Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Analysis on novel coronavirus (COVID-19) using machine learning methods

Milind Yadav[a], Murukessan Perumal[b], Dr. M Srinivas[b,*]

[a] Rajasthan Technical University, Kota, India
[b] National Institute of Technology, Warangal, Telangana, India

ARTICLE INFO

Article history:
Received 9 May 2020
Revised 13 June 2020
Accepted 23 June 2020
Available online 30 June 2020

Keywords:
Novel coronavirus
COVID-19
Simple linear regression
Polynomial regression
Support vector regression model
Pearson
Active cases
Recoveries

ABSTRACT

In this paper, we are working on a pandemic of novel coronavirus (COVID-19). COVID-19 is an infectious disease, it creates severe damage in the lungs. COVID-19 causes illness in humans and has killed many people in the entire world. However, this virus is reported as a pandemic by the World Health Organization (WHO) and all countries are trying to control and lockdown all places. The main objective of this work is to solve the five different tasks such as I) Predicting the spread of coronavirus across regions. II) Analyzing the growth rates and the types of mitigation across countries. III) Predicting how the epidemic will end. IV) Analyzing the transmission rate of the virus. V) Correlating the coronavirus and weather conditions. The advantage of doing these tasks to minimize the virus spread by various mitigation, how well the mitigations are working, how many cases have been prevented by this mitigations, an idea about the number of patients that will recover from the infection with old medication, understand how much time will it take to for this pandemic to end, we will be able to understand and analyze how fast or slow the virus is spreading among regions and the infected patient to reduce the spread based clear understanding of the correlation between the spread and weather conditions. In this paper, we propose a novel Support Vector Regression method to analysis five different tasks related to novel coronavirus. In this work, instead of simple regression line we use the supported vectors also to get better classification accuracy. Our approach is evaluated and compared with other well-known regression models on standard available datasets. The promising results demonstrate its superiority in both efficiency and accuracy.

1. Introduction

COVID-19 is an infectious disease caused by a novel coronavirus which has first been originated in Wuhan city, Hubei Provinces of China [1,2]. Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) is a new type of virus family that has not been earlier identified in people. The virus seems to be transmitted mostly through the minute respiratory droplets via coughing, sneezing or when people interact with each other for some time in close proximity. These droplets can then be inhaled, or they can land on surfaces that others may come into touch with, who can then get contaminate when they contact their eyes, mouth, or nose.

The novel coronavirus can live on different surface like few days (stainless steel and plastic) and few hours (cardboard, and copper). However, the amount of viable virus declines over time and may not always be present in sufficient numbers to cause infection. In humans, the symptoms of this virus can be experienced in between 1 to 14 days from the day of infection. From then it has been spreading at the speed of knots, giving no time to prepare against a newly identified infectious and notorious virus which have compelled the WHO to declare COVID-19 as a pandemic [3] due to its fast human to human transmission and people got infected in every continent and it had already taken so many lives. The statistics and graph for increasing cases, active cases have been shown in the Fig. 1. symptoms of coronavirus change in severity from having no symptoms at all (being asymptomatic) to having fatigue, cough, fever, general weakness, sore throat,muscular pain and in the most extreme cases, sepsis, severe pneumonia, acute respiratory distress syndrome,and septic shock, all potentially leading to death. Reports show that clinical deterioration can occur quickly, often during the 14 days of disease. Of late, anosmia loss of the sense of smell have been reported as a one of the symptom of a coronavirus infection. There is already conformation from many regions such as Italy, China, and South Korea that patients with committed SARS-CoV-2 infection have developed anosmia/hyposmia, in some cases in the absence of any other symptoms. Still, there is no proper treatment, drugs or vaccine for coronavirus disease. Several random drugs are being tried

[*] Corresponding author.

https://doi.org/10.1016/j.chaos.2020.110050

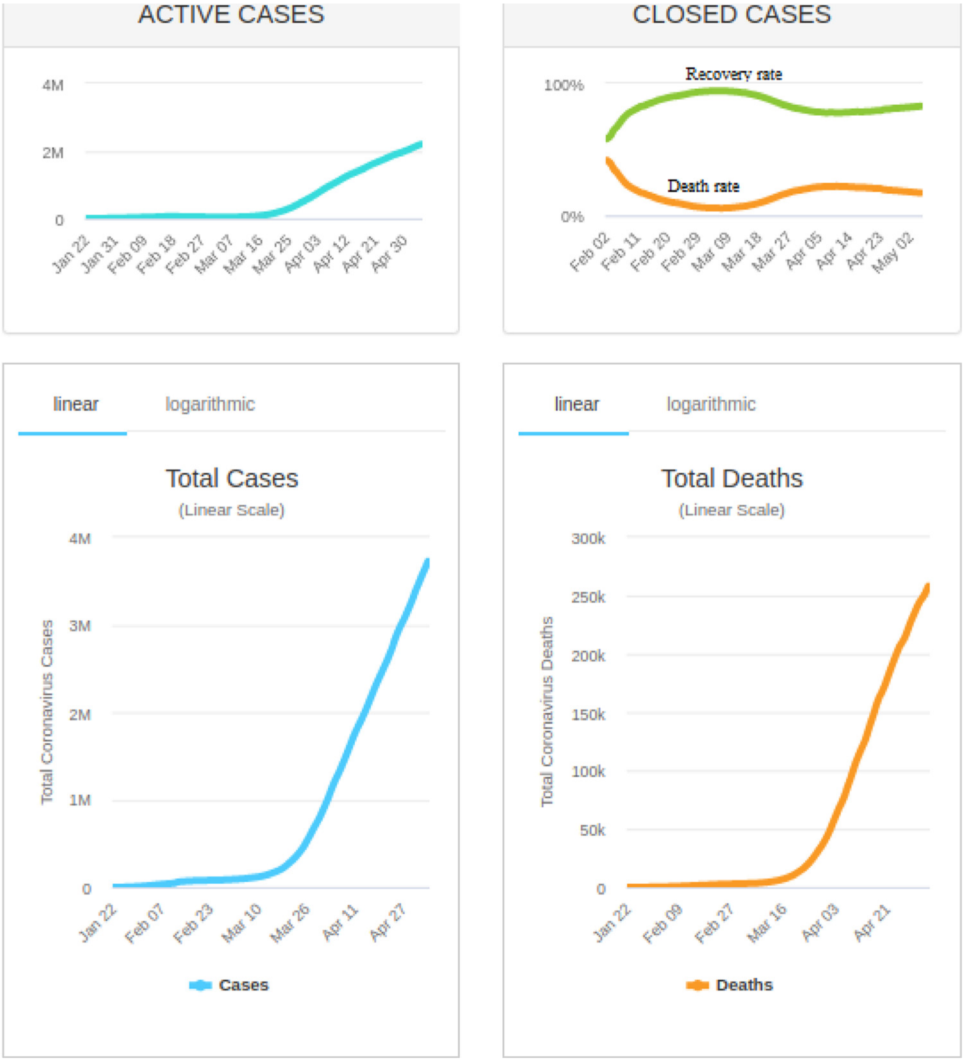


Fig. 1. COVID-19 Statistics published by WHO [4]. (A) Total Number of COVID-19 world-wise effected active cases (B)Closed cases (Death and Recovery rate) weekly wise (C) World-wide total number of COVID-19 cases (D) Total world-wide coronavirus deaths.

to target the virus on severely affected coronavirus patients. However, the use of these need to be more carefully assessed in randomized controlled trials.

Several clinical trials are ongoing to assess their effectiveness but results are not yet available. As this is a new virus, no vaccine is currently available. Although work on a vaccine has already started by several research groups and pharmaceutical companies worldwide, it may be months to more than a year before a vaccine has been tested and is ready for use in humans [5]. Till today COVID-19 has been infected the citizens of more than 212 countries leads to 37,47,356 patients out of which 2,58,970 people had lost their lives and 12,50,693 people [6] gets recovered but due to the virus bi-phasic nature, there may be chances of infection again in those recovered cases.

Due to an insufficient number of test kits, ventilators, oxygen tanks, hospital beds, and unavailability of proper treatment or vaccine, it is very important to analyze the growth rates of positive cases, number of recoveries, and other factors that affect the growth of this virus. To the same extent proper arrangements can be made to prevent losses of lives and to have proper insights of the condition. For example, based on the analysis of data, the government can have the prior information to the number of cases till a particular day, and before that day they can arrange all the necessary medical equipment, or which mitigations to be done to prevent losses of lives. Nowadays, machine learning methods have been widely used in healthcare field [7,8] and for having much faster and efficient prediction of COVID-19 infected person.

In this work, the Support Vector Regression (SVR) [9–11] model is used to solve the four different types of COVID-19 related problems. The proposed method will be fitted into the dataset containing the total number of COVID19 positive cases, and the number of recoveries for different countries like Mainland China, US, Italy, South Korea, and India. And with the help of the proposed method to predict the future number of total cases, active cases, and recoveries. These tasks can help a country/region to understand the spreading of the virus, facilitate/aware people, start mitigations. It'll also help that region/country to be prepared for what's will happen in the future, which may help in saving lives and agony. And compare proposed method results with other well know regression models such as Simple Linear Regression, Polynomial Regression [12]. And another task containing the weather data for regions like New York City(NYC) and Milano (Italy), to analyze the correlation between different weather parameters and the total number of cases Pearson's method is used. This will help in understanding the effects of weather conditions on the virus spread.

2. Proposed Method

This section is describing the motivation and the detailed overview of the tasks and proposed approach. COVID 19 is having dramatic effects among people causing deaths, agony and chaos. Such that to analyze the effects, some tasks can be performed. These tasks can help in understanding and extracting knowledge from COVID 19 data, which can help a country(region) to understand the spreading of the virus, facilitate(aware) people, start mitigations, whether or not mitigation is having some positive effect, other factors affecting the virus, etc. It'll help that country to get prepared for what's coming in the near future, that may help in saving lives and agony. In this paper, mainly five different tasks that were performed [13] and that are as follows:

I) Predicting the spread of corona virus across regions. II) Analysing the growth rates and the types of mitigation across countries. III) Predicting how the epidemic will end. IV) Analysing the transmission rate of virus. V) Correlating the corona virus and weather conditions.

In our proposed work Support Vector Regression (SVR) model is used to work on first four tasks. And Pearson's Correlation [14] method is used for fifth task.

2.1. Support Vector Regression (SVR)

Support vector machines (SVM) [15] is a supervised learning algorithm. This algorithm is used for classification and regression problems. SVR is based on the same principles as SVM for classification i.e. to find a hyperplane in a d-dimensional space (d is the number of features) that uniquely classifies the data points. SVR uses a non-parametric technique, which means, the output from the SVR model does not depend on distributions of the dependent and independent variables. SVR technique is basically dependent on kernel functions, which allows for the construction of a non-linear model without changing the explanatory variables, which helps in better interpretation of the resultant model. In these algorithms, a hyperplane is found that separates the different features. The produced model by SVM does not depend on the training points that lie outside the margin but instead depends on a subset of the training data as the cost function. Similarly, in SVR, support vectors find the closest data points and the actual function represented by them. We get closest to the actual curve if the distance between the support vectors to the regressed curve is maximum.

A hyperplane is a function that classifies the points in a higher dimension or other words hyperplanes are the boundaries that help in the classification of the data points. If the margin for any hyperplane is maximum, then that hyperplane is the optimal hyperplane. The points which are closest to hyperplane are called support vector points and the distance of the vectors from the hyperplane are called the margins, as shown in Fig. 2.

Farther the Support Vector points, from the hyperplane, more is the probability that the points will be correctively classified in their respective region or classes. Thus, the equation of the hyperplane in the d dimension can be given as:

$$\begin{aligned} z &= l_0 + 1_1 x_1 + l_2 x_2 + l_3 x_3 \ldots \\ &= l_0 + \textstyle\sum_{i=1}^{n} l_i x_i \\ &= l_0 + l_1^T x \\ &= b + l_1^T x \end{aligned} \tag{1}$$

where $l_i = \{l_0, l_1, l_2, \ldots\}$, b = biased term (l_0) and x = variables. Kernel is an important part of SVR. The kernel is a way of computing the dot product of two vectors x and y in some high dimensional feature space. Kernel trick is used in SVR which simply means to replace the dot product of two vectors by the kernel function.

3. Experimental Results

We have present the experimental results in detail about each task and at the same time, we have also compared the performance results of our proposed method with the three different well known regression methods such as Simple Linear Regression, Polynomial Regression. The dataset used for the first four tasks consisted of the total number of positive cases, recoveries, deaths from 22/01/2020 (Day 1) to 24/04/2020 (Day 93) in different countries/regions, Country/Region name, and date. However, for all four task dates, region name and, total positive cases information was used. And for these four tasks main regions that were mostly focused on were, Mainland China, US, Italy, South Korea and, India.

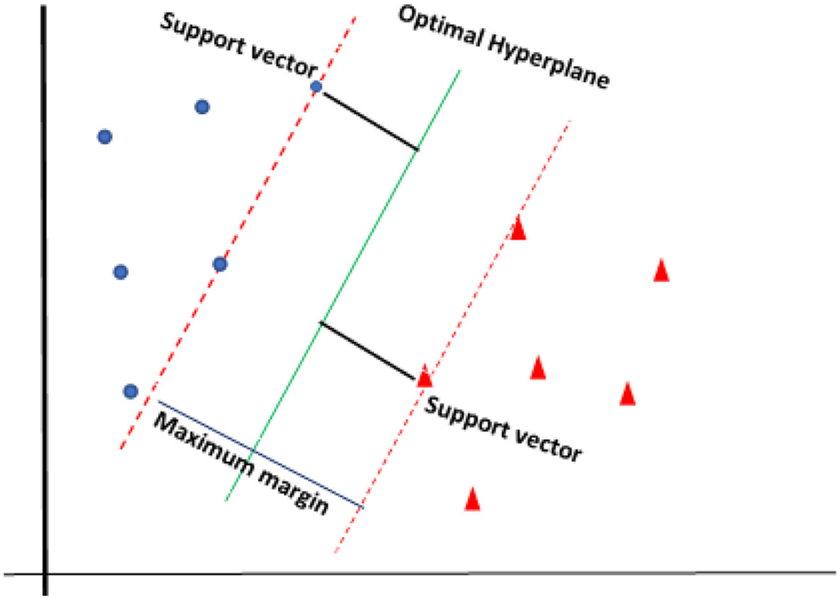


Fig. 2. SVM Model Maximum-margin Hyperplane.

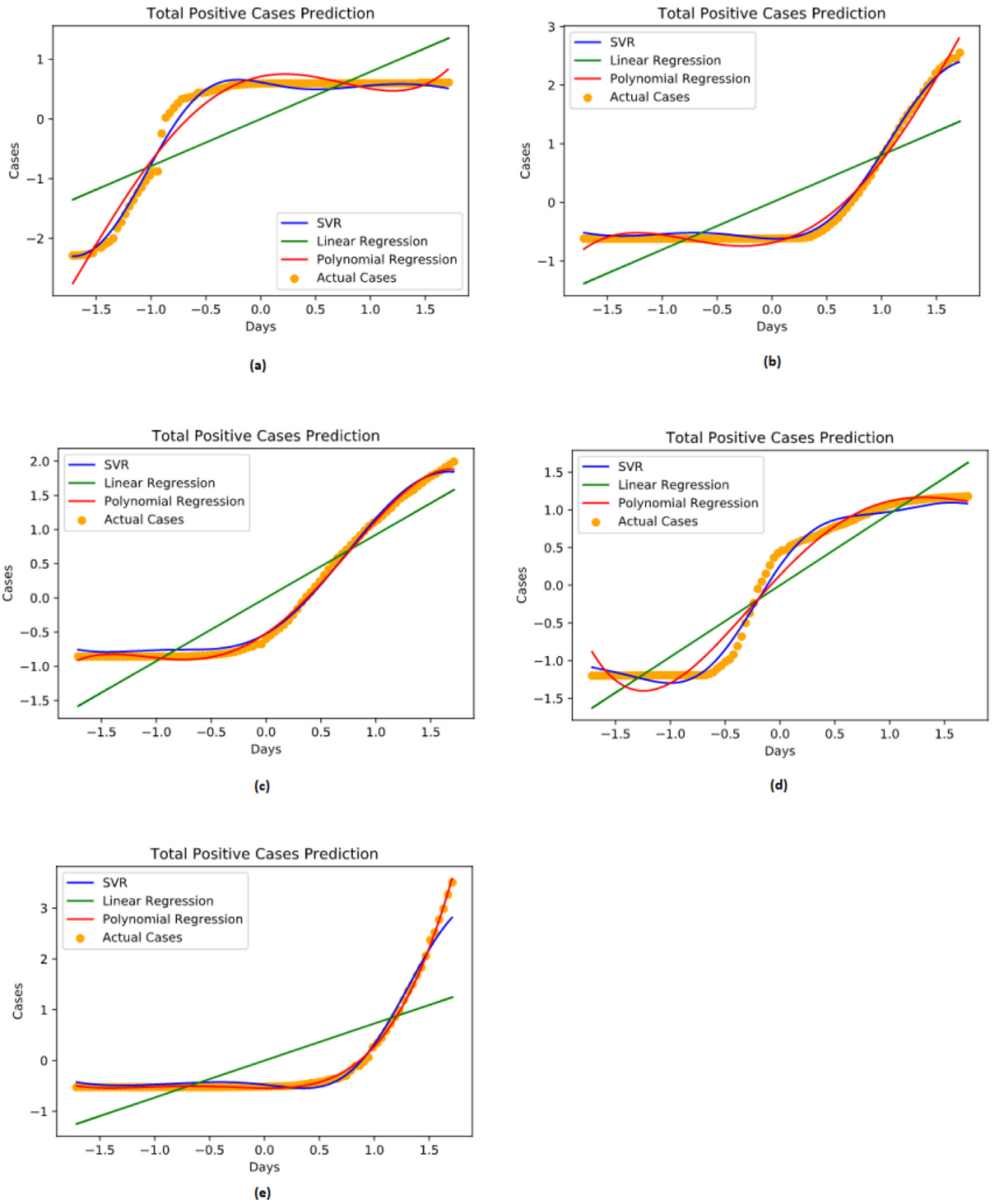
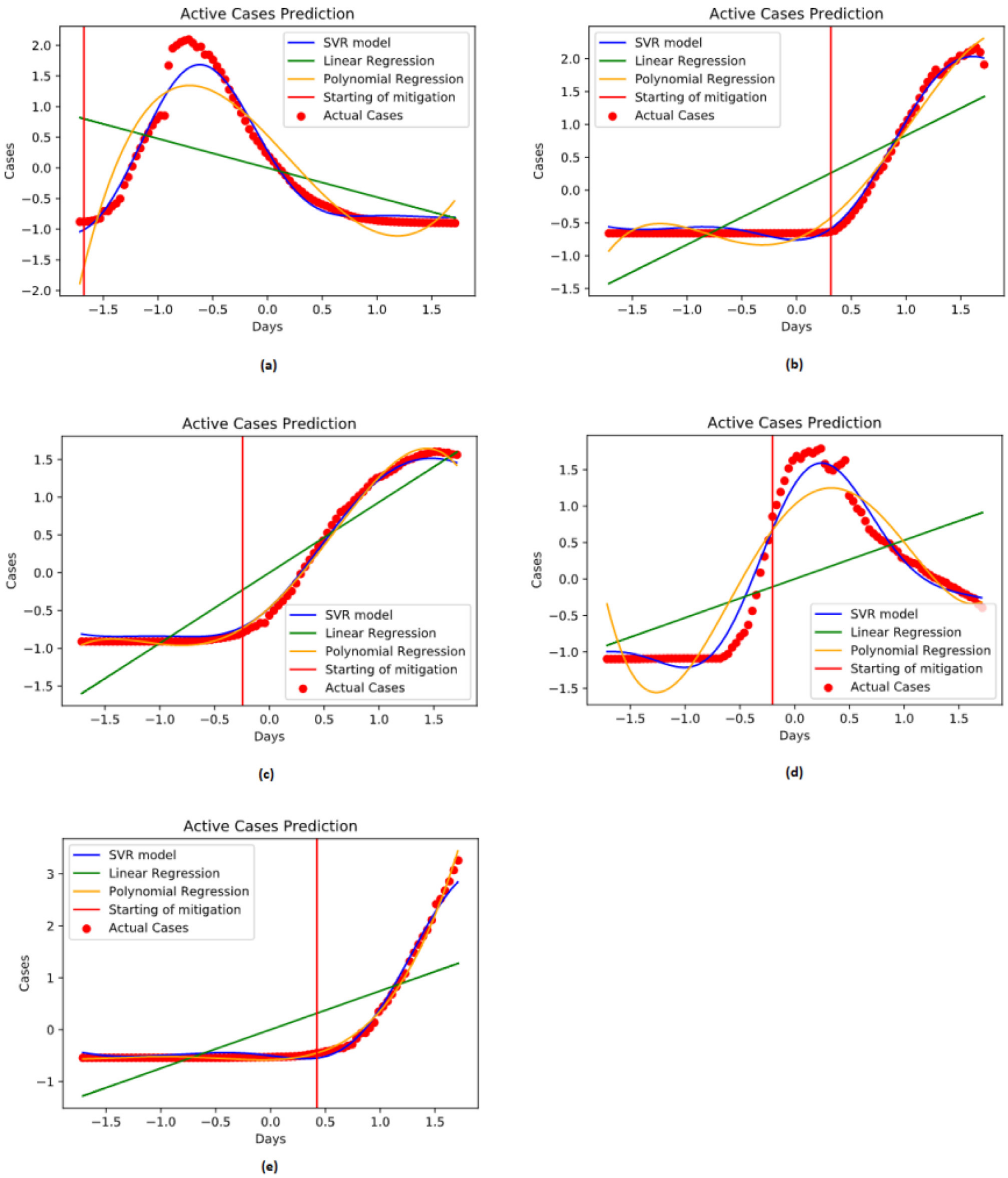


Fig. 3. Accuracy of predicting total number of positive cases in different regions; (a) Mainland China, (b) US, (c) Italy, (d) South Korea, (e) India.

3.1. (I) redicting the spread of corona virus across regions

The outbreak of Covid-19 is developing into a major international crisis, and it's starting to influence important aspects of daily life. For example, Bans have been placed on hotspot countries, international manufacturing operations have often had to throttle back production and many goods solely produced in China have been halted altogether. In highly affected areas, people are starting to stock up on essential goods. Such that, in order to predict how the virus could spread across different countries and regions, different regression models were to be used to predict the total number of positive cases. The main goal of this task was to build and compare regression models that can predict the progression of the total positive COVID 19 cases from different regions, that may help



(a)

(b)

(c)

(d)

(e)

Fig. 4. Predicting total number of active cases in different regions; (a) Mainland China, (b) US, (c) Italy, (d) South Korea, (e) India.

mitigation efforts. The advantage of doing this would be that, we will have an idea about the number that will reach of many cases, this will give the idea about the level of spread, and in accordance to that, the government and the citizens can make proper plans to handle the situation by taking measures to minimize the virus spread by various mitigation and other necessary actions. Predicting total number of cases in different regions are show in Fig. 3.

The accuracy of the total number of positive cases in Mainland China is shown in Fig 3(a). The predicted results for the total number of cases till day 93 with Simple Linear Regression method is 85,492, with the Polynomial Regression method is 73,561, and the proposed SVR method is 65,795, whereas the actual number was 68,128. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 62.1 %, 96.2 %, and 98.8 % respectively. In Fig 3(b) shows the total number of positive cases in the US country. The predicted values for the total number of cases till day 93 with Simple Linear Regression method is 1,71,324, with the Polynomial Regression method is 2,95,006, and the proposed SVR method

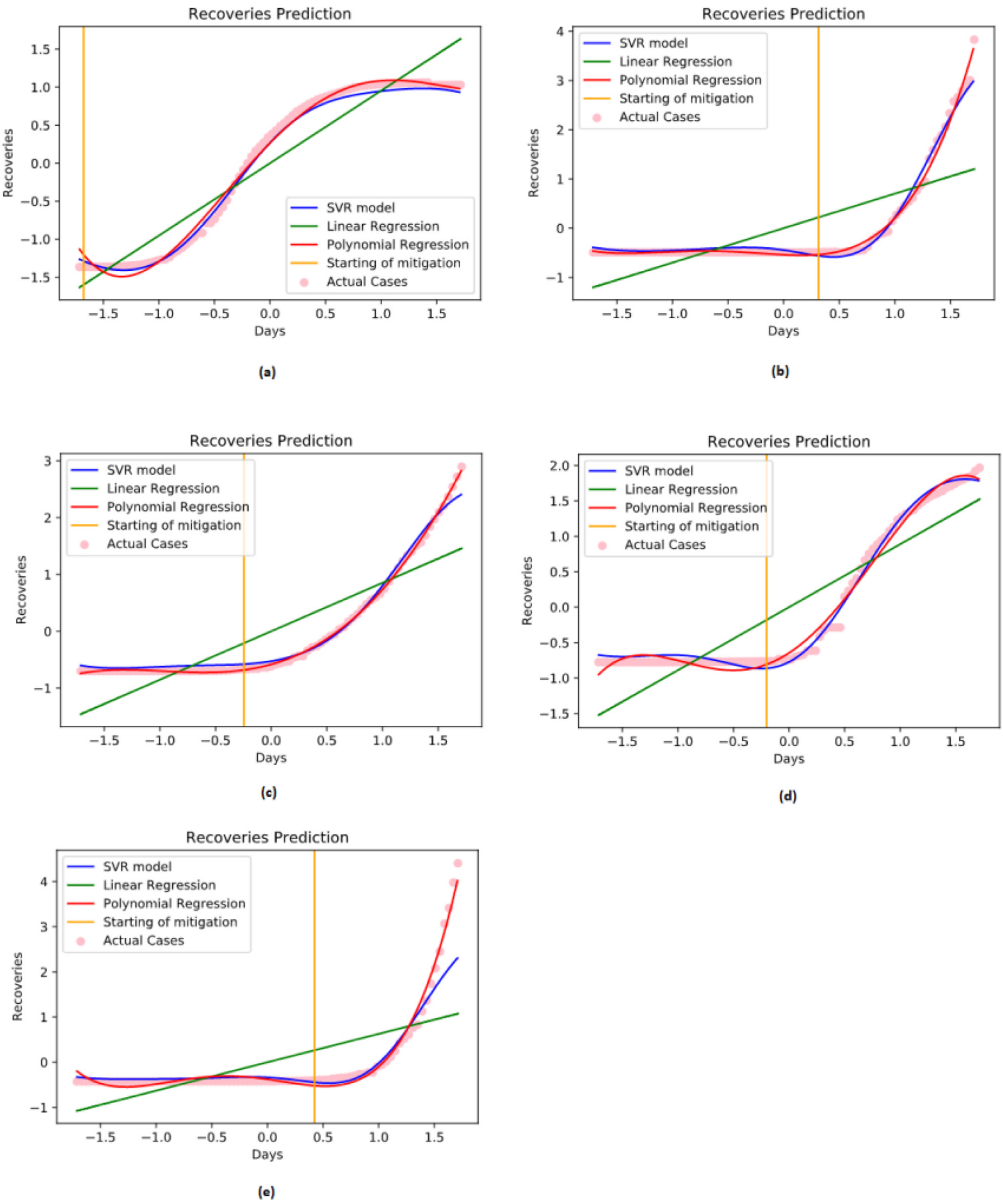


Fig. 5. Predicting recoveries in different regions; (a) Mainland China, (b) US, (c) Italy, (d) South Korea, (e) India.

is 2,58,253, whereas the actual number was 2,71,590. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 65.01 %, 98.82 %, and 99.47 % respectively.

The total number of predicted positive cases in Italy country is shown in Fig 3(c). The predicted values for the total number of cases till day 93 with Simple Linear Regression method is 1,65,154, with the Polynomial Regression method is 1,85,188 and the proposed SVR method is 1,83,007, whereas the actual positive cases was 1,92,994. The accuracy result for Simple Linear Regression, Polynomial Regression, and SVR were 85.36 %, 99.75 % and 99.41 % respectively. In Fig 3(d) shows the total number of positive cases in the South Korea country is plotted. The predicted values for the total number of cases till day 93 with Simple Linear Regression method is 12,733, with the Polynomial Regression method is

10,421 and the proposed SVR method is 10,266, whereas the actual positive cases was 10,718. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 90.10 %, 97.28 % and 99.06 % respectively. In Fig 3(e) shows the total number of positive cases in the Indian country is plotted. The predicted values for the total number of cases till day 93 with Simple Linear Regression method is 10,789, with the Polynomial Regression method is 25,165 and the proposed SVR method is 20,373, whereas the actual number positive cases was 24,530. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 53.04 %, 99.85 % and 98.31 % respectively.

3.2. (II) Analysing the growth rates and the types of mitigation across countries

Over the last few months different countries have performed different forms of mitigation to prevent the spread of COVID 19. These mitigations involved like the ban of large gatherings, closing of schools, banned flights, and other transportation, put cities in lockdown, etc. Such that, in order to see the effects of mitigation, this task was performed. The main objective of the task was to evaluate the effectiveness of mitigation by trying to see if a correlation can be discovered between the different types of mitigation and the growth rate of active cases. Like what measures seem to work, which not and Which ones are the most effective, Keeping in mind that there are numerous factors which might affect the growth rate (e.g. country's general hygiene, population density, how much time they had to prepare for this epidemic before the spread was not much, etc.). In order to predict how the mitigation was affecting the number of active cases across different countries and regions, different regression models, were used to predict the number of active cases in a particular region. The goal of solving this task helps us to understand which mitigations are most effective. The advantage of doing this would be that, we will have an idea about how well the mitigations are working, and the actions that are taken till date, how effective are they, or how many cases have been prevented by this and so on.

To prepare the dataset for this task firstly, dates were converted to day number taking 22 Jan 2020 as Day 1 and 24 April 2020 as Day 93. Then, total cases were extracted region wise in order of day number. To see growth rates, information for the number of active cases were needed. The active cases for every day were calculated by subtracting the total number of deaths and recoveries from the total case count day wise. After this the day number and active case count were scaled to observe clearer results. For different affected countries, a linear regression and a polynomial regression model was fitted and visualized to the dataset to predict the number of active cases in that particular region. In the polynomial regression model, the model was fitted with different polynomial degrees to find which degree curve fits best. Then, in an urge to seek for better results a proposed SVR model with Radial basis function kernel (RBF) was fitted and visualized to predict the number of active cases in that particular region. Then, the accuracy of all the models was calculated and compared using the Coefficient of Determination to see how well each model predicts results.

Fig. 4 shows the total number of active cases of COVID 19 of that particular region, and is plotted with a number of active cases on Y-axis and number of days on X-axis. All the values of X-axis and Y-axis were scaled before use.

In Fig 4(a) shows the growth rate of the total number of active cases in Mainland China region. China started its complete lockdown from early january, and the total number of active cases of Mainland China is decreasing at a satisfying rate. The predicted values for the total number of cases till day 93 with Simple Linear Regression method is 1414, with the Polynomial Regression method is 6371 and the proposed SVR method is 1706, whereas the actual number active cases was 23. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 80 %, 85.46 % and 97.01 % respectively.

The growth rate of the total number of active cases in US country is shown in Fig 4(b). US haven't opted complete lockdown strategy, but instead partially closed schools and other gathering places and promoted social distancing and personal hygiene among citizens (mitigation started around 19th March 2020), and as shown in the figure the total number of active cases of US is not seen to be lowering down. The predicted values for the total number of cases till day 93 with Simple Linear Regression method is 1,22,606, with the Polynomial Regression method is 1,74,969 and the proposed SVR method is 1,56,954, whereas the actual number active cases was 1,51,100. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 69.07 %, 97.82 % and 99.46 % respectively.

In Fig 4(c) shows the growth rate of total number of active cases in Italy is plotted. Italy started its complete lockdown from 9th March 2020, and as seen in this figure the total number of active cases of Italy is not decreasing. But the curve has started to flatten at the top, which is a good sign, as the growth rate of active cases has started to decrease. The predicted values for the total number of cases till day 93 with Simple Linear Regression method is 1,08,114, with the Polynomial Regression method is 1,00,262 and the proposed SVR method is 1,01,910, whereas the actual number active cases was 1,06,527. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 87.22 %, 99.56 % and 99.51 % respectively. The growth rate of total number of active cases in South Korea is shown is Fig 4(d). South Korea started its lockdown from early March, and as seen in this figure the total number of active cases of South Korea is decreasing at a satisfying rate. The predicted values for the total number of cases till day 93 with Simple Linear Regression method is 5270, with the Polynomial Regression method is 2064 and the proposed SVR method is 2200, whereas the actual number active cases was 1843. The accuracy for Simple Linear, Polynomial Regression and SVR were 28.2%, 85.3% and 96.76% respectively.

In Fig 4(e) shows the growth rate of the total number of active cases in India is plotted. India started its complete lockdown from 24th March 2020. It is clear that the total number of active cases of India is not decreasing, but despite such a dense population, the condition of India is better than most other countries. The predicted values for the total number of cases till day 93 with Simple Linear Regression method is 8733, with the Polynomial Regression method is 19,248 and the proposed SVR method is 16,283, whereas the actual number of active cases was 18,252. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 55.5 %, 99.58 % and 99.6 % respectively.

3.3. (III)Predicting how many infected patients will recover

Since no vaccine for the virus has yet been discovered, it is important to see how many of the patients will recover from this virus, and how and when the epidemic will end. The main objective of this task was to predict how many people were going to recover based on old recovery records.

Such that in order to predict how many people will actually recover, records for the number of recovered patients across different countries were taken. The goal of solving this task helps us to understand how the epidemic will end. i.e. how many patients will recover. The advantage of doing this would be that, we will have an idea about the number of patients that will recover from the infection, from the older known methods, since no vaccine or cure is yet discovered. And by predicting the time that will be taken by all the patients to recover, we will be able to understand how much time will it take to for this pandemic to end.

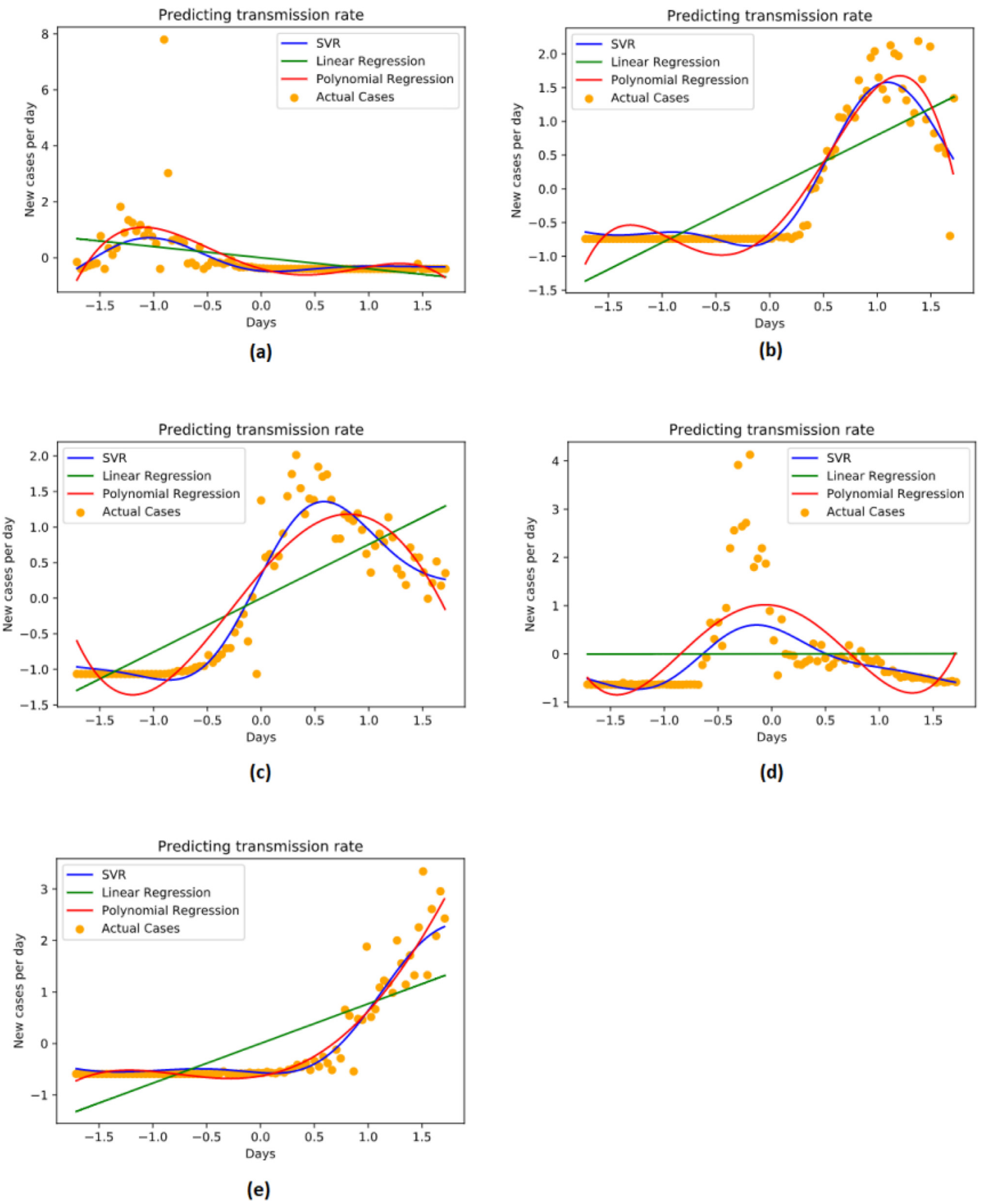


Fig. 6. Result Analysis of transmission of COVID-19 in (a) Mainland China, (b) US, (c) Italy, (d) South Korea, (e) India.

To prepare the dataset for this task firstly, dates were converted to day number taking 22 Jan 2020 as Day 1 and 24 April 2020 as Day 93. Then, total number recoveries till that day were extracted region wise in order of day number. Next, the day number and recovery counts were scaled to observe clearer results. For different affected countries, a linear regression and a polynomial regression model was fitted and visualized to the dataset to predict the number of recoveries of affected patients in that particular region. In the polynomial regression model, the model was fitted with different polynomial degrees to find which degree curve fits best. Then, in the urge to seek for better results an SVR model with Radial basis function kernel (RBF) was fitted and visualized to predict the number of recoveries in that particular region. Then, the accuracy of all the models was calculated and compared using the Coefficient of Determination to see how well each model predicts results.

Fig. 5 shows the growth rate of the number of recovered patients from COVID 19 of that particular region, and is plotted with a number of recoveries on the *Y*-axis and number of days on the *X*-axis.

In Fig 5(a)shows the total number of recoveries in Mainland China country. The curve is dropping after a certain point because most of the affected people are either already recovered or dead, at that point of time. The predicted values for the total recoveries till day 93 with Simple Linear Regression method is 79,606, with the Polynomial Regression method is 62,232 and the proposed SVR method is 60,951, whereas the actual number of active cases was 63,593. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 90.92%, 99.32% and 99.47% respectively. The total number of recoveries in US country is shown in Fig 5(b). The rate of recoveries is mostly constant in the entire curve, but rises near the end. The predicted values for the total recoveries till day 93 with Simple Linear Regression method is 38,817, with the Polynomial Regression method is 96,164 and the proposed SVR method is 80,101, whereas the actual number of active cases was 99,079. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 49.14%, 99.30% and 98.54% respectively.

In Fig 5(c) shows the total number of recoveries in Italy region is plotted. The number of recoveries is increasing every day. The predicted values for the total recoveries till day 93 with Simple Linear Regression method is 36,324, with the Polynomial Regression method is 59,536 and the proposed SVR method is 52,309, whereas the actual number of recoveries was 60,498. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 72.6%, 99.92% and 99% respectively.

The total number of recoveries in South Korea country is shown Fig 5(d). The rate of recovering patients are decreasing after reaching local maxima because most of the patients are already recovered or dead. The predicted values for the total recoveries till day 93 with Simple Linear Regression method is 7238, with the Polynomial Regression method is 8118 and the proposed SVR method is 8054, whereas the actual number of recoveries was 8635. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 78.9%, 98.75% and 99.27% respectively.

In Fig 5(e) shows the total number of recoveries in India is plotted. The number of recovering patients is increasing. The predicted values for the total recoveries till day 93 with Simple Linear Regression method is 1710, with the Polynomial Regression method is 5107 and the proposed SVR method is 3122, whereas the actual number of recoveries was 5498. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 39.5%, 98.36% and 86.39% respectively. Almost all regions proposed method given best prediction accuracy except india region due to less number of samples.

3.4. (IV) Analysing the transmission rate of virus

The COVID19 virus is spreading at immense rates among humans all around the globe. Since, no vaccine for the virus is yet discovered, so it is important to understand how the virus is transmitting i.e. how fast or how slow the virus is spreading among different countries. In order to predict how many persons are infected each day, records for the number of total positive cases across different countries were taken. The goal of solving this task helps us to understand in which countries the transmission is faster or slower.

The advantage of doing this would be that, we will be able to observe and analyze how fast or slow the virus is spreading among regions therefore, which areas needs more attention or not. To prepare the dataset for this task firstly, dates were converted to day number taking 22 Jan 2020 as Day 1 and 24 April 2020 as Day 93. Then, the total number of newly found cases per day was calculated by subtracting the total number of cases a day before to the total number of cases on the present day and, we get the number of newly found cases per day. Then, the day number and new cases per day were scaled to observe clearer results.

Fig. 6 shows different graphs that represent the rate of transmission of COVID19 virus among people of different regions, and is plotted with the number of newly found cases on each day on the *Y*-axis and number of days on the *X*-axis.

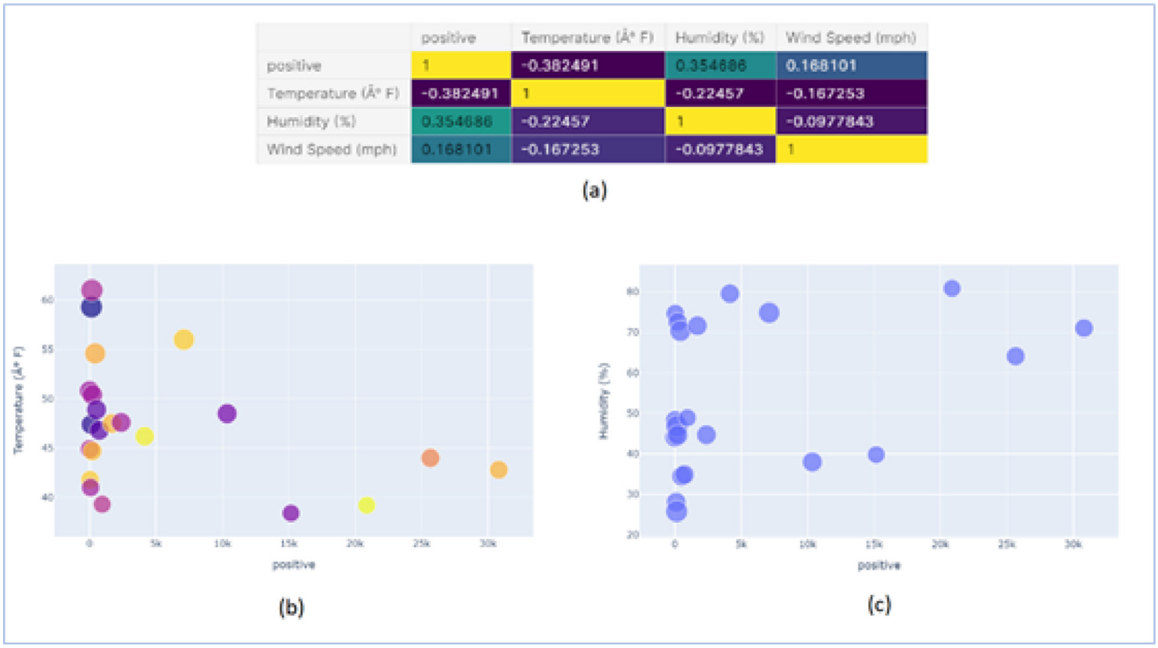
In Fig 6(a) shows the total number of newly found cases per day in Mainland China. The curve is dropped to zero after a certain point because according to the data, no new cases are now found in China. The predicted values for the newly found cases for day 93 with Simple Linear Regression method is −505, with the Polynomial Regression method is −586 and the proposed SVR method is 128, whereas the actual number of newly found cases was 0. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 15.7%, 34.05% and 37.9% respectively. The total number of newly found cases per day in US region is shown in Fig 6(b). The curve is rose to local maxima and is seen to be dropping with the number of days. The predicted values for the newly found cases for day 93 with Simple Linear Regression method is 8214, with the Polynomial Regression method is 3569 and the proposed SVR method is 4560, whereas the actual number of newly found cases was 8130. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 63.4%, 89.93% and 92.1% respectively.

In Fig 6(c) shows the total number of newly found cases per day in Italy is plotted. The curve is rose to local maxima and is seen to be dropping with the number of days. The predicted values for the newly found cases for day 93 with Simple Linear Regression method is 5030, with the Polynomial Regression method is 1920 and the proposed SVR method is 2831, whereas the actual number of newly found cases was 3021. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 57.33%, 83.51% and 91.51% respectively. The total number of newly found cases per day in South Korea is shown in Fig 6(d). The curve once rose to local maxima for limited time but dropped since then and did not rise with the increment of days. The predicted values for the newly found cases for day 93 with Simple Linear Regression method is 114, with the Polynomial Regression method is 121 and the proposed SVR method is 8, whereas the actual number of newly found cases was 10. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 0.0006%, 43.2% and 43.7% respectively.

In Fig 6(e) shows the total number of newly found cases per day in India is plotted. The curve keeps on rising to attain local maxima. The predicted values for the newly found cases for day 93 with Simple Linear Regression method is 920, 1643 with the Polynomial Regression method is and the proposed SVR method is 1378, whereas the actual number of newly found cases was 1453. The accuracy for Simple Linear Regression, Polynomial Regression, and SVR were 59.4%, 91.1% and 91.1% respectively.

3.5. (V) Correlating the corona virus and weather conditions

The weather conditions that affect the spread of COVID 19 is not the same for all regions. Such that for different regions, it is important to analyze which weather conditions mostly affect the spread. The main objective of this task was to get data regarding the count of infected people of New York City and Milan city (Italy) date wise with temperature and then analyze it with humidity perception and wind. The main motive was to observe if these factors



	positive	Temperature (Â° F)	Humidity (%)	Wind Speed (mph)
positive	1	-0.382491	0.354686	0.168101
Temperature (Â° F)	-0.382491	1	-0.22457	-0.167253
Humidity (%)	0.354686	-0.22457	1	-0.0977843
Wind Speed (mph)	0.168101	-0.167253	-0.0977843	1

Fig. 7. Result Analysis of different correlation information of New York City (a) correlation table for temperature, humidity, wind speed, and total positive COVID19 cases, (b) correlation between temperature and total positive cases, (c) correlation between total positive cases and humidity.

contribute to the spread in these cities, although the effects may be tiny but one cannot ignore the effects.

The advantage of doing this would be that we'll be able to create better surroundings for the infected patient to reduce the spread. The people can also be warned if they should avoid humidity or not, or high temperature or not, etc. To understand the correlation between the spread and weather conditions Pearson's correlation method is using.

3.5.1. *Pearson's correlation*

Correlation is a measure of association between two variables and the direction of their relationship. The value of the correlation is always lesser than +1 and greater than −1. A value +1 or −1, means a perfect correlation between two variables. If the coefficient approaches 0, the relationship between the two variables becomes weaker. '+' sign indicates a positive relationship and '-' sign indicates a negative relationship. Most used types of correlations: Pearson correlation, Kendall rank correlation, Spearman correlation, and the Point-Biserial correlation.

Pearson correlation: It is the measure of the degree of the relationship between linearly related variables. Pearson correlation is the most widely used correlation. For the Pearson correlation, both the variables whose correlation is to be found are assumed to be normalized, if not normalized, then normalization should be performed first. Also, the relationship between both the variables should be a straight line, assuming that data is equally distributed about the regression line. Following is the formula which is used to calculate the Pearson r correlation:

$$r = \frac{N\sum ab - (\sum a)(\sum b)}{\sqrt{[N\sum a^2 - (\sum a)^2][N\sum b^2 - (\sum b)^2]}} \tag{2}$$

Where, r is Pearson r correlation coefficient between a, b, N is number of observations, a indicate value of x and b is a value of y.

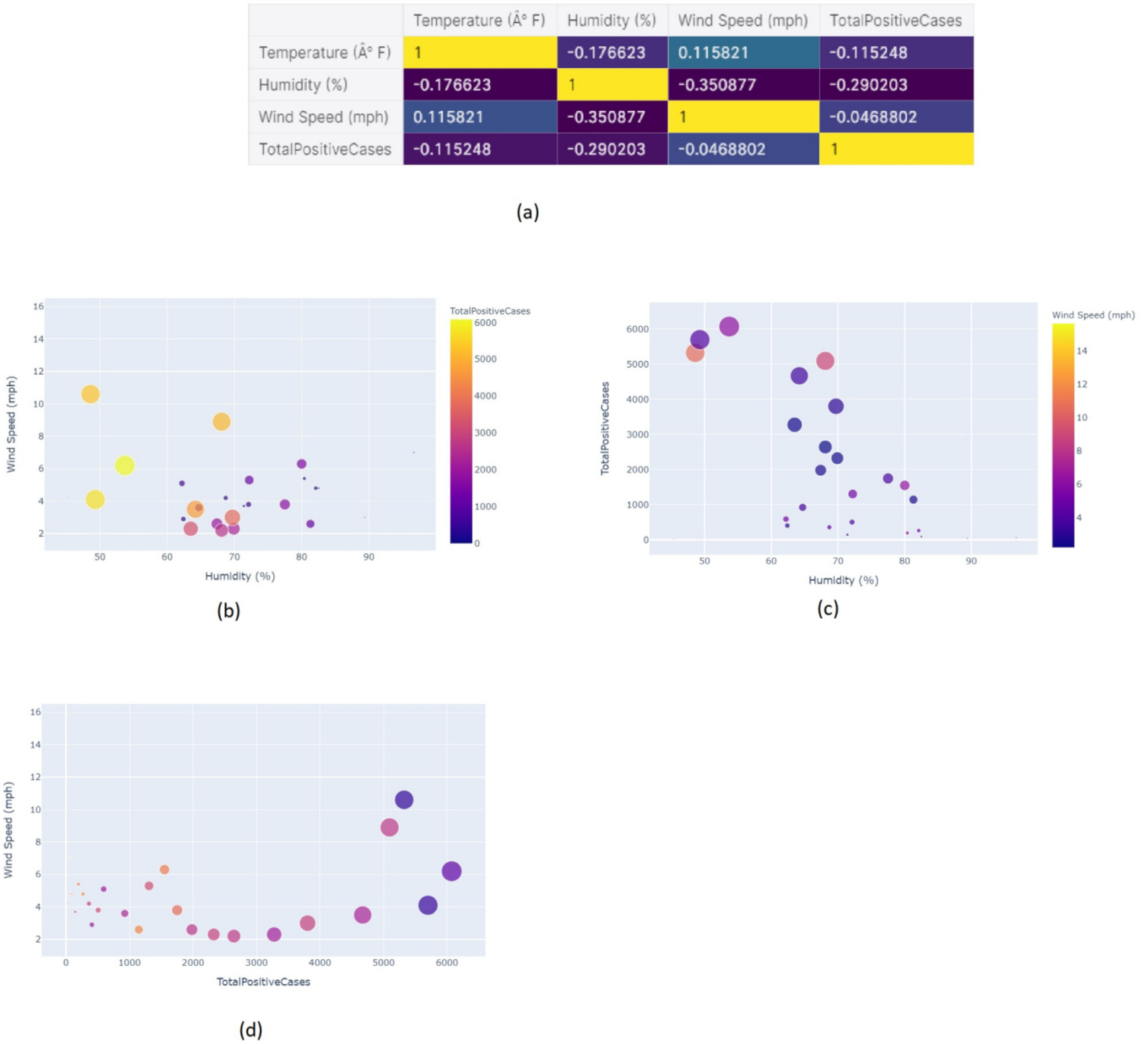
To perform this task weather data for New York City (US) and Milan (Italy) were scraped from a weather website [16,17], which consisted of Wind Speed, Humidity, and Temperature for the month of March. Then the total number of cases in these two cities was plotted on a graph for visualization. Then using Pearson method correlation was found between Wind Speed, Humidity, Temperature, and Total Covid-19 Cases. Then, to visualize the effects of correlated weather conditions and total positive cases, different graphs were plotted. Fig. 7 represents the correlation information of New York City and Fig. 7(a) shows the correlation table for temperature, humidity, wind speed, and total positive COVID19 cases in New York City. It is observed from the correlation chart, the number of positive cases is mostly affected by Temperature and Humidity with relatively high modulus of coefficient value as 0.382491 and 0.354686 respectively.

Fig. 7 (b) is a graph that shows how temperature and total positive cases in NYC affect each other. And it is observed that the number of positive cases decreases with an increase in temperature.

Fig. 7 (c) is a graph that shows how humidity and total positive cases in NYC affect each other. And it is observed that the number of positive cases increases with an increase in humidity.

Fig. 8 represents the correlation information of Milan city and Fig. 8(a) shows the correlation table for temperature, humidity, wind speed, and total positive COVID19 cases in Milan city. It is observed from the correlation chart the number of positive cases is mostly affected by Humidity, as the modulus of coefficient value is 0.290203. And for wind speed and humidity modulus of the correlation coefficient is 0.350877. In Fig. 8(b) a correlation graph between humidity and wind Speed in Milan city, is plotted. In this figure dark coloured bubbles show lower number of positive cases, whereas lighter bubbles represent comparatively higher number of positive cases. It is observed that, despite the small correlation that exists between these variables it is found that the number of affected cases increases per day while humidity decreases. In Fig. 8(c) a correlation graph between total positive cases and humidity in Milan city, is plotted. In this figure, dark-colored bubbles show lower wind speed, whereas lighter bubbles represent comparatively higher wind speed. It is observed that as the number of positive cases increases and the humidity decreases; the wind speed gradually increases with the number of positive cases.

In Fig. 8(d) a correlation graph between total positive cases and wind speed in Milan city, is plotted. It is observed that, wind speed and total positive cases have a nonlinear shape.



	Temperature (Â° F)	Humidity (%)	Wind Speed (mph)	TotalPositiveCases
Temperature (Â° F)	1	-0.176623	0.115821	-0.115248
Humidity (%)	-0.176623	1	-0.350877	-0.290203
Wind Speed (mph)	0.115821	-0.350877	1	-0.0468802
TotalPositiveCases	-0.115248	-0.290203	-0.0468802	1

(a)

(b)

(c)

(d)

Fig. 8. Result Analysis of different correlation information of Milan city (a) correlation table for temperature, humidity, wind speed, and total positive COVID19 cases, (b) correlation between humidity and wind Speed, (c) correlation between total positive cases and humidity, (d) correlation graph between total positive cases and wind speed.

4. Conclusion

COVID-19 causes illness in humans and creates severe damage in the lungs. However, COVID-19 has killed many people in the entire world. In this paper, we are proposing the Support Vector Regression method based navel coronavirus analysis on five different tasks. Main novelty in this work is instead of simple regression line we use supported vectors also to get better classification accuracy.

The main advantage of doing the first task would be that, this will give the idea about the level of spread, and in accordance to that, the government and the citizens can make proper plans to handle the situation by taking measures to minimize the virus spread by various mitigation and other necessary actions. With the help of second task will have an idea about how well the mitigation's are working, and the actions that are taken till date how effective are they, or how many cases have been prevented by this. The advantage of doing third task would be that, we will have an idea about the number of patients that will recover from the infection, from the older known methods, since no vaccine or cure is yet discovered. And by predicting the time that will be taken by all the patients to recover, we will be able to understand how much time will it take to for this pandemic to end. With fourth task we will be able to observe and analyze how fast or slow the virus is spreading among regions therefore, which areas needs more attention or not. Finally, Fifth task create better surroundings for the infected patient to reduce the spread. The people can also be warned if they should avoid humidity or not and high temperature or no. Pearson's correlation method gives a clear understanding of the correlation between the spread and weather conditions. In all tasks, the proposed Support Vector Regression method based coronavirus analysis given promising results compared with other well know regression methods on the first four tasks.

Credit Author Statement

Author Agreement Statement We the undersigned declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We understand that the Corresponding Author is the sole contact for the Editorial process. He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs

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.

Acknowledgement

The authors would like to thank the anonymous reviewers for their thorough review and valuable comments. This work was supported in part by grants Government of India, Ministry of Human Resource Development and NIT Warangal under NITW/CS/CSE-RSM/2018/908/3118 project.

References

[1] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020.

[2] Tan W., Zhao X., Ma X., Wang W., Niu P., Xu W., et al. A novel coronavirus genome identified in a cluster of pneumonia cases–Wuhan, China 2019–2020China CDC Weekly2020; 2(4):61-62.

[3] Gorbalenya AE. Severe acute respiratory syndrome-related coronavirus–the species and its viruses, a statement of the coronavirus study group. BioRxiv 2020.

[4] Confirmed cases and deaths by country, territory, or conveyance2020a;http://www.worldometers.info/coronavirus/.

[5] mode of transmission by country, person infectious2020b; https://www.ecdc.europa.eu/en/covid-19/questions-answers.

[6] Confirmed cases country, territory, or conveyance2020;www.who.int/docs/default-source/coronaviruse/situation-reports/20200327-sitrep-67-covid-19.pdf?sfvrsn=b65f68eb_4.

[7] Srinivas M, Lin Y-Y, Liao H-YM. Deep dictionary learning for fine-grained image classification. In: 2017 IEEE International conference on image processing (ICIP). IEEE; 2017. p. 835–9.

[8] Srinivas M, Mohan CK. Classification of medical images using edge-based features and sparse representation. In: 2016 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE; 2016. p. 912–16.

[9] Awad M., Khanna R.. Support vector regression. efficient learning machines. 2015.

[10] Wu C-H, Ho J-M, Lee D-T. Travel-time prediction with support vector regression. IEEE Trans Intell Transp Syst 2004;5(4):276–81.

[11] Smits GF, Jordaan EM. Improved SVM regression using mixtures of kernels. In: Proceedings of the 2002 international joint conference on neural networks. IJCNN'02 (Cat. No. 02CH37290), vol. 3. IEEE; 2002. p. 2785–90.

[12] Ostertagová E. Modelling using polynomial regression. Procedia Eng 2012;48:500–6.

[13] Covid-19 challenge tasks2020c;https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset/tasks.

[14] Sedgwick P. Pearsons correlation coefficient. BMJ 2012;345:e4483.

[15] Müller K-R, Smola AJ, Rätsch G, Schölkopf B, Kohlmorgen J, Vapnik V. Predicting time series with support vector machines. In: International conference on artificial neural networks. Springer; 1997. p. 999–1004.

[16] For data extraction2020a;https://www.kaggle.com/lumierebatalong/tutorial-extract-data-from-html-file-using-pandas.

[17] For weather2020b;https://www.wunderground.com/.