Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

An Approach to Forecast Impact of COVID-19 Using Supervised Machine Learning Model

Senthilkumar Mohan[1] | John A[2] | Ahed Abugabah[3] | Adimoolam M[4] | Shubham Kumar Singh[5] | Ali kashif Bashir[6] | Louis Sanzogni[7]

[1]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore 632014, India
[2]Department of Computer Science & Engineering,Galgotias University, Noida (India) Email: johnmtech@gmail.com
[3]College of Technological Innovation, Zayed University,Dubai, United Arab Emirates,Email: ahed.abugabah@zu.ac.ae
[4]Saveetha school of Engineering, Chennai (India) Email: m.adimoolam@gmail.com
[5]Indiana University, Bloomington (USA) Email: ksingh.shubh@gmail.com
[6]Department of Computing and Mathematics, Manchester Metropolitan University,Manchester, United Kingdom Email: dr.alikashif.b@ieee.org
[7]Griffith University, Brisbane, Australia,Email: l.sanzogni@griffith.edu.au

Correspondence
*senthilkumar mohan, Email: senthilkumar.mohan@vit.ac.in

Summary

The COVID-19 pandemic has emerged as one of the most disquieting worldwide public health emergencies of the 21st century and has thrown into sharp relief, among other factors, the dire need for robust forecasting techniques for disease detection, alleviation as well as prevention. Forecasting has been one of the most powerful statistical methods employed the world over in various disciplines for detecting and analyzing trends and predicting future outcomes based on which timely and mitigating actions can be undertaken. To that end, several statistical methods and machine learning techniques have been harnessed depending upon the analysis desired and the availability of data. Historically speaking, most predictions thus arrived at have been short term and country-specific in nature. In this work, multi-model machine learning technique is called EAMA for forecasting Covid-19 related parameters in the long-term both within India and on a global scale have been proposed. This proposed EAMA hybrid model is well-suited to predictions based on past and present data. For this study, two data sets from the Ministry of Health & Family Welfare of India and Worldometers respectively, have been exploited. Using these two data sets, long-term data predictions for both India and the world have been outlined, and observed that predicted data being very similar to real-time values. The experiment also conducted for state-wise predictions of India and the country-wise predictions across the world and it has been included in the Appendix.

KEYWORDS:
COVID-19, Prediction, Machine Learning, Ensemble learning, Healthcare

1 | INTRODUCTION

The current focus on the analysis of environmental data for the prediction of future trends has made it a prime area of research worldwide. Based on the kind of data prediction and analytical techniques employed, the present as well as the future state of the data can be forecasted or predicted. Different techniques related to modeling, statistics, data mining, artificial intelligence, and machine learning are used for the analysis of data from the past or present in order to forecast future trends. The various stages involved in such analysis and predictions include defining the task, collecting related data from various sources, analyzing the data, statistical analysis, data modeling, deployment of the collected data using multiple techniques, and finally, model monitoring. This kind of predictive analysis is frequently applied to various use case scenarios such as market sales prediction,

customer requirement prediction, healthcare status prediction, collection analysis, fraud detection, etc. Among these, the analysis and prediction of healthcare data is considered to be an important area of application, especially for predicting the future state of proliferation of highly infectious diseases. In this context, the analysis of Covid-19 related data for predicting its proliferation and containment trends is of utmost importance for arresting this ongoing pandemic across the world. Its highly infectious nature and high mortality rates make every second a valuable one, as its infection and mortality rates continue to burgeon every single day.

Countries across the world have adopted certain protocols to arrest the spread of the disease like staying indoors, social distancing, hand washing, travel restrictions, lockdowns, etc. Some of these measures such as lockdowns are quite severe and affect normal human activities in unprecedented ways and have severe economic ramifications. For instance, the spate of lockdowns across the world recently has severely affected the GDP of the entire world, making robust forecasting of Covid-19 related parameters even more crucial. In order to meet this requirement, various denominations of analyses and predictions using information gathered from different sources such as daily updated websites, Kaggle, Orange, and Weka can be seen. As a result, various techniques and methodologies introduced by different researchers for forecasting the future effects of the Covid-19 pandemic can be seen competing with each other, with each having their unique strengths and weaknesses. Advanced artificial intelligence techniques such as machine learning and deep learning have also being used to undertake such forecasting, with each technique having its own unique approach. In machine learning for instance, different approaches and techniques such as the regression model, the auto-regressive model, the classification model, etc. are used.

The novelty of the work proposed and outlined in this article is that it considers an approach that combines non-linear transmission and social-spatial and temporal transmission along with the month-wise prediction of future data. Most of the historical data-driven approaches have been linear methods, and do not consider the temporal or time-based transmission methods.

As such, the contribution of the Paper can be outlined thus:

1. Building of a computational hybrid model with long term predictions for India, with state-wise and date-wise views.
2. Proposing the Ensemble learning hybrid model that integrates different machine learning techniques and improves prediction accuracy.
3. Applying an auto-regressive correlation model to predict the future behavior of Covid-19 data using past and seasonal data.
4. Use of the hybrid model to predict the future Covid-19 status of various countries, along with state-wise, date-wise data views predicted with the help of seasonal data such as heat, air quality, location and other dynamically updated inputs.

The organization of the paper is as follows: Section 2 provides a brief background on related works and existing research pertaining to the field. Section 3 provides information about the data set as well as the proposed supervised model. Section 4 provides a gist of the working methodology and section 5 presents the forecast and prediction results and discussion and is followed by the conclusion as well as the future directions of the research.

2 | RELATED WORKS

The authors Abu Kaisar Mohammad et al.,[1] presented the forecasting and predictions for Covid-19 in Bangladesh and used a linear regression model in order to train with 25 days of data. The data was validated using root mean square error and produced forecast data for a month. The authors lamo T, Reina et al.,[2] carried out data forecasting based on mobility, demographic variables, government measures, weather conditions, and described various data sets linked to various countries. The Authors Ahmad Alimadadi et al.,[3] developed a text and data mining technique to predict Covid-19 parameters and this work was analyzed with the help of a machine learning method for predicting spread, accuracy, and speed of diagnosis. The authors Mohammed A. A. Al-qaness et al.,[4] proposed a model for forecasting that used an adaptive neuro-fuzzy inference system (ANFIS). In this work an enhanced flower pollination algorithm (FPA) and a salp swarm algorithm (SSA) based method were proposed. This method demonstrated the best performance with the use of Root Mean Squared Relative Error (RMSRE), Root Mean Squared Relative Error (RMSRE), Mean Absolute Percentage Error (MAPE), and coefficient of determination. The implementation of this method involved usage of two data sets from China and the USA. The authors S. F. Ardabili et al.,[5] proposed models based on machine learning techniques such as a multi-layered perceptron and an adaptive network-based fuzzy inference system. The proposed model predicted behavior with nation-wise and day-wise views. The authors M. Azarafza et al.,[6] predicted forecasting results using deep learning techniques for the Covid-19 datasets from Iran. The proposed model used the Long Short Term Memory (LSTM) neural network for forecasting scenarios for the entire country.

The authors Mouhamadou and A.M.T. Balde proposed comparative forecasting results using machine learning methods. The classical SIR model[7] was used to fit Covid-19 data using different techniques and tools for forecasting including machine learning with fitting functions. The authors Zlatan Car et al.,[8] proposed a Multi-Layer Perceptron (MLP) and an Artificial Neural Network (ANN) for predicting the spread of Covid-19. This model had 48384 ANNs of trained data from 16128 data sets and this model also cross-validated using a k-fold algorithm with the ReLU function used for activation. Authors Raj Dandekar et al.,[9] proposed a model for Covid-19 prediction with data from four countries, namely China, Italy, South Korea, and the United States of America, based on the neural network augmented model. Quarantine and isolation-related criterion were used to analyze and forecast the Covid-19 related parameters. The SIR model was used for the assumption of direct transmission. The authors Malvika, S. Marimuthu et al.,[7] proposed a model for short-term projections and the prediction of maximum number of active cases. The logistic and SIR growth model was used for the prediction of state-wise data with actual and future predictions with data from four main states, as shown in the implementation.

The authors[10] proposed a model for cumulative forecasting using a modified stacked auto-encoder. Using this method, multiple-step forecasting was predicted and the trajectory of the forecast was shown to be highed. In this work Artificial Intelligence (AI)[11] with an encoder for the entire world's Covid-19 data from WHO was used and forecasted. The Modified Auto-encoder was used for real-time predictions around 30 countries. The authors Chiou-Jye Huang et al.,[12] proposed Deep Convolutional Neural Network (DCNN) for Covid forecasts for the datasets from China. The DCNN forecast based on the state-wise data of china and this model produced more accurate predictions. Authors Pavan Kumar et al.,[13] proposed the Auto-Regressive Integrated Moving Average Model (ARIMA) model for the top 15 countries in the world. Based on the predictions, China was demonstrated to be a faster recovering country compared to all the other countries included. Dianbo Liu et al., proposed a model[14] to forecast Covid-19 related parameters based on different types of input data such as (a) official health reports from China CDC, (b) Covid-19-related internet search activity from Baidu, (c) news media activity reported by Media Cloud, and (d) daily forecasts of Covid-19 activity from GLEAM, an agent-based mechanistic model. The agent-based Augmented ARGONet and Global Epidemic and Mobility models were used for geospatial data predictions. The authors Nick Altieri et al.,[15] proposed a model for short-term forecasting based on the trajectory. The Combined Linear and Exponential Predictors were used for the forecast and using this, the prediction of mortality rates and interval errors, etc. were calculated. The authors Babacar Mbaye Ndiaye et al.,[16] proposed a model using the SIR model and machine learning techniques to analyze and present forecasts based on public data. The SIR model was used for predictions of transmission using populations with the help of parametric and non-parametric methods.

The authors[17]Ajitesh Srivastava et al., proposed a heterogeneous prediction model based on human mobility and quick adaption trends. The training model of the proposed work was originally fit into the initial values. The results of the heterogeneous model were based on the fixed and variable schemes that were fixed with travel data and movement data. Rajan Gupta et al.,[18] proposed a model for predictions in India using the SEIR model as well as the Regression model with the help of John Hopkins University dataset. The proposed model had a lower error prediction rate and made predictions upto two weeks. Vasilis Papastefanopoulos et al.,[19] proposed six time series forecasting methods for the prediction of active case populations. The six time series methods were ARIMA, Holt-Winters Additive Model, TBAT, Automatic Forecasting Procedure, DeepAR, and N-Beat. With the help of these six methods, deaths as well as recovered and confirmed cases were predicted and compared using data from various countries. The authors Gergo Pinter et al.,[20] proposed a hybrid machine learning approach for pandemic predictions in which adaptive network-based fuzzy inference system (ANFIS) and multi-layered perceptron-imperialist competitive algorithm (MLP-ICA) were proposed to predict the time series of infected individuals and mortality rate. The experimental results were predicted for one-month. FURQAN RUSTAM et al.,[21] proposed different supervised machine learning algorithms for forecasting future Covid-19 trends.

The Linear Regression (LR), Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine (SVM), and Exponential Smoothing (ES) were used for forecasting, but the predictions were made for only 5 to 10 days. The authors of[13] developed a novel hybrid ARIMA-WBF model for forecasting of Covid-19 cases. Real-time forecasts of the daily Covid-19 cases in Canada, France, India, South Korea, and the UK were presented. The forecasts reflected the impact of the broad spectrum of social distancing measures implemented by the governments. Four control variables were also identified with powerful associations with cases, fatality rates using an optimal regression tree. The authors of[22,23] used Covid-19 data from the USA and Canada to predict and forecasting was done using deep learning with short term memory for only two successive days.

Sohini Sengupta et al.,[24] proposed machine learning and k-means clustering and hierarchical clustering methods that were used to forecast pandemic situations. In this method, Indian state-wise and month-wise data were analysed for predictions. The authors Natasha Sharma et al.,[25] proposed a spatial-based transmission and forecasting with the help of the SEIQRD method.

Using this method of spatial heterogeneity, the entire population of India in a region was divided into small distinct geographical sub-regions. The authors Narinder Singh Punn et al.,[26] proposed a technique based on machine learning and deep learning algorithms for the analysis of Covid-19 data. This techniques is used polynomial regression and RMSE methods. L. Yan et al.,[27] proposed a machine learning method for infections of Covid-19 with the help of different clinical features such as fever and cough that was used for prediction. The authors of[28] presented different survey methods and it was demonstrated that forecasting methods using AI, machine learning, and deep learning methods. These methods provide the best resolutions for prediction.

The authors of[29] presented the state transition matrix model for date-wise predictions with data predicted for short durations. The authors of[30] proposed a wide-range Covid-19 data to be predicted across various countries with the help of machine learning. The authors of[31] proposed a deep learning model for long time predictions using LSTM, GRU and Bi-LSTM. But this model did not include any dynamic parameters. The authors of[32] proposed a Weibull based Long-Short-Term-Memory approach (W-LSTM) model for predictions across various countries. The authors of[33] described the drone and body area network model to predict Covid-19 spread in the long term. But this work similarly failed to include various dynamic and influential factors such as heat, climate change, air quality etc. The geographical based parameters are influenced the Covid-19. Especially the air pollution affects the lung inflammation. So, the location-based factors are considered for the prediction of the Covid-19.

Above reviewed methods predicted only short term future data and did not consider spatial transmission. And also long term predictions also were not proposed. The authors et al.,[25] proposed month-wise and state-wise data but did not consider spatial transmission. Clearly, advanced predictions are required to include month-wise and spatial-wise predictions for the best data driven decision-making. In this article, it has been planned to incorporate month-wise predictions based on social-spatial transmissions and multiple changing parameters. Also, most of the previous works cited have failed to consider dynamically updating location-based parameters. In this proposed work additionally, dynamic parameters such as heat, air quality and other location-based factors have been duly considered for predictions. The recent denominating methods for Covid-19 prediction and limitation represented in Table 1.

TABLE 1 COVID-19 prediction and limitation

S.No	Model	Advantages	Limitations
1.	State Transition Matrix Model [29]	Date-wise data predicted	Dynamic parameters not included and predicted short range.
2.	Machine Learning [30]	predicted the growth of the wide-ranging in various countries.	Dynamic Climate not included and date-wise not predicted.
3.	Deep learning using LSTM, GRU and Bi-LSTM [31]	Predicted long time prediction using deep learning	Dynamic parameters are not included.
4.	Long-Short-Term-Memory approach (W-LSTM) model [32]	Predicted using socio-economic factors.	Not described the socio-economic parameters.
5.	Drone based network model for predictions [33].	Predicted Body Area Network and drone-based network model to predict the model	Influencing factors are not described in this model.

3 | PROPOSED METHOD

A Covid-19 forecasting model has important ramifications since based on the direction of future predictions, different decisions will be triggered. This proposed prediction model is based on a hybrid model and is also called the Ensemble Learning, Auto-Regressive, and Moving Regressive (EAMA) model. This EAMA model consists of an Ensemble Learning, an auto-regression model, as well as a moving average model. The ensemble Learning is used to combine multiple features and inputs, and considerably improves the accuracy of the predictions. The auto-regressive model is used to make future predictions based on previous trends and past data measurements. The moving average model similarly forecasts data by using current and past Covid-19 data. The proposed hybrid model therefore consists of various steps that are involved in the forecasting of Covid-19 related trends.

Steps involved in the proposed method:

1. Collect datasets with different parameters from Ministry of India and worldometers.
2. Train 80%, validate 10% and test on 10% of the collected sample datasets.
3. Apply the proposed hybrid model to forecast the Covid-19 data trends.
    i) Ensemble Learning: combining multiple features, inputs, and past data to predict future data trends.
    ii) Auto-regressive: forecasting future Covid-19 data using present and past measurement data.
    iii) The moving average regressive: forecasting Covid-19 data using current and future Covid-19 data trends.
4. Forecast final Covid data scenarios including dynamic parameters.

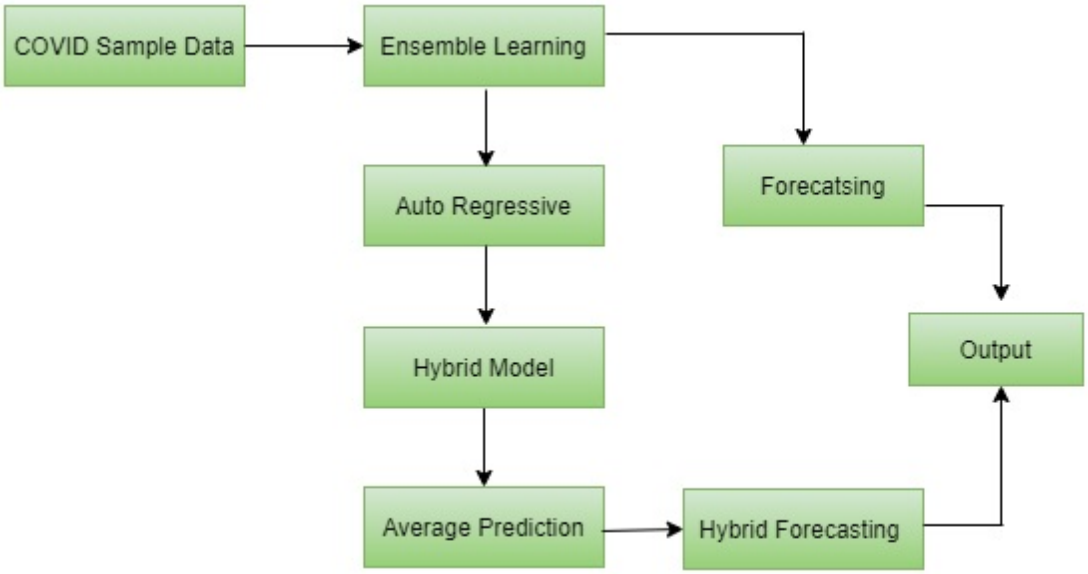

FIGURE 1 : EAMA hybrid model for forecasting

The diagram of the proposed architecture is shown in Figure 1 where the hybrid model based on the ensemble learning is demonstrated. The essence of the hybrid approach is the combination of the sequential and non-sequential data models used to predict the Covid-19 data scenarios. The EAMA model aims to incorporate features like timing, moving patterns of patients, past data, and past behavior of the state and country data using ensemble learning. The ensemble learning examines the state and location-specific data. The location and timing features are incorporated in a weight matrix in a supervised learning model. The past inputs, past measurements, and the time-series based predictions, etc. are saved in the weight matrix based on the locations. The future inputs and a training model are then used to yield the state-wise predictions for a single country or across different countries. The EAMA model continuously incorporates new location and timing-based data and find the errors in these predictions. The errors are then rectified in subsequent predictive iterations. The machine learning method and the EAMA hybrid models yield different predictions in the date-wise and location-wise views. The proposed model focuses on a hybrid statistical time series method to predict future data based on patterns, time, linear predictive models, and non-linear input and output data. The main advantages of the proposed work are improved prediction rates over a longer time period and increased prediction accuracy. The governing Equation (1) of the proposed hybrid model is as given below:

$$Z_i(t) = Y_i(t) + X_i(t) + e_i(t) \tag{1}$$

$Z_i(t)$ denotes the EAMA forecast model and $Y_i(t)$ denotes the current forecast data ensemble learning model.
$X_i(t)$ denotes the observation of data on a time series at time t and location with various seasonal parameters.
$e_i(t)$ is the data error in the model.

The main advantages of the proposed hybrid model compared to previously existing works are predictions over a longer time period and increased accuracy. The materials, the combined supervised hybrid model, and the working procedures of the proposed hybrid method are presented in the upcoming sections.

3.1 | Materials

Dataset: The dataset used in this research work played a crucial role in the accurate prediction of Covid-19 cases. We have collected data mainly from two sources:
1. Ministry of Health & Family Welfare, Government of India:[34] Data pertaining to all the Covid-19 cases in India was taken from the website of the Ministry of Health and Family Welfare which is maintained by the Indian Government.
2. Worldometer:[35] Covid-19 data for the rest of the world was taken from the 'Worldometer' website which is run by an international team of developers, researchers, and volunteers. This website is recognized by the American Library Association.
The aforementioned data sets are updated regularly. Consequently, the nature of this data is highly dynamic. We took the latest data available at this point to plot all the graphs and to create various kinds of Tables. For plotting the graphs, data up till the month of July 2020 was used. Furthermore, we predicted the Covid-19 cases for the next three months (until October 2020) using different Machine Learning methods for Time-Series Forecasting. Duration of Data:
1. India: From 30 January 2020 till July 2020.
2. World: From 22 January 2020 till July 2020.
Classification of Data:
The data is classified into three categories:
1. Confirmed Cases 2. Recovered Cases 3. Deaths

We have incorporated the above 3 categories of data in our datasets. To calculate the 'Active Cases', we added the total 'Recovered Cases' and 'Deaths' and then subtracted the resulting number from 'Confirmed Cases'. Additionally, we have added two new columns in our dataset for India viz; 'Death rate per 100' and 'Cure rate per 100'.

3.2 | Supervised Machine Learning methods

Different supervised machine learning models have been used to predict and analyze future data. Some of the models mainly used in this study are ensemble learning, auto-regressive model, and moving average regressive model.
Ensemble learning: Ensemble learning is a combination of multiple models such as experts or classifiers that are generated in order to solve a particular intelligence problem. The main usage of ensemble learning is to improve the prediction, classification, and to construct better approximations of functions that need to be learned. Using this method, the prediction performance is improved and unfavorable circumstances arising from the use of poor predictions are eliminated. This learning model is used for decision-making processes, incremental learning, and error correction. In this learning model, 'boosting' is employed to increase the weightage for training data that is misclassified so that the existing weak classifier can be strengthened. Use of this boosting concept produces better accuracy. The boosting can be represented as Equation (1) and (2).

$$\mathrm{D}_1 = \mathrm{L}_1 + L_2 \tag{2}$$

Where D_1 - denotes the base leaner or training model, L_1- denotes the weight assigned with the corrected classifier, and L_2- denotes the boosted classifier.

$$\mathrm{D}_2 = \mathrm{L}_1 + L_2 \tag{3}$$

D_2 - denotes the second base learner. Using this boosted D_1 and D_2, the best results are achieved with help of voting between these two base learners or averaging them.

Auto-regressive model : An auto-regressive model predicts the data based on time and measurements taken from previous actions. The previous actions and the statistical correlation between the observations in the past are used to predict the data (in this case Covid-19 data) at future instances. The governing Equation of the auto-regressive model is represented as Equation (4).

$$P(t+1) = C + \sum_{i=0}^{T} W(t-i).P(t-i) + E(t) \tag{4}$$

$P(t+1)$ - denotes the prediction value at time t+1, C - constant location,
T - denotes various logged features, W(t-i) - denotes weight, t - denotes time.

The rate at which Covid-19 spreads is changing every day. With seasonal changes, Covid-19 predictions fluctuate every day. Ensemble learning is used to train the model and update the effects of data automatically. Ensemble learning takes into account the seasonal changes and corresponding factors are used to boost the training model automatically. The parameters encapsulating seasonal changes such as heat, humidity and air quality are updated in every round of boosting. Based on the boosting, the quality of prediction is improved and incorrectly predicted data from previous iterations is also updated. Equations (2) and (3) are used for training and updates every day. Equation (5) is used to represent a fixed region and changing seasonal parameters represented in Equation (6).

$$X_i(t) = \Sigma(X_1, X_2, ..., X_n) + \Sigma(CP_1, CP_2, ..., CP_n) \tag{5}$$

$X_i(t)$ - denotes the observation of data on a time series at time t and region.
$\sum(X)$ - represents different regions and including X are constants.
$\sum(CP)$ - represents different changing parameters based on seasonal changes.

$$\sum(CP) = \left(CH^j(t) + CHU^2(t) + CA^3(t), \ldots, CN''(t)\right) + \left(CH^j(t+1) + CHU^2(t+1) + \mu, CP^n(t+1)\right) \tag{6}$$

Σ(CP) - represents different parameters at the initial time, $CH^1(t)$ - denotes heat, $CH^2(t)$ - denotes humidity,$CH^3(t)$ - denotes air quality, $CH^n(t)$ - denotes upcoming parameters, $CP^n(t+1)$ - denotes next time continuous updating values. In equation (4), $\sum(X)$ represents different locations and Σ(CP) represents the changing parameters with respect to region and time. These parameters are changing every day. The final prediction of Covid-19 spread per day is as shown in Equation (7).

$$CF(X) = \sum_{t=1}^{T} dt.j(Xi(t)) \tag{7}$$

As mentioned in Equation (5), $\Sigma(X)$ - denote represents different locations with X as constants and $\Sigma(CP)$ - denote represents different dynamic parameters based on seasonal changes.The updated prediction represented in Equation (8).

$$P(t+1) = w(t).f(t) + w(t) + f(t) \tag{8}$$

P(t+1) - denote the prediction value of t+1, f(t) - denote various features of the seasonable change, w(t).f(t) - denote predicted features data f(t) with respect to weighted data w(t). Using Equation (4), the past trends in Covid-19 data can be extrapolated to predict future qualitative and quantitative behavior.

Ensemble learning is an automated learning model and it supports different parameters and combines different parameters. The two main parameters that are used for training and testing are location and dynamic seasonal parameters. The (X) is considered as location and is essentially constant. The (CP) encodes dynamic parameters with each and every parameter being updated for every iteration of the computations. The initial data was employed as the training data, but later, all the parameters are automatically trained and updated.

Moving average model: The moving average model is a common model for predicting data based on the linear dependence existing between the current and past values. The forecasting model is represented in Equation (9).

$$y_t = C + \mathscr{C}_t + \theta_1 \mathscr{C}_{t-1} + \theta_2 \mathscr{C}_{t-2} + ... + \theta_p \mathscr{C}_{t-q} \tag{9}$$

where,y_t denote the weighted moving average.
$\mathscr{C}_t$ denote the error rate.
$\theta_1 + ... + \theta_p$ denote the different time-series pattern.
q- denote moving average data.

Various steps for Hybrid EAMA model:

In the Algorithm 1 represented the various procedure and steps for prediction of the Covid-19. Initially the various input dataset sources are included to the training and testing of the model. Second part of the proposed work, included various parameters such as location and other features are incremented. The help of D1 and D2 simultaneously all the features are boosted continuously.
Step 1: The collected Covid-19 sample data from February to July duration is used as training data.
Step 2: Determine the best sample using the EAMA model based on current sample and training data.

- Ensemble learning combines different inputs such as location, migration data, and past data, and is described in section 3.2.

- Auto-regressive model measures the current and past data and based on this, the future data is predicted as explained in section 3.2.
- Moving average model, forecasts the future data using past and present values, and produces average values from multiple instances of data.

Step 3: Train the model using the series of data generated and produce boosted classifiers with the help of D1 and D2 recursively.

- The future data is correlated using Equation (3) and weight W(t-i) and time series t is updated continuously.
- Obtain the weighted and fine grained solution using Equations (5-8).

The sample is validated, tested, and predictions are generated.
Step 4: Final predictions are generated using moving average model (with the help of Equation (5)). Continuous prediction and testing are performed in different iterations.

Algorithm 1 Hybrid EAMA model

Require: Input: Automated trained datasets, India State wise datasets, Countries wise datasets.
Output: Forecasted Data
Initialize EAMA model
if $X_i(t) = 0$ **then** $\sum(X)$ = fixed location
$\sum(CP)$ = CP +1
T= 0; T = T+ 1 : Increment feature value by 1.
P (t+1) = W(t-i).p(t) : Current Prediction.
P(t+1) + 1 : Increment of prediction value of 1.
end if
if y(t)=0 **then**
y(t)= y (t) + 1 : Increment of moving Average value by 1.
$D_1 = D_1 + 1$: Increment of base learner value by 1.
$D_2 = D_2 + 1$: Increment of boosted learner value by 1.
$X_i(t) = X_i(t) + 1$; : Increment of prediction value.
end if

4 | RESULTS AND DISCUSSION

This section describes the achieved results and presents them in a diagrammatic way. For the implementation of the algorithm and the discussion of the results, two data sets are used, and these are described in detail in section 3.1. The proposed EAMA hybrid model achieved significant predictive accuracy on three parameters, namely the number of people affected, the number of recoveries, and mortality Figures. The predicted results are shown in Figures 2 and 3 and are also fully tabulated in the appendix. The predictions pertaining to India and various countries from the months of July to October 2020 are shown in Figure 2 and Figure 3. The proposed hybrid model was predicted from January to July data. For implementation, different metrics such as Reproductive Number (R0), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), Prediction Error (PE) are employed. The R0 value is used to measure the closest value and to find the related values. Some of the R0 values that are used in the prediction are as follows: Italy (0.95), France (0.85), Germany (0.85), Spain (0.85), India ($0.9 - 1.25$), United states (0.9), United Kingdom (0.85). Among Indian states the maximum R value is more than 1 as seen in Kerala. These regressive

base values are also used in the implementation of the model for the prediction of Covid-19 cases. Some other metrics used in the prediction are as shown Equations (10) - (13).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{|\text{ observed}(i) - \text{predicted }(i)|^2}{n}} \tag{10}$$

$$PE(t) = \left(\frac{(\text{ observed }(t) - \text{ predicted }(t))}{\text{observed }(t)} \right) \tag{11}$$

$$MAPE = \left(\frac{1}{n}\right) \left(\sum_{t=1}^{n} |PE(t)| \right) \tag{12}$$

$$R_0 = 1 + r_0 v + f(1-f)\left(r_0 v\right)^2 \tag{13}$$

The output of the prediction results are based mainly on the twin parameters of affected and forecast data for the India and the world. For the implementation of the proposed EAMA model, we used 80% as training data, 10% for testing, and 10% for validating. The proportions of data used for training, testing, and validation produced the best results compared to other proportions of training, testing and validating data.

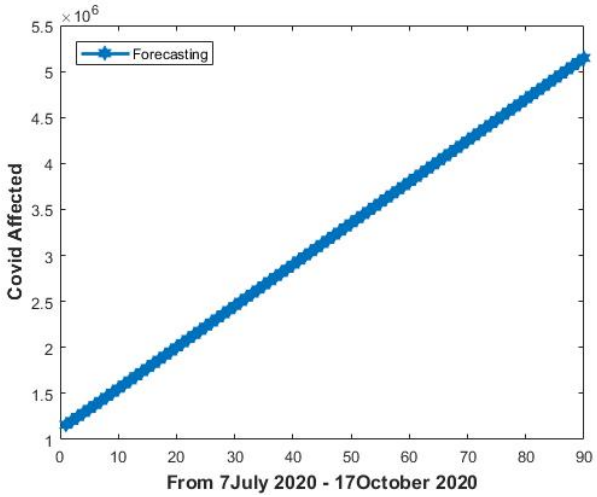

FIGURE 2 : Covid affected in India July to October.

The prediction anticipates a linear growth in the spread of Covid-19 cases. This predicted model is based on the location, past data for the number of people affected in the particular geographical area, the movement patterns of people, etc. The scope of the prediction is gradually expanded and corresponding real-world data is also gradually increased. The prediction of the corresponding parameters at an international level is shown in Figure 4.

In this prediction, the positive cases are decreasing in some periods, and again increasing in certain other periods. The overall cases in the world gradually increased and then gradually decreased. In the middle of August month, the positive cases spiked up suddenly due to the renewed migration of people. Table - 1 1 shows the state-wise prediction in the India and Table - 2 2 shows daily forecasting data for the India as well as the world. The prediction values are seen to have increased gradually.

Figure 6 shows the prediction Covid-19 affected cases worldwide from July to October 2020. The worldwide positive cases, death cases, recovery cases and the confirmed cases are shown in Figure 4. The confirmed cases and active cases are seen to be high again in the worldwide predictions. In this prediction, the recovery cases are seen to gradually decrease and death cases are also found to gradually increase because the number of affected people in advanced age groups are very high worldwide.

Figure 5 shows the forecast predictions of various parameters in the India. For the India, the prediction of the following parameters are made viz; cure rate per 100 people, death rate per 100 people, active cases, cured cases, confirmed cases, etc. Figure7, confirmed Covid-19 cases seem to be always very high when compared with other parameters. The exact values of the

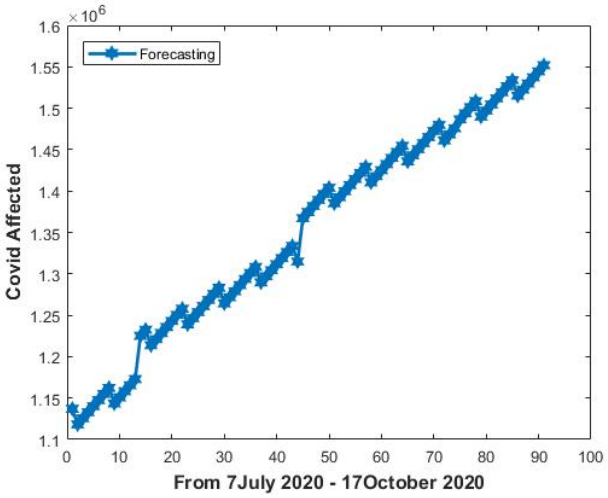

FIGURE 3 : Covid affected daily.

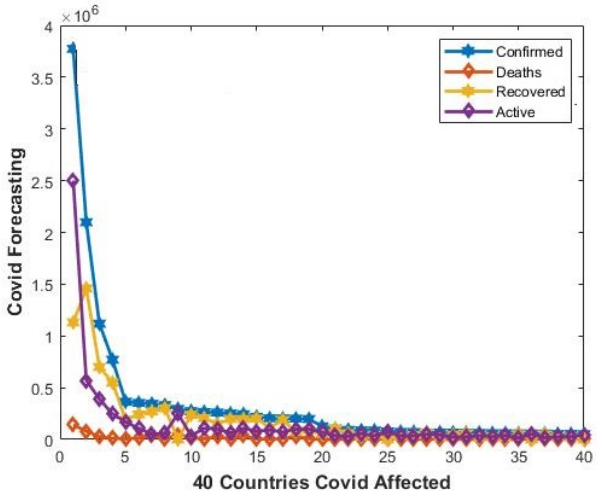

FIGURE 4 : Covid -19 Prediction of 40 countries.

cured and affected numbers are mentioned in Appendix 1. The number of deaths in the India have been quite low but at the end of October, it is bound to have a huge jump due to climatic change, geographical changes, and the number of people affected. In the north India and the south India, cases increased due to the geographical and climate changes. Especially in the north India end of the October and starting of the November the huge number of cases are increased due winter starting. But in the south India, due to weather changes and reduction of migration of the people, cases are reduced. So, automatically the affecting people ratio is reduced in the south India. The active cases are gradually decreasing compared to the number of affected patients. In the India, the real prediction and the forecast are both very similar. Table 1 shows the data forecast state-wise and as opposed to the previous methods, it shows highly accurate predictions. Similarly, Figure 7 shows the world-wide predictions and it uses four parameters viz,. confirmed cases, deaths, recoveries and active cases. The number of confirmed cases, active cases, and deaths are seen to gradually increase but the recovery rate exhibits a marked reduction.The daily international forecasting values and Indian forecasting values are shown in Table 2 and the country-wise forecasting values are shown in Table 3. In this way, all the predictions are compared to the actual daily data and the results are also shown to be fairly close. Compared to the previous methods, the main advantages of this work are the high accuracy of the long-term predictions and their marked similarity with real-world data.

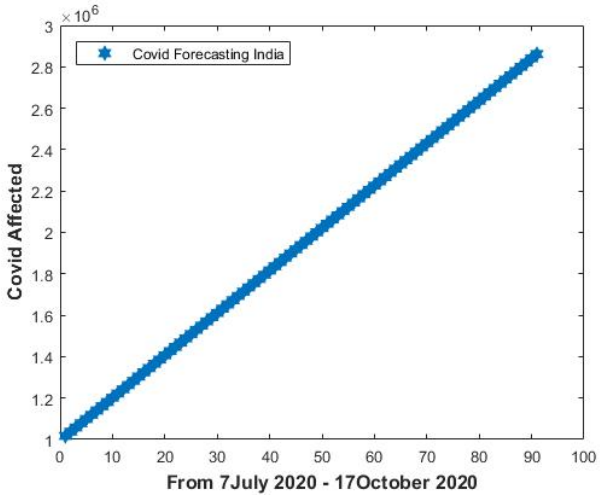

FIGURE 5 : Covid forecasting in India.

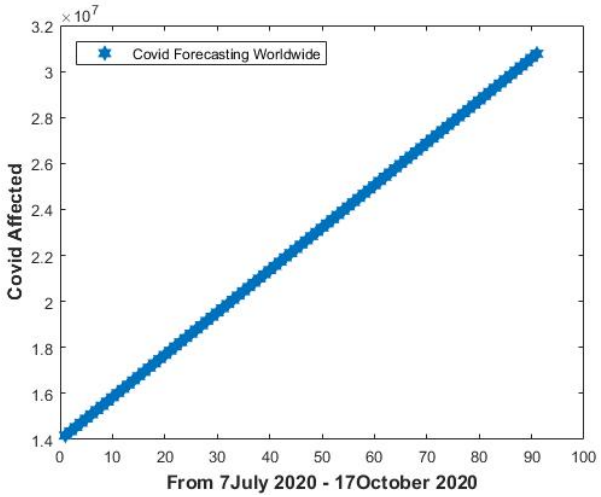

FIGURE 6 : COVID forecasting in Worldwide.

5 | CONCLUSION

The proposed hybrid model makes it possible to acquire novel features of the data because it predicts Covid-19 cases based on different scenarios such as location, past data, movement patterns of citizens etc,. The proposed hybrid supervised model EAMA employed a combination of ensemble learning techniques, auto-regressive, and moving regressive models. Using this mixture of techniques helped to easily combine multiple features and inputs, predict future data using past data trends, and produce average aggregate results. The proposed model had 80%, 10%, and 10% of the overall data for training, testing, and validation respectively. Therefore, the prediction performance was seen to be high, as was the validation accuracy.

For implementation, two different datasets were used, the ministry of India dataset and the Worldometer dataset from the months of February to July 2020. Using this hybrid EAMA model, different parameters were predicted at an international and national level such as the number of affected cases, confirmed cases, and deaths. Especially in the India, the number of active cases and deaths were predicted at a state-wise granularity. The main novelty of the proposed work lies in its long-term prediction accuracy, as opposed to other methods which work only for a short duration. Using these predictions, we can easily measure the future trajectory of Covid-19 cases and employ it in decision and policy making processes. The future work

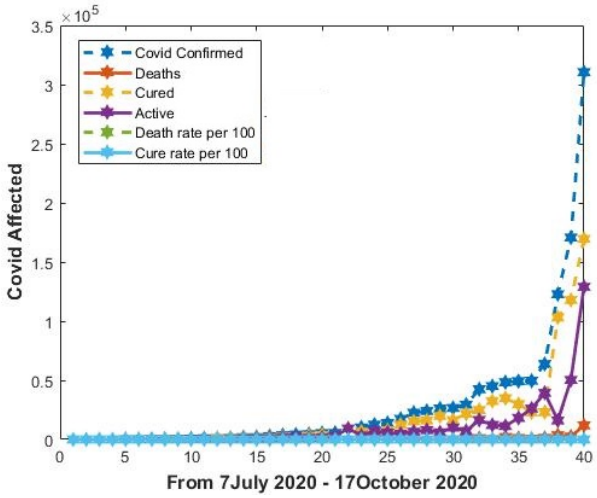

FIGURE 7 : COVID forecasting and impact.

should incorporate different models such as recurrent neural networks and also use non-linear methods for predictions tailored to specific geographical locales.

ACKNOWLEDGMENTS

This work was supported in part by Zayed University, office of research under Grant No. R17089.

Conflict of interest

The authors declare no potential conflict of interests.

How to cite this article: Senthilkumar Mohan, John A, Adimoolam M, Shubham Kumar Singhand A.K.Bashir (2020), COVID-19 Future Forecasting Using Supervised Machine Learning Models, *Trans Emerging Tel Tech.*, *2020*.

APPENDIX

TABLE 1 Forecasting Table

State/UnionTerritory	Confirmed	Deaths	Cured	Active	Death rate per 100	Cure rate per 100
Daman & Diu	2	0	0	2	0	0
Dadar Nagar Haveli	26	0	2	24	0	7.69
Unassigned	77	0	0	77	0	0
Andaman and Nicobar Islands	203	0	145	58	0	71.43
Sikkim	283	0	92	191	0	32.51
Mizoram	284	0	167	117	0	58.8
Meghalaya	450	2	66	382	0.44	14.67
Dadra Nagar Haveli& Daman & Diu	605	2	414	189	0.33	68.43
Chandigarh	717	12	488	217	1.67	68.06
Arunachal Pradesh	740	3	282	455	0.41	38.11
Nagaland	988	0	445	543	0	45.04
Ladakh	1178	2	1003	173	0.17	85.14
Himachal Pradesh	1483	11	1059	413	0.74	71.41
Manipur	1911	0	1213	698	0	63.47
Puducherry	1999	28	1154	817	1.4	57.73
Tripura	2878	5	1759	1114	0.17	61.12
Goa	3657	22	2218	1417	0.6	60.65
Telengana	4111	156	1817	2138	3.79	44.2
Uttarakhand	4515	52	3116	1347	1.15	69.01
Chhattisgarh	5407	24	3775	1608	0.44	69.82
Jharkhand	5535	49	2716	2770	0.89	49.07
Cases being reassigned to states	9265	0	0	9265	0	0
Punjab	10100	254	6535	3311	2.51	64.7
Kerala	12480	42	5371	7067	0.34	43.04
Jammu and Kashmir	13899	244	7811	5844	1.76	56.2
Odisha	17437	91	12453	4893	0.52	71.42
Madhya Pradesh	22600	721	15311	6568	3.19	67.75
Assam	23999	57	16023	7919	0.24	66.77
Haryana	26164	349	19793	6022	1.33	75.65
Bihar	26569	217	16308	10044	0.82	61.38
Rajasthan	29434	559	21730	7145	1.9	73.83
West Bengal	42487	1112	24883	16492	2.62	58.57
Telangana	45076	415	32438	12223	0.92	71.96
Gujarat	48355	2142	34901	11312	4.43	72.18
Uttar Pradesh	49247	1146	29845	18256	2.33	60.6
Andhra Pradesh	49650	642	22890	26118	1.29	46.1
Karnataka	63772	1331	23065	39376	2.09	36.17
Delhi	122793	3628	103134	16031	2.95	83.99
Tamil Nadu	670693	9481	597915	50297	1.45	69.08
Maharashtra	1410455	38854	1069566	129032	3.82	54.62

TABLE 2 COVID-19 Forecasting

Date	India forecasting	World forecasting
8/30/2020	1877722	21916827
8/31/2020	1898218	22101241
9/1/2020	1918713	22285655
9/2/2020	1939209	22470069
9/3/2020	1959705	22654482
9/4/2020	1980201	22838896
9/5/2020	2000697	23023310
9/6/2020	2021192	23207724
9/7/2020	2041688	23392138
9/8/2020	2062184	23576552
9/9/2020	2082680	23760965
9/10/2020	2103176	23945379
9/11/2020	2123671	24129793
9/12/2020	2144167	24314207
9/13/2020	2164663	24498621
9/14/2020	2185159	24683035
9/15/2020	2205655	24867448
9/16/2020	2226150	25051862
9/17/2020	2246646	25236276
9/18/2020	2267142	25420690
9/19/2020	2287638	25605104
9/20/2020	2308134	25789518
9/21/2020	2328629	25973931
9/22/2020	2349125	26158345
9/23/2020	2369621	26342759
9/24/2020	2390117	26527173
9/25/2020	2410613	26711587
9/26/2020	2431108	26896001
9/27/2020	2451604	27080414
9/28/2020	2472100	27264828
9/29/2020	2492596	27449242
9/30/2020	2513092	27633656
10/1/2020	2533587	27818070
10/2/2020	2554083	28002484
10/3/2020	2574579	28186897
10/4/2020	2595075	28371311
10/5/2020	2615571	28555725
10/6/2020	2636066	28740139
10/7/2020	2656562	28924553
10/8/2020	2677058	29108967
10/9/2020	2697554	29293381
10/10/2020	2718050	29477794
10/11/2020	2738545	29662208
10/12/2020	2759041	29846622
10/13/2020	2779537	30031036
10/14/2020	2800033	30215450
10/15/2020	2820529	30399864
10/16/2020	2841024	30584277
10/17/2020	2861520	30768691

TABLE 3 COVID-19 Forecasting

Country	Confirmed	Deaths	Cured	Active
US	8973260	240534	4131121	4501605
Brazil	2098389	79488	1459072	570479
India	6218206	98,497	5,281,200	990622
Russia	770311	12323	549387	245382
South Africa	364328	5033	191059	168236
Peru	353590	13187	241955	108616
Mexico	344224	39184	271239	50099
Chile	330930	8503	301794	59099
United Kingdom	294792	45300	529	249492
Iran	273788	14188	237788	34887
Pakistan	265083	5599	205929	108642
Spain	260255	28752	150376	101617
Saudi Arabia	250920	2486	197735	63026
Italy	244434	35045	196949	108257
Turkey	219641	5491	202010	80808
Bangladesh	204525	2618	111642	90790
Germany	202735	9092	187400	72864
France	201448	30049	72408	99600
Colombia	197278	6736	91793	98749
Argentina	126755	2260	54105	70390
Qatar	106648	157	103377	35634
Iraq	92530	3781	60528	29632
Egypt	87775	4302	28380	55093
Indonesia	86521	4143	45401	37598
Sweden	77281	5619	0	71662
Ecuador	74013	5313	31901	36799
Kazakhstan	71838	375	43029	34497
China	68135	4512	64435	50633
Philippines	67456	1831	22465	43160
Oman	66661	318	44004	22445
Belarus	66095	499	58204	25477
Belgium	63706	9800	17289	36617
Ukraine	60077	1504	31836	26737
Bolivia	59582	2151	18553	38878
Kuwait	59204	408	49687	15831
Canada	57466	5655	0	51811
United Arab Emirates	56922	339	49269	17173
Panama	53468	1096	28482	23890
Dominican Republic	52855	981	25094	26780
Netherlands	51725	6138	100	45589

References

1. Abu Kaisar Mohammad Masum MKSA, Hossain SA. COVID-19 in Bangladesh: A Deeper Outlook into The Forecast with Prediction of Upcoming Per Day Cases Using Time Series. *Procedia Computer Science* 2020; 178: 291-300. doi: https://doi.org/10.1016/j.procs.2020.11.031

2. Alamo T, Reina DG, Mammarella M, Abella A. Covid-19: Open-Data Resources for Monitoring, Modeling, and Forecasting the Epidemic. *Electronics* 2020; 9: 1-30. doi: https://doi.org/10.3390/electronics90508271

3. Alimadadi A, Aryal S, Manandhar I, Munroe PB, Joe B, Cheng X. Artificial intelligence and machine learning to fight COVID-19. *Physiological Genomics* 2020; 52(4): 200-202. doi: 10.1152/physiolgenomics.00029.2020

4. Al-qaness MAA, Ewees AA, Fan H, Abd El Aziz M. Optimization Method for Forecasting Confirmed Cases of COVID-19 in China. *Journal of Clinical Medicine* 2020; 9(3). doi: 10.3390/jcm9030674

5. Ardabili SF, Mosavi A, Ghamisi P, et al. COVID-19 Outbreak Prediction with Machine Learning. *medRxiv* 2020. doi: 10.1101/2020.04.17.20070094

6. Azarafza M, Azarafza M, Tanha J. COVID-19 Infection Forecasting based on Deep Learning in Iran. *medRxiv* 2020. doi: 10.1101/2020.05.16.20104182

7. Baldé MA. Fitting SIR model to COVID-19 pandemic data and comparative forecasting with machine learning. *medRxiv* 2020. doi: 10.1101/2020.04.26.20081042

8. Zlatan C, segota SB, Andelic N, Lorencin I, Mrzljak V. Modeling the Spread of COVID-19 Infection Using a Multilayer Perceptron. *Computational and Mathematical Methods in Medicine* 2020; 2020: 1–10.

9. Dandekar R, Barbastathis G. Quantifying the effect of quarantine control in Covid-19 infectious spread using machine learning. *medRxiv* 2020. doi: 10.1101/2020.04.03.20052084

10. Iwendi C, Bashir AK, Peshkar A, et al. COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Frontiers in Public Health* 2020; 8: 357. doi: 10.3389/fpubh.2020.00357

11. S. Khan; A. K. Bashir; C. Iwendi; T. R. Gadekallu; N. Deepa BP. A Feature Extraction based approach to Detect Covid-19 related Fake News. *Applied Soft Computing* 2020.

12. Huang CJ, Chen YH, Ma Y, Kuo PH. Multiple-Input Deep Convolutional Neural Network Model for COVID-19 Forecasting in China. *medRxiv* 2020. doi: 10.1101/2020.03.23.20041608

13. Kumar P, Kalita H, Patairiya S, et al. Forecasting the dynamics of COVID-19 Pandemic in Top 15 countries in April 2020: ARIMA Model with Machine Learning Approach. *medRxiv* 2020. doi: 10.1101/2020.03.30.20046227

14. Liu D, Clemente L, Poirier C, et al. A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. *arXiv preprint arXiv:2004.04019* 2020.

15. Altieri N, Barter RL, Duncan J, et al. a COVID-19 data repository and forecasting county-level death counts in the United States. *arXiv preprint arXiv:2005.07882* 2020.

16. Liu D, Clemente L, Poirier C, et al. A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. *arXiv preprint arXiv:2004.04019* 2020.

17. Srivastava A, Prasanna VK. Data-driven Identification of Number of Unreported Cases for COVID-19: Bounds and Limitations. *arXiv preprint arXiv:2006.02127* 2020.

18. Pandey G, Chaudhary P, Gupta R, Pal S. SEIR and Regression Model based COVID-19 outbreak predictions in India. *arXiv preprint arXiv:2004.00958* 2020.

19. Papastefanopoulos V, Linardatos P, Kotsiantis S. COVID-19: A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population. *Applied Sciences* 2020; 10(11). doi: 10.3390/app10113880

20. Pinter G, Felde I, Mosavi A, Ghamisi P, Gloaguen R. COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach. *Mathematics* 2020; 8(6). doi: 10.3390/math8060890

21. Rustam F, Reshi AA, Mehmood A, et al. COVID-19 Future Forecasting Using Supervised Machine Learning Models. *IEEE Access* 2020; 8: 101489-101499. doi: 10.1109/ACCESS.2020.2997311

22. Chimmula VKR, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons Fractals* 2020; 135: 109864. doi: https://doi.org/10.1016/j.chaos.2020.109864

23. Chakraborty T, Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *Chaos, Solitons Fractals* 2020; 135: 109850. doi: https://doi.org/10.1016/j.chaos.2020.109850

24. Sengupta S, Mugde S, Sharma G. Covid-19 Pandemic Data Analysis and Forecasting using Machine Learning Algorithms. *medRxiv* 2020. doi: 10.1101/2020.06.25.20140004

25. Sharma N, Verma AK, Gupta AK. Spatial Network based model forecasting transmission and control of COVID-19. *medRxiv* 2020. doi: 10.1101/2020.05.06.20092858

26. Punn NS, Sonbhadra SK, Agarwal S. COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms. *medRxiv* 2020. doi: 10.1101/2020.04.08.20057679

27. Sun C, Tang X, Jing L, et al. An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence* 2020; 2(May). doi: 10.1038/s42256-020-0180-7

28. Shinde GR, Kalamkar AB, Mahalle PN, Dey N, Chaki J. Forecasting Models for Coronavirus Disease (COVID 19): A Survey of the State of the Art. *SN Computer Science* 2020: 1–15. doi: 10.1007/s42979-020-00209-9

29. Zheng Z, Wu K, Yao Z, Zheng J, Chen J. The Prediction for Development of COVID-19 in Global Major Epidemic Areas Through Empirical Trends in China by Utilizing State Transition Matrix Model. *medRxiv* 2020. doi: 10.1101/2020.03.10.20033670

30. Tuli S, Tuli S, Tuli R, Gill SS. Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing. *medRxiv* 2020. doi: 10.1101/2020.05.06.20091900

31. Shahid F MM. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos Solitons Fractals* 2020. doi: 10.1016/j.chaos.2020.110212

32. Tuli S, Tuli S, Verma R, Tuli R. Modelling for prediction of the spread and severity of COVID-19 and its association with socioeconomic factors and virus types. *medRxiv* 2020. doi: 10.1101/2020.06.18.20134874

33. Kumar A, Sharma K, Singh H, Naugriya S, Gill S, Buyya R. A drone-based networked system and methods for combating coronavirus disease (COVID-19) pandemic. *Future Generation Computer Systems* 2021; 115. doi: 10.1016/j.future.2020.08.046

34. Ministry of Health & Family Welfare, Gov. of Indias. "https://www.mohfw.gov.in"; .

35. Worldometer. "https://www.worldometers.info/coronavirus"; .

36. Dhanwant JN, Ramanathan V. Forecasting covid 19 growth in india using susceptible-infected-recovered (sir) model. *arXiv preprint arXiv:2004.00696* 2020.

37. Chakraborty T, Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *Chaos, Solitons Fractals* 2020; 135: 109850. doi: https://doi.org/10.1016/j.chaos.2020.109850

38. Yan L, Zhang HT, Xiao Y, et al. Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. *medRxiv* 2020. doi: 10.1101/2020.02.27.20028027

39. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020; 369. doi: 10.1136/bmj.m1328

40. Shinde GR, Kalamkar AB, Mahalle PN, Dey N, Chaki J, Hassanien AE. Forecasting models for coronavirus disease (COVID-19): a survey of the state-of-the-art. *SN Computer Science* 2020; 1(4): 1–15.

41. Srivastava A, Prasanna VK. Learning to Forecast and Forecasting to Learn from the COVID-19 Pandemic. *arXiv preprint arXiv:2004.11372* 2020.

42. Hu Z, Ge Q, Li S, Boerwincle E, Jin L, Xiong M. Forecasting and evaluating intervention of Covid-19 in the World. *arXiv preprint arXiv:2003.09800* 2020.

43. Hu Z, Ge Q, Jin L, Xiong M. Artificial intelligence forecasting of covid-19 in china. *arXiv preprint arXiv:2002.07112* 2020.

44. Grasselli G, Ospedale G, Policlinico M, et al. Critical Care Utilization for the COVID-19 Outbreak in Lombardy , Italy Early Experience and Forecast During an Emergency Response. 2020; 323(16): 1545–1546. doi: 10.1056/NEJMoa2002032