G.M. Leung, Estimating clinical severity of COVID-19 from the transmission dynamics in wuhan, China, *Nature Med.* (2020) 1–5.
- [40] P. Wu, X. Hao, E.H. Lau, J.Y. Wong, K.S. Leung, J.T. Wu, B.J. Cowling, G.M. Leung, Real-time tentative assessment of the epidemiological characteristics of novel coronavirus infections in Wuhan, China, as at january 22 2020, *Eurosurveillance* 25 (3) (2020).
- [41] S.J. Fong, N.Li.G. Dey, R. González-Crespo, E. Herrera-Viedma, Finding an accurate early forecasting model from small dataset: A case of 2019-nCoV novel coronavirus outbreak, *Int. J. Interact. Multimed. Artif. Intell.* 6 (1) (2020) 132–140, <http://dx.doi.org/10.9781/ijimai.2020.02.002>.
- [42] S. Yang, P. Cao, P. Du, Z. Wu, Z. Zhuang, L. Yang, X. Yu, Q. Zhou, X. Feng, X. Wang, W. Li, Early estimation of the case fatality rate of COVID-19 in mainland China: a data-driven analysis, *Ann. Transl. Med.* 8 (4) (2020).
- [43] J. Shaman, W. Yang, S. Kandula, Inference and forecast of the current West African Ebola outbreak in Guinea, Sierra Leone and Liberia, *PLoS Curr.* 6 (2014).
- [44] A. Mahajan, N.A. Sivadas, Solanki. R, An epidemic model SIPHERD and its application for prediction of the spread of COVID-19 infection in India, *Chaos Solitons Fractals* 140 (2020) 110156, <http://dx.doi.org/10.1016/j.chaos.2020.110156>.
- [45] S.I. Alzahrani, I.A. Aljamaan, E.A. Al-Fakih, Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions, *J. Infect. Public Health* 13 (7) (2020) 914–919, <http://dx.doi.org/10.1016/j.jiph.2020.06.001>.
- [46] B. Rostami-Tabar, J.F. Rendon-Sanchez, Forecasting COVID-19 daily cases using phone call data, *Appl. Soft Comput.* 100 (2021) 106932, <http://dx.doi.org/10.1016/j.asoc.2020.106932>.
- [47] A. Tomar, N. Gupta, Prediction for the spread of COVID-19 in India and effectiveness of preventive measures, *Sci. Total Environ.* 728 (2020) 138762, <http://dx.doi.org/10.1016/j.scitotenv.2020.138762>.
- [48] SenPei-CU/COVID-19, Data and code posting, 2020, <https://github.com/datasets/covid-19/tree/master/data>. (Accessed 5 October 2020).
- [49] Imdevskp/ebola\_outbreak\_dataset, Data and code posting, 2020, [https://github.com/imdevskp/ebola\\_outbreak\\_dataset](https://github.com/imdevskp/ebola_outbreak_dataset). (Accessed 5 October 2020).
- [50] Imdevskp/mers\_outbreak\_dataset, Data and code posting, 2020, <https://www.kaggle.com/imdevskp/mers-outbreak-dataset-20122019>. (Accessed 5 October 2020).
- [51] Centers for Disease Control and Prevention (CDC), Coronavirus (COVID-19), 2020, <https://www.cdc.gov/coronavirus/2019-nCoV/index.html>. (Accessed 3 October 2020).
- [52] World Health Organization, 2020. Coronavirus disease 2019 (COVID-19): situation report, 67.
- [53] Centers for Disease Control and Prevention (CDC), Severe acute respiratory syndrome (SARS), 2020, <https://www.cdc.gov/sars/index.html>. (Accessed 5 October 2020).
- [54] World Health Organization, Severe acute respiratory syndrome (SARS), 2020, <https://www.who.int/csr/sars/en/>. (Accessed 5 October 2020).
- [55] 2009 H1N1 Pandemic (H1N1pdm09 virus), Centres for diseases Control and Prevention. Available at: <https://www.cdc.gov/flu/pandemic-resources/2009-h1n1-pandemic.html>.
- [56] K. Khalili-Damghani, F. Abdi, S. Abolmakarem, Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries, *Appl. Soft Comput.* 73 (2018) 816–828.
- [57] M. Thakur, D. Kumar, A hybrid financial trading support system using multi-category classifiers and random forest, *Appl. Soft Comput.* 67 (2018) 337–349.