IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 4, 2020, accepted May 13, 2020, date of publication May 25, 2020, date of current version June 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2997311

COVID-19 Future Forecasting Using Supervised Machine Learning Models

FURQAN RUSTAM[1], AIJAZ AHMAD RESHI[2], (Member, IEEE), ARIF MEHMOOD[3], SALEEM ULLAH[1], BYUNG-WON ON[4], WAQAR ASLAM[3], (Member, IEEE), AND GYU SANG CHOI[5]

[1]Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan
[2]Department of Computer Science, The College of Computer Science and Engineering, Taibah University, Medina 42353, Saudi Arabia
[3]Department of Computer Science and IT, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan
[4]Department of Software Convergence Engineering, Kunsan National University, Gunsan 54150, South Korea
[5]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, South Korea

Corresponding authors: Arif Mehmood (arifnhmp@gmail.com) and Byung-Won On (bwon@kunsan.ac.kr)

This work was supported by the National Research of Korea (NRF) grant funded by Korea government (MSIT) (No. NRF-2019R1F1A1060752).

ABSTRACT Machine learning (ML) based forecasting mechanisms have proved their significance to anticipate in perioperative outcomes to improve the decision making on the future course of actions. The ML models have long been used in many application domains which needed the identification and prioritization of adverse factors for a threat. Several prediction methods are being popularly used to handle forecasting problems. This study demonstrates the capability of ML models to forecast the number of upcoming patients affected by COVID-19 which is presently considered as a potential threat to mankind. In particular, four standard forecasting models, such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES) have been used in this study to forecast the threatening factors of COVID-19. Three types of predictions are made by each of the models, such as the number of newly infected cases, the number of deaths, and the number of recoveries in the next 10 days. The results produced by the study proves it a promising mechanism to use these methods for the current scenario of the COVID-19 pandemic. The results prove that the ES performs best among all the used models followed by LR and LASSO which performs well in forecasting the new confirmed cases, death rate as well as recovery rate, while SVM performs poorly in all the prediction scenarios given the available dataset.

INDEX TERMS COVID-19, exponential smoothing method, future forecasting, adjusted R^2 score, supervised machine learning.

I. INTRODUCTION

Machine learning (ML) has proved itself as a prominent field of study over the last decade by solving many very complex and sophisticated real-world problems. The application areas included almost all the real-world domains such as healthcare, autonomous vehicle (AV), business applications, natural language processing (NLP), intelligent robots, gaming, climate modeling, voice, and image processing. ML algorithms' learning is typically based on trial and error method quite opposite of conventional algorithms, which follows the programming instructions based on decision statements like if-else [1]. One of the most significant areas of ML is forecasting [2], numerous standard ML algorithms have been used in this area to guide the future course of actions needed in many application areas including weather forecasting, disease forecasting, stock market forecasting as well as disease prognosis. Various regression and neural network models have wide applicability in predicting the conditions of patients in the future with a specific disease [3]. There are lots of studies performed for the prediction of different diseases using machine learning techniques such as coronary artery disease [4], cardiovascular disease prediction [5], and breast cancer prediction [6]. In particular, the study [7] is focused on

The associate editor coordinating the review of this manuscript and approving it for publication was Utku Kose.

live forecasting of COVID-19 confirmed cases and study [8] is also focused on the forecast of COVID-19 outbreak and early response. These prediction systems can be very helpful in decision making to handle the present scenario to guide early interventions to manage these diseases very effectively.

This study aims to provide an early forecast model for the spread of novel coronavirus, also known as SARS-CoV-2, officially named as COVID-19 by the World Health Organization (WHO) [9]. COVID-19 is presently a very serious threat to human life all over the world. At the end of 2019, the virus was first identified in a city of China called Wuhan, when a large number of people developed symptoms like pneumonia [10]. It has a diverse effect on the human body, including severe acute respiratory syndrome and multi-organ failure which can ultimately lead to death in a very short duration [11]. Hundreds of thousands of people are affected by this pandemic throughout the world with thousands of deaths every coming day. Thousands of new people are reported to be positive every day from countries across the world. The virus spreads primarily through close person to person physical contacts, by respiratory droplets, or by touching the contaminated surfaces. The most challenging aspect of its spread is that a person can possess the virus for many days without showing symptoms. The causes of its spread and considering its danger, almost all the countries have declared either partial or strict lockdowns throughout the affected regions and cities. Medical researchers throughout the globe are currently involved to discover an appropriate vaccine and medications for the disease. Since there is no approved medication till now for killing the virus so the governments of all countries are focusing on the precautions which can stop the spread. Out of all precautions, ''be informed'' about all the aspects of COVID-19 is considered extremely important. To contribute to this aspect of information, numerous researchers are studying the different dimensions of the pandemic and produce the results to help humanity.

To contribute to the current human crisis our attempt in this study is to develop a forecasting system for COVID-19. The forecasting is done for the three important variables of the disease for the coming 10 days: 1) the number 0f New confirmed cases. 2) the number of death cases 3) the number of recoveries. This problem of forecasting has been considered as a regression problem in this study, so the study is based on some state-of-art supervised ML regression models such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES). The learning models have been trained using the COVID-19 patient stats dataset provided by Johns Hopkins. The dataset has been preprocessed and divided into two subsets: training set (85% records) and testing set (15% records). The performance evaluation has been done in terms of important measures including R-squared score (R^2 score), Adjusted R-squared Score ($R^2_{adjusted}$), mean square error (MSE), mean absolute error (MAE), and root mean square error (RMSE).

This study has some key findings which are listed below:

- ES performs best when the time-series dataset has very limited entries.
- Different ML algorithms seem to perform better in different class predictions.
- Most of the ML algorithms require an ample amount of data to predict the future, as the size of the dataset increases the model performances improve.
- ML model based forecasting can be very useful for decision-makers to contain pandemics like COVID-19.

The rest of the paper consists of six sections. Section I presents the introduction, section II contains the description of the dataset and methods used in this study. Section III presents the methodology, section IV presents the results, and section V summarizes the paper and presents the conclusion.

II. MATERIALS AND METHODS

A. DATASET

The aim of this study is the future forecasting of COVID-19 spread focusing on the number of new positive cases, the number of deaths, and the number of recoveries. The dataset used in the study has been obtained from the GitHub repository provided by the Center for Systems Science and Engineering, Johns Hopkins University [12]. The repository was primarily made available for the visual dashboard of 2019 Novel Coronavirus by the university and was supported by the ESRI Living Atlas Team. Dataset files are contained in the folder on the GitHub repository named (csse_covid_19_time_series). The folder contains daily time series summary tables, including the number of confirmed cases, deaths, and recoveries. All data are from the daily case report and the update frequency of data is one day. Data samples from the files are shown in Tables 1, 2, 3 respectively.

TABLE 1. COVID-19 patient death cases time-series worldwide.

Province /State	Country /Region	Lat	Long	1/22/20	1/23/20	. . .	3/27/20
Northern Territory	Australia	-12.46	130.84	0	0	. . .	0
Diamond Princess	Canada	0.000	0.000	0	0	. . .	1
NaN	Algeria	28.03	1.65	0	0	. . .	19

TABLE 2. COVID-19 new confirmed cases time-series worldwide.

Province /State	Country /Region	Lat	Long	1/22/20	1/23/20	. . .	3/27/20
NaN	Afghan	33.00	65.00	0	0	. . .	74
Victoria	Australia	-37. 81	144. 96	0	0	. . .	411
NaN	Algeria	28.03	1.65	0	0	. . .	264

B. SUPERVISED MACHINE LEARNING MODELS

A supervised learning model is built to make a prediction when it is provided with an unknown input instance. Thus in this learning technique, the learning algorithm takes a dataset with input instances along with their corresponding

TABLE 3. COVID-19 recovery cases time-series worldwide.

Province /State	Country /Region	Lat	Long	1/22/20	1/23/20	...	3/27/20
Colombia	Canada	49. 28	-123. 1	0	0	...	4
Victoria	Australia	-37. 81	144. 96	0	0	...	70
NaN	Algeria	28.03	1.65	0	0	...	65

regressor to train the regression model. The trained model then generates a prediction for the given unforeseen input data or test dataset [13]. This learning method may use regression techniques and classification algorithms for predictive models' development

Four regression models have been used in this study of COVID-19 future forecasting:

- Linear Regression
- LASSO Regression
- Support Vector Machine
- Exponential Smoothing

1) LINEAR REGRESSION

In regression modeling, a target class is predicated on the independent features [14]. This method can be thus used to find out the relationship between independent and dependent variables and also for forecasting. Linear regression a type of regression modeling is the most usable statistical technique for predictive analysis in machine learning. Each observation in linear regression depends on two values, one is the dependent variable and the second is the independent variable. Linear regression determines a linear relationship between these dependent and independent variables. There are two factors (x, y) that are involved in linear regression analysis. The equation below shows how y is related to x known as regression.

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

or equivalently

$$E(y) = \beta_0 + \beta_1 x \tag{2}$$

Here, ε is the error term of linear regression. The error term here uses to account the variability between both x and y, β_0 represents y-intercept, β_1 represents slope.

To put the concept of linear regression in the machine learning context, in order to train the model x is represented as input training dataset, y represents the class labels present in the input dataset. The goal of the machine learning algorithm then is to find the best values for β_0 (intercept) and β_1(coefficient) to get the best-fit regression line. To get the best fit implies the difference between the actual values and predicted values should be minimum, so this minimization problem can be represented as:

$$minimize \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \tag{3}$$

$$g = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \tag{4}$$

Here, g is called a cost function, which is the root mean square of the predicted value of y ($pred_i$) and actual y (y_i), n is the total number of data points.

2) LASSO

LASSO is a regression model belongs to the linear regression technique which uses shrinkage [15]. Shrinkage in this context refers to the shrinking of extreme values of a data sample towards central values. The shrinkage process thus makes LASSO better and more stable and also reduces the error [16]. LASSO is considered as a more suitable model for multicollinearity scenarios. Since the model performs L1 regularization and the penalty added in this case is equal to the magnitude of coefficients. So LASSO makes the regression simpler in terms of the number of features it is using. It uses a regularization method for automatically penalizing the extra features. That is, the features that cannot help the regression results enough can be set to a very small value potentially zero.

An ordinary multivariate regression uses all the features available to it and will assign each one a coefficient of regression. In contrast, the LASSO regression attempts to add them one at a time and if the new feature does not improve the fit enough to out-way the penalty term by including that feature then it could not be added meaning as zero. Thus the power of regularization by applying the penalty term for the extra features is that it can automatically do the selection for us. Thus the models are made sparse with few coefficients in this case of regularization since the process eliminates the coefficients when their values are equal to zero. That means LASSO regression works on an objective to minimize the following:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \tag{5}$$

It sets the coefficient, which can be interpreted as min(sum of square residuals + λ |slopel), where, λ |slopel is penalty term.

3) SUPPORT VECTOR MACHINE

A support vector machine (SVM) is a type of supervised ML algorithm used for both regression and classification [17], [18]. SVM regression being a non-parametric technique depends on a set of mathematical functions. The set of functions called kernel transforms the data inputs into the desired form. SVM solves the regression problems using a linear function, so while dealing with problems of non-linear regression, it maps the input vector(x) to n-dimensional space called a feature space (z). This mapping is done by non-linear mapping techniques after that linear regression is applied to space. Putting the concept in ML context with a multivariate training dataset (x_n) with N number of observations with y_n as a set of observed responses. The linear function can be depicted as:

$$f(x) = x'\beta + b \tag{6}$$

The objective is to make it as flat as possible thus to find the value of $f(x)$ with $(\beta'\beta)$ as minimal norm values. So the problem fits in minimization function as:

$$J(\beta) = \frac{1}{2}\beta'\beta \tag{7}$$

with a special condition of the values of all residuals not more than ε, as in the following equation:

$$\forall_n : |y_n - (x'_n\beta + b)| \leq \varepsilon \tag{8}$$

4) EXPONENTIAL SMOOTHING

In exponential smoothing family methods, forecasting is done based on previous periods' data. The past data observations' influence is decaying exponentially as they become older. Thus the weight assigned to different lag values is geometrically declined. ES is a very simple powerful time series forecasting method specifically for univariate data [7], [19]. The forecast for the current time (F_t) in ES is given by:

$$F_t = \alpha A_{t-1} + (1-\alpha)F_{t-1} \tag{9}$$

Here, α smoothing cost where $0 \leq \alpha \leq 1$, A_{t-1} is the actual value of the previous period in time series, F_{t-1} is the forecast value of the previous forecast.

C. EVALUATION PARAMETERS

In this study, we evaluate the performance of each of the learning models in terms of R-squared (R^2) score, Adjusted R-Square ($R^2_{adjusted}$), mean square error (MSE), mean absolute error (MAE), and root mean square error (RMSE).

1) R-SQUARED SCORE

R-squared (R^2) score is a statistical measure used to evaluate the performance of regression models [20], [21]. The statistic shows the dependent variable's variance percentage that collectively determines the independent variable. It measures the relationship strength between the dependent variable and regression models on a convenient 0 – 100% scale. After training the regression model, we can check the goodness-of-fit of trained models by using the R^2 score. R^2 score finds the scatteredness of data points around the regression line which can also be referred to as the coefficient of determination. Its score always between 0 and 100%. 0% score implies the response variable has no variability around its mean explained by the model, and 100% implies that the response variable has all the variability around its mean. The high R^2 score shows the goodness of the trained model. R^2 is a linear model that explains the percentage of variation independent variable. It can be found as:

$$R^2 = \frac{Varianceexplainedbymodel}{Totalvariance} \tag{10}$$

2) ADJUSTED R-SQUARED SCORE

The Adjusted R-squared ($R^2_{adjusted}$) is a modified form of R^2, which also like R^2 shows how well the data points fit the curve. The primary difference between R^2 and $R^2_{adjusted}$ is that the later adjusts for the number of features in a prediction model. In the case of $R^2_{adjusted}$, the increase in new features can lead to its increase if the newly added features are useful to the prediction model. However, if the newly added features are useless, its value will decrease. The $R^2_{adjusted}$ can be defined as:

$$R^2_{adjusted} = 1 - (1-R^2)\frac{n-1}{n-(k+1)} \tag{11}$$

Here, n is the sample size and k is the number of independent variables in the regression equation.

3) MEAN ABSOLUTE ERROR (MAE)

The mean absolute error is the average magnitude of the errors in the set of model predictions [22], [23]. This is an average on test data between the model predictions and actual data where all individual differences have equal weight. Its matrix value range is from 0 to infinity and fewer score values show the goodness of learning models that's the reason it's also called negatively-oriented scores [24].

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \tag{12}$$

4) MEAN SQUARE ERROR (MSE)

Mean square error is another way to measure the performance of regression models [22]. MSE takes the distance of data points from the regression line and squaring them. Squaring is necessary because it removes the negative sign from the value and gives more weight to larger differences. The smaller mean squared error shows the closer you are to finding the line of best fit. MSE can be calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{13}$$

5) ROOT MEAN SQUARE ERROR (RMSE)

Root mean square error can be defined as the standard deviation of the prediction errors. Prediction errors also known as residuals is the distance from the best fit line and actual data points. RMSE is thus a measure of how concentrated the actual data points are around the best fit line. It is the error rate given by the square root of MSE given as follows.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{14}$$

III. METHODOLOGY

The study is about novel coronavirus also known as COVID-19 predictions. The COVID-19 has proved a present potential threat to human life. It causes tens of thousands of deaths and the death rate is increasing day by day throughout the globe. To contribute to this pandemic situation control, this study attempts to perform future forecasting on the death rate, the number of daily confirmed infected cases and the number of recovery cases in the upcoming 10 days.

The forecasting has been done by using four ML approaches that are appropriate to this context. The dataset used in the study contains daily time series summary tables, including the number of confirmed cases, deaths, and recoveries in the past number of days from which the pandemic started. Initially, the dataset has been preprocessed for this study to find the global statistics of the daily number of deaths, confirmed cases, and recoveries. The resulred time-series has been extracted from the reported data as shown in Table 4, the samples of the resulted dataset are shown in Tables 5, 6, 7 respectively.

TABLE 4. Sample of data from worldwide cases time-series.

Category	Province /State	Country /Region	Lat	Long	1/22/20	. . .	3/27/20
Death	Victoria	Australia	-12.4	130.84	0	. . .	0
	Nan	Canada	0.000	0.000	0	. . .	1
	NaN	Algeria	28.03	1.65	0	. . .	19
Recovery	Colombia	Canada	49. 28	-123. 1	0	. . .	4
	Victoria	Australia	-37. 8	144. 96	0	. . .	70
	NaN	Algeria	28.03	1.65	0	. . .	65
New	NaN	Afghan	33.00	65.00	0	. . .	74
Confirmed	Victoria	Australia	-37. 8	144. 96	0	. . .	411
	NaN	Algeria	28.03	1.65	0	. . .	264

TABLE 5. Day wise total death cases sample data.

Day 1 deaths	Day 2 deaths	. . .	Day 66 deaths
0	4	. . .	20

TABLE 6. Day wise total recoveries rate sample data.

Day 1 recoveries	Day 2 recoveries	. . .	Day 66 recoveries
0	6	. . .	139

TABLE 7. Day wise total new confirmed cases sample data.

Day 1 new cases	Day 2 new cases	. . .	Day 66 new cases
0	21	. . .	749

After the initial data preprocessing step, the dataset has been divided into two subsets: a training set (56 days) to train the models and testing set (10 days). The learning models such as SVM, LR, LASSO, and ES have been used in this study. These models have been trained on the days and newly confirmed cases, recovery, and death patterns. The learning models have then been evaluated based on important metrics such as R^2-score, $R^2_{adjusted}$ score MSE, RMSE, and MAE and reported in the results. The proposed approach used in the study has been shown as a block diagram Figure 1.

IV. RESULTS AND DISCUSSION

This study attempts to develop a system for the future forecasting of the number of cases affected by COVID-19 using machine learning methods. The dataset used for the study contains information about the daily reports of the number of newly infected cases, the number of recoveries, and the number of deaths due to COVID-19 worldwide. As the death rate and confirmed cases are increasing day by day which is an alarming situation for the world. The number of people who can be affected by the COVID-19 pandemic in different countries of the world is not well known. This study is an attempt to forecast the number of people that can be affected in terms of new infected cases and deaths including the number of expected recoveries for the upcoming 10 days. Four machine learning models LR, LASSO, SVM, and ES have been used to predict the number of newly infected cases, the number of deaths, and the number of recoveries.

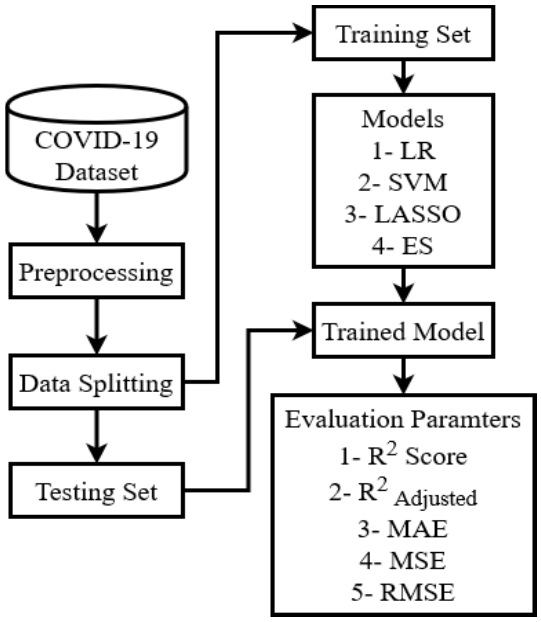

FIGURE 1. Proposed workflow.

A. DEATH RATE FUTURE FORECASTING

The study performs predictions on death rate and according to results ES performs better among all the models, LR and LASSO perform equally well and achieve almost the same R^2 score. In comparison, SVM performs worst in this situation. The results are shown in Table 8.

TABLE 8. Models performance on future forecasting for death rate.

Model	R^2 Score	$R^2_{Adjusted}$	MSE	MAE	RMSE
LR	0.96	0.95	840240.11	723.11	916.64
LASSO	0.85	0.81	3244066.79	1430.29	1801.12
SVM	0.53	0.39	16016210.98	3129.74	4002.02
ES	0.98	0.97	662228.72	406.08	813.77

Figures 2, 3, 4 and 5 show the performance of LR, LASSO, SVM, and ES models respectively in the form of graphs. Graphs in all figures predict that the death rate will increase in upcoming days which is a very alarming sign. The current mortality rate plotted in the graph in Figure 14 shows the models' predictions correct.

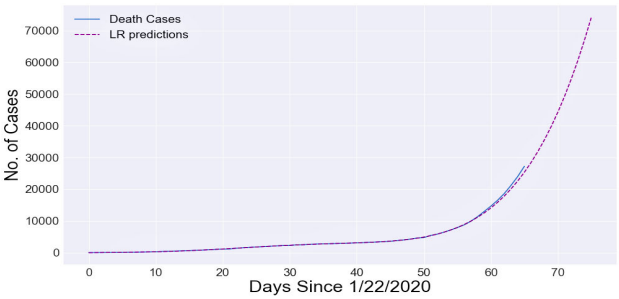

FIGURE 2. Death prediction by LR for the upcoming 10 days.

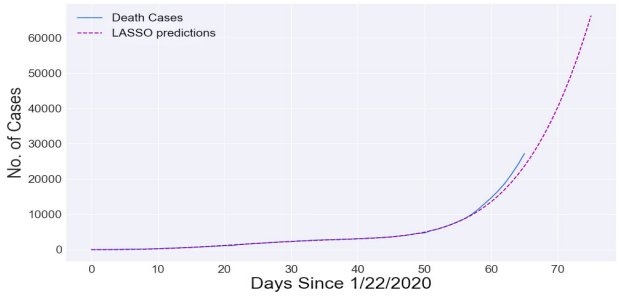

FIGURE 3. Death prediction by LASSO for the upcoming 10 days.

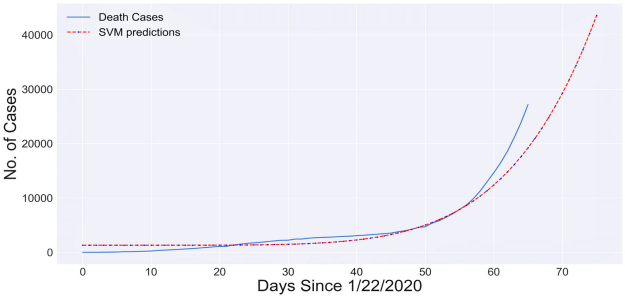

FIGURE 4. Death prediction by SVM for the upcoming 10 days.

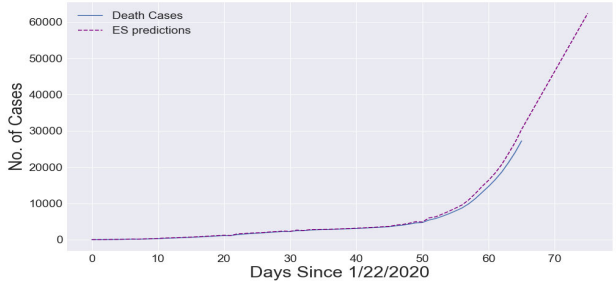

FIGURE 5. Death prediction by ES for the upcoming 10 days.

B. NEW INFECTED CONFIRM CASES' FUTURE FORECASTING

The new confirmed cases of COVID-19 increase day by day Table 9 shows the forecasting results of the models used in this study. ES and LASSO lead the table in terms of performace, LR also performed good, while SVM performs very poorly in terms of all the evaluation metrics. Graphs in figures 6, 7, 8, 9 show the predictions of learning models.

TABLE 9. Models performance on future forecasting for new infected confirm cases.

Model	R^2 Score	$R^2_{Adjusted}$	MSE	MAE	RMSE
LR	0.83	0.79	1472986504.96	30279.55	38390.51
LASSO	0.98	0.97	234489560.99	11693.97	15322.11
SVM	0.59	0.47	5760890969.30	60177.90	75911.28
ES	0.98	0.97	283201302.2	8867.43	16828.58

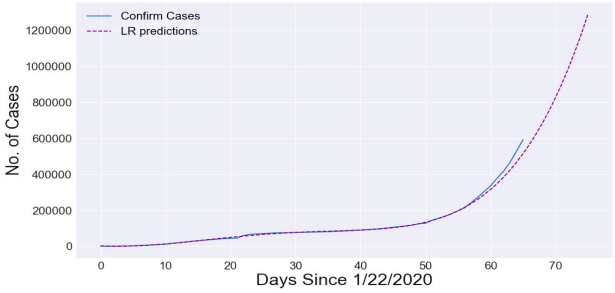

FIGURE 6. New infected confirm cases prediction by LR for the upcoming 10 days.

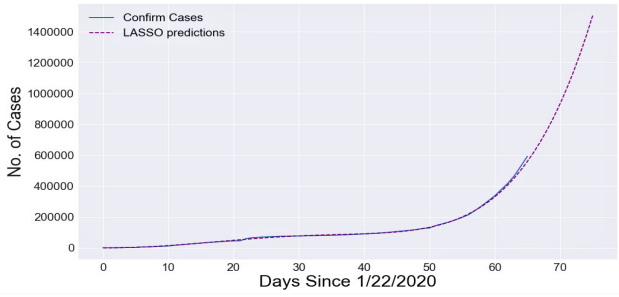

FIGURE 7. New infected confirm cases prediction by LASSO for the upcoming 10 days.

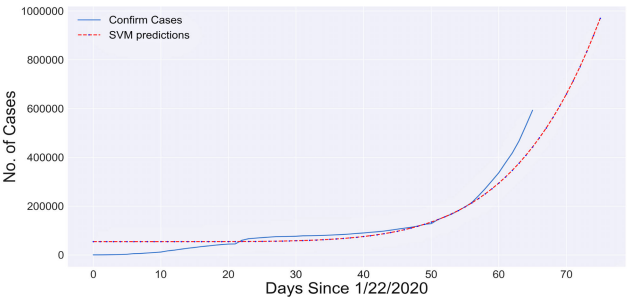

FIGURE 8. New infected confirm cases prediction by SVM for the upcoming 10 days.

C. RECOVERY RATE FUTURE FORECASTING

In recovery rate future forecasting the ES again performs better among all the other models. All other models perform poorly, the order of performance from best to worst is ES is best followed by LR, LASSO and SVM due to the nature

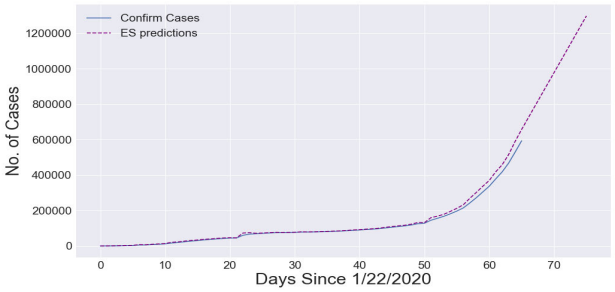

FIGURE 9. New infected confirm cases prediction by ES for the upcoming 10 days.

of available time-series data. The prediction trends for the coming days are shown in Figures 10, 11, 12, and 13. The performance results of learning models are shown in Table 10 below:

TABLE 10. Models performance on future forecasting for recovery rate.

Model	R^2 Score	$R^2_{Adjusted}$	MSE	MAE	RMSE
LR	0.39	0.21	480922814.51	17016.08	21929.95
LASSO	0.29	0.08	1462144344.82	30705.27	38237.99
SVM	0.24	0.02	13121148615.72	106739.82	114547.58
ES	0.99	0.99	5970634.07	1827.85	2443.48

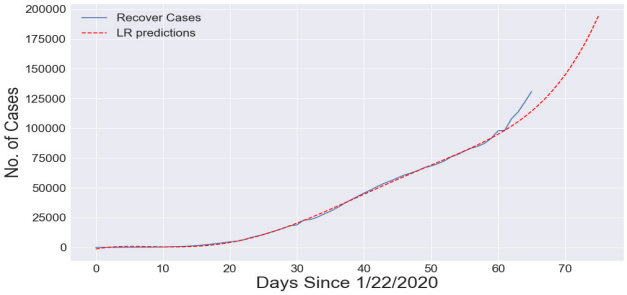

FIGURE 10. Recovery rate prediction by LR for the upcoming 10 days.

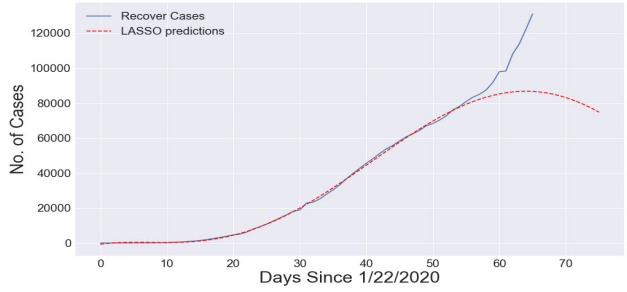

FIGURE 11. Recovery rate prediction by LASSO for the upcoming 10 days.

However, comparing the current recovery statistics (Figure 19) with our models' predictions, the ES prediction is following the trends which are very close to the actual situation.

Besides, some more analysis has been performed after 5 days of experiments on the updated dataset and some important statistics have been found as shown in

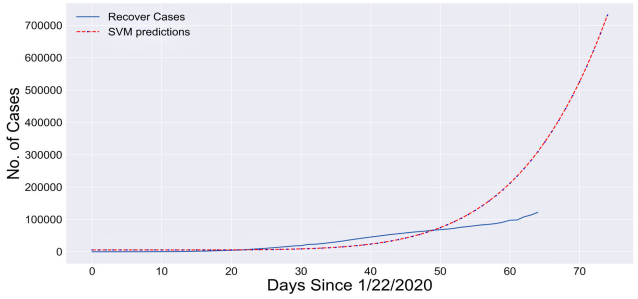

FIGURE 12. Recovery rate prediction by SVM for the upcoming 10 days.

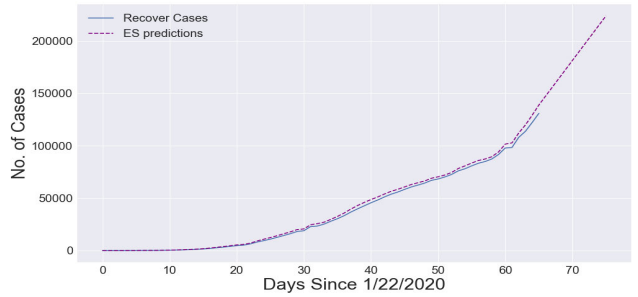

FIGURE 13. Recovery rate prediction by ES for the upcoming 10 days.

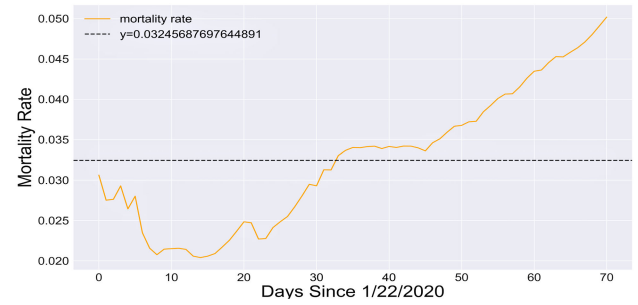

FIGURE 14. Mortality rate after 5 days of this study prediction.

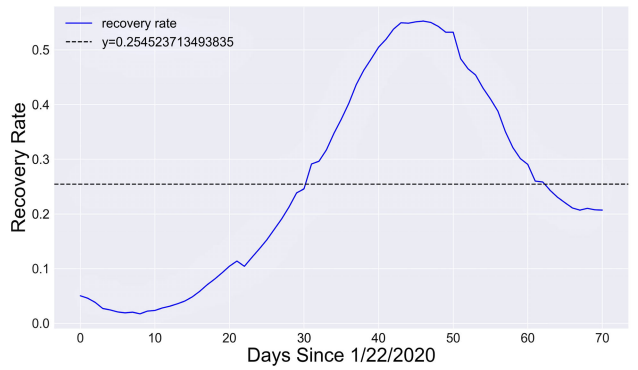

FIGURE 15. Recovery rate after 5 days of this study prediction.

Figure14, 15, 16, and 17. Figure 14 and 15 show that our model predictions are quite promising, because the models predict that in upcoming days death rate will be increased and the graph of mortality rate shows the same pattern and in recovery scenario models predict that recoveries rate will

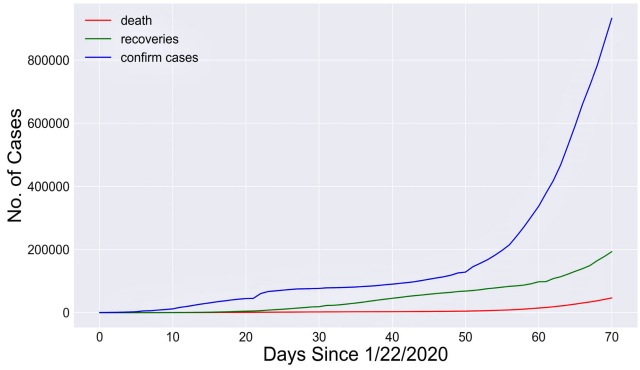

FIGURE 16. Comparison between death rate, recovery rate and confirm case rate after 5 days of this study prediction.

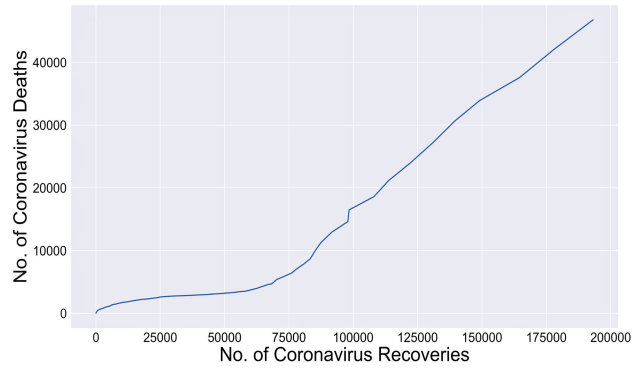

FIGURE 17. Ratio between recovery rate and death rate after 5 days of this study prediction.

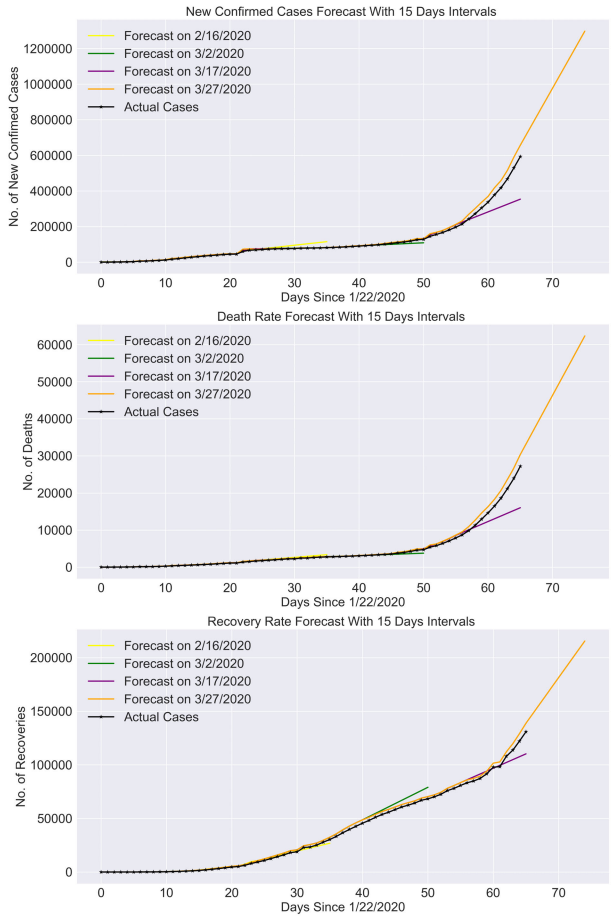

FIGURE 18. ES performances on death rate, recovery rate and new confirmed case with 10-15 days intervals.

be slowed down and recovery graph in Figure 15 follows the same pattern which proves the model predictions correct.

D. MODEL PERFORMANCES WITH 10-15 DAYS PREDICTION INTERVALS

As shown in the previous sections, ES performed best in all three cases such as, death rate forecasting, the number of new confirmed cases forecasting, and recovery rate forecasting. Considering the best performance given by ES model in all the three forecasting cases among all the four models, the model has been used for further analysis with interval prediction [7]. Figure 18 presents the model performance on the death rate, recovery rate, and new confirmed cases with 15 days interval period.

First, all the models have been trained from the dataset of 22 Jan 2020 to 16 Feb 2020, and predictions were made for the upcoming 10 days from 16/02/2020. Since the data available in this dataset was of only 26 days. Due to the availability of a very small sized dataset, three models LR, LASSO, and SVM couldn't perform very well in prediction results as reported in Table 11. However, ES performs better even on the limited number of records in the dataset as shown in the graphs of Figure 18.

In the second model training interval, the models were trained from the dataset of 22 Jan 2020 to 02 Mar 2020, data of 15 more days were added to the training set to predict the outcome of the upcoming 10 days from 02 Mar 2020. Now the dataset contained data of 41 days, the models LR, LASSO, and SVM still could not perform well in all prediction classes. However, the ES in this phase also performed very well as can be seen in graphs of Figure 18.

In the third interval next 15 days were added to the dataset. The size of the training dataset in this interval was 56, as can be seen in the results LR was significantly improved and also the LASSO had shown some improvement. ES in this interval while performing good shows some deviation as shown in the graphs of Figure 18, from the actual data series because of a sudden rise in all the three cases in this period.

In the fourth Interval data of 10 more days have been added increasing the size of the training set to 66, in this interval all the models can be seen as improved very significantly and making the overall results very near to the actual situation. However, ES outperforms all the models in the prediction of all three cases.

In general, ES performed best followed by LR performed followed by LASSO and then SVM. The prediction results have been compared with the actual data reports of these particular day intervals. The predictions results provided by these

TABLE 11. Models performance on future forecasting for recovery rate.

Interval	Dataset Size (Number of Days)	Dates (From 22 Jan 2020) To	LASSO Performance	LR Performance	SVM Performance	ES Performance
1.	26	16 Feb 2020	Very poor	Very poor	Very poor	Best
2.	41	2 Mar 2020	Very poor	Very poor	Very poor	Best
3.	56	17 Mar 2020	Poor	Good	Very poor	Best
4.	66	27 Mar 2020	Better	Best	Well improved	Best

models have been found very closer to the actual reports. The interval details have been compiled and given in Table 11.

To continue and extend further the scope of the of this study in forecasting. The same methodology has been applied to further forecast the number of confirmed cases, deaths, and recoveries up to 6 Apr 2020. Figure 19 presents the plots of confirmed cases, deaths, and recoveries on the first four panes followed by the plot of actual situation gathered from the actual data reports of the sampling period of the study in the fifth pane. The results in the graphs indicate that the ML models used in this study befit the forecasting task making the way towards the usability of the study and future research of the similar nature.

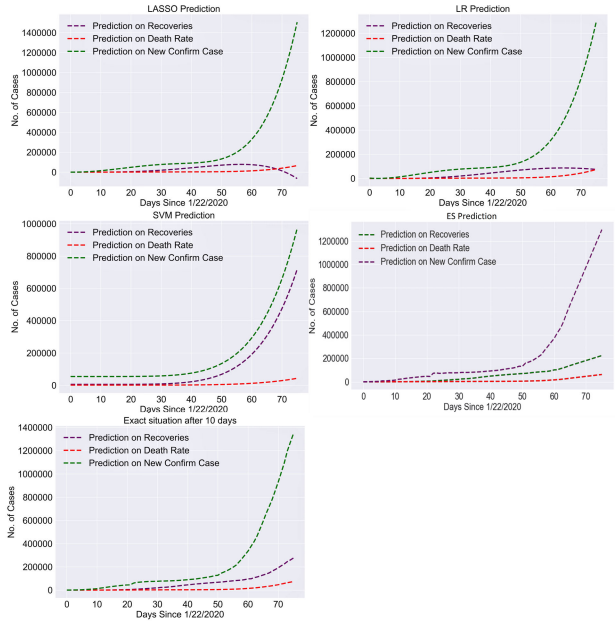

FIGURE 19. All models predictions form 1/22/2020 to 4/6/2020 and real situation form 1/22/2020 to 4/6/2020.

E. PREDICTION INTERVALS OF LR FOR FORECASTING UNCERTAINTY

A prediction interval is a quantification of the uncertainty on a prediction. It provides a probabilistic upper and lower bounds on the estimate of an outcome variable [25]. To evaluate this uncertainty we perform prediction intervals on LR, because among three regression models (LR, LASSO, and SVM), in general, LR performs better in all three cases (death rate forecasting, new confirmed cases forecasting, recovery rate forecasting). The results can be seen in Table 12 showing the prediction intervals along with ranges and true values. Graphs in Figures 20, 21, and 22 show the prediction with interval, actual value, and error bar for newly confirmed cases, death rate, and recovery rate respectively.

TABLE 12. Prediction intervals using LR in all three cases (death rate, new confirmed cases, recovery rate).

Cases	Significance Level	Prediction Interval	Range	True value
Recovery	80%	14403.95	-14375.95 to 14431.95	-23474.35
	90%	18511.33	-18483.33 to 18539.33	-23474.35
	95%	22056.05	-22028.05 to 22084.05	-23474.35
New Confirmed	80%	94320.84	-93765.84 to 94875.84	-64476.148
	90%	121217.02	-120662.02 to 121772.02	-64476.14
	95%	144428.79	-143873.79 to 144983.79	-64476.14
Death	80%	4719.35	-4702.35 to 4736.35	-3488.37
	90%	6065.10	-6048.10 to 6082.10	-3488.37
	95%	7226.50	-7209.50 to 7243.50	-3488.37

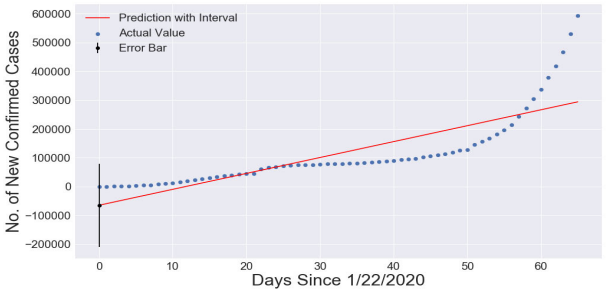

FIGURE 20. Prediction intervals using LR for new confirmed forecasting.

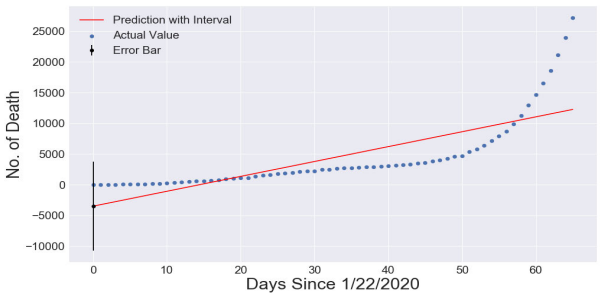

FIGURE 21. Prediction intervals using LR for death rate forecasting.

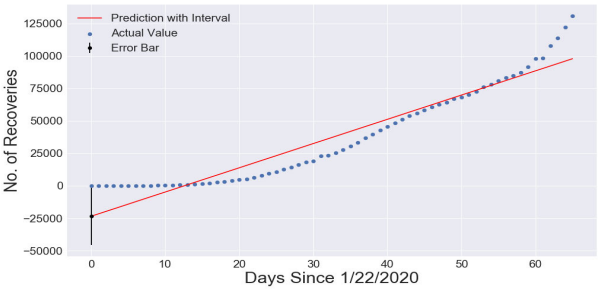

FIGURE 22. Prediction intervals using LR for recovery rate forecasting.

V. CONCLUSION

The precariousness of the COVID-19 pandemic can ignite a massive global crisis. Some researchers and government

agencies throughout the world have apprehensions that the pandemic can affect a large proportion of the world population [26], [27]. In this study, an ML-based prediction system has been proposed for predicting the risk of COVID-19 outbreak globally. The system analyses dataset containing the day-wise actual past data and makes predictions for upcoming days using machine learning algorithms. The results of the study prove that ES performs best in the current forecasting domain given the nature and size of the dataset. LR and LASSO also perform well for forecasting to some extent to predict death rate and confirm cases. According to the results of these two models, the death rates will increase in upcoming days, and recoveries rate will be slowed down. SVM produces poor results in all scenarios because of the ups and downs in the dataset values. It was very difficult to put an accurate hyperplane between the given values of the dataset. Overall we conclude that model predictions according to the current scenario are correct which may be helpful to understand the upcoming situation. The study forecasts thus can also be of great help for the authorities to take timely actions and make decisions to contain the COVID-19 crisis. This study will be enhanced continuously in the future course, next we plan to explore the prediction methodology using the updated dataset and use the most accurate and appropriate ML methods for forecasting. Real-time live forecasting will be one of the primary focuses in our future work.

REFERENCES

[1] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PLoS ONE*, vol. 13, no. 3, Mar. 2018, Art. no. e0194889.
[2] G. Bontempi, S. B. Taieb, and Y.-A. Le Borgne, "Machine learning strategies for time series forecasting," in *Proc. Eur. Bus. Intell. Summer School*. Berlin, Germany: Springer, 2012, pp. 62–77.
[3] F. E. Harrell Jr, K. L. Lee, D. B. Matchar, and T. A. Reichert, "Regression models for prognostic prediction: Advantages, problems, and suggested solutions," *Cancer Treat. Rep.*, vol. 69, no. 10, pp. 1071–1077, 1985.
[4] P. Lapuerta, S. P. Azen, and L. Labree, "Use of neural networks in predicting the risk of coronary artery disease," *Comput. Biomed. Res.*, vol. 28, no. 1, pp. 38–52, Feb. 1995.
[5] K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel, "Cardiovascular disease risk profiles," *Amer. heart J.*, vol. 121, no. 1, pp. 293–298, 1991.
[6] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Comput. Sci.*, vol. 83, pp. 1064–1069, Jan. 2016.
[7] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus COVID-19," *PLoS ONE*, vol. 15, no. 3, Mar. 2020, Art. no. e0231236.
[8] G. Grasselli, A. Pesenti, and M. Cecconi, "Critical care utilization for the COVID-19 outbreak in lombardy, italy: Early experience and forecast during an emergency response," *JAMA*, vol. 323, no. 16, p. 1545, Apr. 2020.
[9] WHO. *Naming the Coronavirus Disease (Covid-19) and the Virus That Causes it*. Accessed: Apr. 1, 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technic% alguidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-cau%ses-it
[10] C. P. E. R. E. Novel, "The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (Covid-19) in China," *Zhonghua Liu Xing Bing Xue Za Zhi= Zhonghua Liuxingbingxue Zazhi*, vol. 41, no. 2, p. 145, 2020.
[11] L. van der Hoek, K. Pyrc, M. F. Jebbink, W. Vermeulen-Oost, R. J. Berkhout, K. C. Wolthers, P. M. Wertheim-van Dillen, J. Kaandorp, J. Spaargaren, and B. Berkhout, "Identification of a new human Coronavirus," *Nature Med.*, vol. 10, no. 4, pp. 368–373, 2004.
[12] Johns Hopkins University Data Repository. *Cssegisanddata*. Accessed: Mar. 27, 2020. [Online]. Available: https://github.com/CSSEGISandData
[13] M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, "Analytics defined," in *Information Security Analytics*, M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, Eds. Boston, MA, USA: Syngress, 2015, pp. 1–12. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128002070000010
[14] H.-L. Hwa, W.-H. Kuo, L.-Y. Chang, M.-Y. Wang, T.-H. Tung, K.-J. Chang, and F.-J. Hsieh, "Prediction of breast cancer and lymph node metastatic status with tumour markers using logistic regression models," *J. Eval. Clin. Pract.*, vol. 14, no. 2, pp. 275–280, Apr. 2008.
[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., Ser. B, Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.
[16] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970.
[17] X. F. Du, S. C. H. Leung, J. L. Zhang, and K. K. Lai, "Demand forecasting of perishable farm products using support vector machine," *Int. J. Syst. Sci.*, vol. 44, no. 3, pp. 556–567, Mar. 2013.
[18] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. Choi, "Tweets classification on the base of sentiments for US airline companies," *Entropy*, vol. 21, no. 11, p. 1078, Nov. 2019.
[19] E. Cadenas, O. A. Jaramillo, and W. Rivera, "Analysis and forecasting of wind velocity in chetumal, quintana roo, using the single exponential smoothing method," *Renew. Energy*, vol. 35, no. 5, pp. 925–930, May 2010.
[20] J. Lupón, H. K. Gaggin, M. de Antonio, M. Domingo, A. Galán, E. Zamora, J. Vila, J. Peñafiel, A. Urrutia, E. Ferrer, N. Vallejo, J. L. Januzzi, and A. Bayes-Genis, "Biomarker-assist score for reverse remodeling prediction in heart failure: The ST2-R2 score," *Int. J. Cardiol.*, vol. 184, pp. 337–343, Apr. 2015.
[21] J.-H. Han and S.-Y. Chi, "Consideration of manufacturing data to apply machine learning methods for predictive manufacturing," in *Proc. 8th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Jul. 2016, pp. 109–113.
[22] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, no. 1, pp. 79–82, 2005.
[23] R. Kaundal, A. S. Kapoor, and G. P. Raghava, "Machine learning techniques in disease forecasting: A case study on rice blast prediction," *BMC Bioinf.*, vol. 7, no. 1, p. 485, 2006.
[24] S. Baran and D. Nemoda, "Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting," *Environmetrics*, vol. 27, no. 5, pp. 280–292, Aug. 2016.
[25] Y. Grushka-Cockayne and V. R. R. Jose, "Combining prediction intervals in the m4 competition," *Int. J. Forecasting*, vol. 36, no. 1, pp. 178–185, Jan. 2020.
[26] N. C. Mediaite. *Harvard Professor Sounds Alarm on 'Likely' Coronavirus Pandemic: 40% to 70% of World Could be Infected This Year*. Accessed: Feb. 18, 2020. [Online]. Available: https://www.mediaite.com/news/harvard-professor-sounds-alarm-on-likely-%coronavirus-pandemic-40-to-70-of-world-could-be-infected-this-year/
[27] BBC. *Coronavirus: Up to 70% of Germany Could Become Infected—Merkel*. Accessed: Mar. 15, 2020. [Online]. Available: https://www.bbc.com/news/world-us-canada-51835856

FURQAN RUSTAM received the M.C.S. degree from the Department of Computer Science, The Islamia University of Bahawalpur, Pakistan, from October 2015 to October 2017. He is currently pursuing the master's degree in computer science with the Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Pakistan. He is also serving as Research Assistant with the Fareed Computing and Research Center, KFUEIT. His recent research interests include data mining, machine learning, and artificial intelligence, mainly working on creative computing and supervised machine learning.

AIJAZ AHMAD RESHI (Member, IEEE) received the Ph.D. degree from the Department of Computer Science, BIHER, Bharath University, Chennai, India, in 2015. He is currently working as an Assistant Professor with the Department of Computer Science, College of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia. His recent research interests include machine learning, deep learning, the Internet of Things (IoT), Web of Things (WoT), and wireless sensor and actuator networks.

ARIF MEHMOOD received the Ph.D. degree from the Department of Information and Communication Engineering, Yeungnam University, South Korea, from February 2014 to November 2017. He is currently working as an Assistant Professor with the Department of Computer Science and IT, The Islamia University of Bahawalpur, Pakistan. His recent research interests include data mining, mainly working on AI and deep learning-based text mining, and data science management technologies.

SALEEM ULLAH was born in AhmedPur East, Pakistan, in 1983. He received the B.Sc. degree in computer science from The Islamia University Bahawalpur, Pakistan, in 2003, the M.I.T. degree in computer science from Bahauddin Zakariya University, Multan, in 2005, and the Ph.D. degree from Chongqing University, China, in 2012. From 2006 to 2009, he worked as a Network/IT Administrator in different companies. From August 2012 to February 2016, he worked as an Assistant Professor with The Islamia University Bahawalpur. He has been working as an Associate Dean with the Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, since February 2016. He has almost 14 years of industry experience in the field of IT. He is currently an Active Researcher in the field of *ad hoc* networks, the Internet of Things, congestion control, data science, and network security.

BYUNG-WON ON received the Ph.D. degree from the Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, USA, in 2007. Then, he worked as a full-time Researcher with the Advanced Digital Sciences Center, University of British Columbia, and the Advanced Institutes of Convergence Technology, for seven years. Since 2014, he has been a Faculty Member with the Department of Software Convergence Engineering, Kunsan National University, South Korea. His recent research interests include data mining, especially on probability theory and applications; machine learning; and artificial intelligence, mainly working on abstractive summarization, creative computing, and multiagent reinforcement learning.

WAQAR ASLAM (Member, IEEE) received the M.Sc. degree in computer science from Quaid-i-Azam University, Islamabad, Pakistan, the Ph.D. degree in computer science from the Eindhoven University of Technology, The Netherlands, and the Ph.D. degree from the Overseas Scholarship, HEC, Pakistan. He is currently an Assistant Professor with the Department of Computer Science and IT, The Islamia University of Bahawalpur, Pakistan. His research interests include performance modeling and QoS of wireless/computer networks, performance modeling of (distributed) software architectures, radio resource allocation, the Internet of Things, fog computing, effort/time/cost estimation of software development in (distributed) agile setups, social network data analysis, and DNA/chaos-based information security.

GYU SANG CHOI received the Ph.D. degree from the Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, USA, in 2005. He was a Research Staff Member with the Samsung Advanced Institute of Technology (SAIT), Samsung Electronics, from 2006 to 2009. Since 2009, he has been a Faculty Member with the Department of Information and Communication, Yeungnam University, South Korea. His research interests include nonvolatile memory and storage systems.