Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 1679835, 19 pages
https://doi.org/10.1155/2021/1679835

Research Article

On-Ground Distributed COVID-19 Variant Intelligent Data Analytics for a Regional Territory

Umrah Zadi Khuhawar,[1] **Isma Farah Siddiqui**,[2] **Qasim Ali Arain**,[2] **Mokhi Maan Siddiqui**,[3] **and Nawab Muhammad Faseeh Qureshi**[4]

[1]*Department of Computer Systems Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan*
[2]*Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan*
[3]*Department of Electrical Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan*
[4]*Department of Computer Education, Sungkyunkwan University, Seoul, Republic of Korea*

Correspondence should be addressed to Nawab Muhammad Faseeh Qureshi; faseeh@skku.edu

Received 4 August 2021; Revised 9 October 2021; Accepted 20 October 2021; Published 10 December 2021

Academic Editor: Hasan Ali Khattak

The onset of the COVID-19 pandemic and the subsequent transmission among communities has made the entire human population extremely vulnerable. Due to the virus's contagiousness, the most powerful economies in the world are struggling with the inadequacies of resources. As the number of cases continues to rise and the healthcare industry is overwhelmed with the increasing needs of the infected population, there is a requirement to estimate the potential future number of cases using prediction methods. This paper leverages data-driven estimation methods such as linear regression (LR), random forest (RF), and XGBoost (extreme gradient boosting) algorithm. All three algorithms are trained using the COVID-19 data of Pakistan from 24 February to 31 December 2020, wherein the daily resolution is integrated. Essentially, this paper postulates that, with the help of values of new positive cases, medical swabs, daily death, and daily new positive cases, it is possible to predict the progression of the COVID-19 pandemic and demonstrate future trends. Linear regression tends to oversimplify concepts in supervised learning and neglect practical challenges present in the real world, often cited as its primary disadvantage. In this paper, we use an enhanced random forest algorithm. It is a supervised learning algorithm that is used for classification. This algorithm works well for an extensive range of data items, and also it is very flexible and possesses very high accuracy. For higher accuracy, we have also implemented the XGBoost algorithm on the dataset. XGBoost is a newly introduced machine learning algorithm; this algorithm provides high accuracy of prediction models, and it is observed that it performs well in short-term prediction. This paper discusses various factors such as total COVID-19 cases, new cases per day, total COVID-19 related deaths, new deaths due to the COVID-19, the total number of recoveries, number of daily recoveries, and swabs through the proposed technique. This paper presents an innovative approach that assists health officials in Pakistan with their decision-making processes.

1. Introduction

The COVID-19 was declared a deadly virus by the World Health Organization (WHO) [1, 2]. There is a need for countries to act in unison to prevent further transmission of the disease. A pandemic is a disease that spreads worldwide [3]. Throughout history, the world has witnessed many pandemics. The most recent was in the year 2009 due to the H1N1 flu. The first few cases of COVID-19 were reported to the WHO on 31 December 2019 in the city of Wuhan, Hubei province in China, wherein several people were afflicted with pneumonia, and the cause could not be determined. In January 2020, officials identified a novel virus that was not named yet [4, 5], which was subsequently popularized as the 2019 novel Coronavirus [6]. Upon obtaining the samples and analyzing the virus genetics, it was established that it caused the outbreak. The virus was named Coronavirus 2019 (COVID-19) by the WHO in February 2020 [7],

while some studies found that this deadly COVID-19 virus is associated with SARS-CoV-2 [8, 9]. With its 204 million population, Pakistan saw first of its case in February 2020 [10]. With the 5th largest population globally, it became essential to understand how the virus will progress in this vast population and how it will progress in Pakistan. Therefore, it has become essential to address the problem of the future trend of COVID-19-positive cases in Pakistan by using the COVID-19 dataset from [11]. Machine learning is widely used to handle large data, and it can help in this regard. We specifically test three methods, namely, linear regression, random forest, and XGBoost algorithm. In this paper, we predict positive COVID19 cases in Pakistani regions of Sindh, Punjab, Gilgit Baltistan, Balochistan, Khyber Pakhtunkhwa, Azad Jammu, and Kashmir using three ML algorithms, and we compare the results; in order to find out the optimal algorithm for the dataset which gives the highest accuracy for the forecast of COVID-19-positive cases. A real-time forecasting scheme is presented based on ML models, which provides real-time prediction allowing citizens and the government of Pakistan to take actions proactively [12, 13]. This paper effectively predicts future COVID-19 pandemic trends by employing open-source data science libraries and machine learning tools in Python. The primary objectives of this study are as follows:

(i) To source [12], preprocess, visualize, and analyze the data of COVID-19 in Pakistan

(ii) To recognize the various parameters required for COVID-19 modeling and drive these variables for all the three forecasting algorithms used

(iii) To rectify and eliminate biases

(iv) Model and predict future the trend of the COVID-19 pandemic

(v) Visualize and discuss the results

The COVID-19 dataset of Pakistan has not been tested on a large scale by using machine learning algorithms. This paper contributes to using machine learning algorithms on indigenous datasets in Pakistan, which can significantly help in assessing and planning to take actions accordingly. The paper is structured in the following manner: Section 2 presents an overview of literature related to COVID-19 forecasting; Section 3 explains the methodology for predicting COVID-19; Section 4 shows the results for all three machine learning models; and Section 5 illustrates the relationship between parameters. In Section 5, we summarized this work and presented various results.

2. Related Work

Kavadi et al. [1] developed a mathematical model to assess and estimate the growth of the worldwide COVID-19 pandemic. Machine learning generalized inverse Weibull model has been implemented to evaluate the potential risks associated with the Coronavirus. In order to ensure precise and real-time prediction on the growth of the pandemic, cloud computing was employed. A model was implemented by Nemati et al. [3] to highlight the efforts of the Pakistan government to fight with COVID-19. This paper presents the current scenario of the Coronavirus situation in Pakistan and provides information about the hospital facilities provided for COVID-19 patients. The results show that the recovery rate is higher than the mortality rate in Pakistan, and Balochistan has more hospitals for COVID-19 patients. Azad Jammu and Kashmir have the least hospitals for COVID-19 patients. Isolation zones were built in Pakistan, and this study shows that Punjab and Khyber Pakhtunkhwa regions have more isolation wards and better medical facilities. Ardabili et al. [4] proposes the PDR-NML method (partial derivative regression and nonlinear machine learning) to predict the pandemic trends of COVID-19. The results show that the proposed ML method is more effective than other state-of-the-art methods in the Indian population. Thus, it can be an innovative tool in helping other countries make their predictions. The authors of this study have also used PPDLR for normalizing the features required for timely prediction and PDLFR for robust and accurate prediction and observed that machine learning performed well for data analysis than artificial intelligence. Lalmuanawma et al. [5] predicted the trend of COVID-19. The Fb-prophet model is used to establish the pandemic curve and forecasting its direction. The disadvantage of this study is that they have used the limited dataset this work is integrated into the logistic model. Three significant points have been summarized based on the modeling results related to Indonesia, Peru, Brazil, India, and Russia. According to estimations based purely on mathematical aspects, the peak of the virus will be witnessed globally in late October, and it is expected that 14.12 million people will be impacted on a cumulative basis. Rustam et al. [7] implemented the autoregressive integrated moving average (ARIMA) model to predict the new COVID-19 cases each day in Saudi Arabia for four weeks. The authors have summarized four different prediction models in this study, including autoregressive model, moving average, a combination of both (ARMA), and integrated ARMA (ARIMA), to identify the apt model fit. The results show that the ARIMA model is more effective in comparison to the other models. Pandey et al. [8] aim to forecast the COVID-19-positive cases in India and Odisha by using linear regression and multiple linear regression. Therefore, it is observed that both models provided remarkable accuracy for the prediction of the COVID-19 pandemic. Roy et al. [10] summarized four machine learning algorithms to forecast COVID-19-infected people. The data of COVID-19 between 20/01/2020 and 18/09/2020 for the USA, Germany, and global were obtained from the World Health Organization. The performance of all algorithms is compared according to the RMSE, APE, and MAPE criteria, and it was observed that these models could be used to diagnose the COVID-19 data over time. To predict the future forecast of the COVID-positive case, Ayyoubzadeh et al. [11] used XGBoost, K-means, and long short-term memory (LSTM) neural networks to construct a prediction model. Therefore, it was observed that K-means–LSTM provides higher accuracy with an error score of 601.20%.

3. Methodology

In this study, classification algorithms were applied, and an evaluation process is done for each algorithm based on different parameters shown in Figure 1. This research work involves few significant steps like data collection, data preprocessing, applying machine learning algorithms, evaluation, and comparative analysis.

3.1. Data Collection. The data used in this work is accessed from http://covid.gov.pk [12, 14]. The information related to COVID-19 cases in Pakistan has been compiled from different sources, including Kaggle and World Health Organization (WHO) [6, 15–17]. A cumulative data set is created from a mix of the above resources. The information taken from http://covid.gov.pk/ data is not in a required CSV format. It also contained some unnecessary data that was not needed to predict positive cases in Pakistan data preprocessing was done. The dataset includes the hospital data of COVID-19-positive patients, deceased patients, recovered patients, total deaths of patients, and the number of swab tests conducted every day in each region of Pakistan. The dataset contains all the COVID-19 data of the patients in the specified data collection period.

3.2. Data Preprocessing. After the collection of information, the data was transformed into the required CSV format. In order to rectify the issue of systemic bias, a feasible methodology was adopted. The moving-average method, which is typically used to assess time-series through the computation of averages of various subsets within the complete dataset, was adopted for this purpose. The moving-average method, which is typically used to assess time-series through the computation of averages of various subsets within the complete dataset, was adopted for this purpose. In this context, seven days were taken as the complete dataset. Initially, the moving average was computed by finding the average of the first subset over seven days. Then, the subset was altered as the following fixed subset was chosen. This went on till all the subsets were subjected to this method. Essentially, this method tends to smoothen the data by mitigating anomalies, the weekend bias. In Figures 2–5, the dataset variables are plotted as time series depicting total COVID-19-positive cases across Pakistan, total COVID-19 deaths across Pakistan, new COVID-19-positive cases in Pakistan regions, and COVID-19 patients who are in serious condition. Figure 2 displays the daily new COVID-19-positive cases in Pakistan as it is essential for forecasting. Figure 3 displays the average of COVID-19-positive cases in a week. And, Figure 4 represents total COVID-19-positive cases across Pakistan. Also, Figure 5 represents total deaths across Pakistan. Figure 6 displays daily new reported COVID-19 cases in Pakistan regions, whereas Figure 7 illustrates the COVID-19 patients' data who are in serious condition.

3.3. Applying Machine Learning Algorithms. After preprocessing, random forest, XGBoost, and linear regression models were applied to predict COVID-19-positive cases in Pakistan [18]. A linear regression model was employed to model the COVID-19 trend. It was trained using positive cases and new positive cases data on both the national and provincial levels in Pakistan. In regression, the R^2 coefficient of determination is a statistical measure that informs the preciseness of the regression predictions by comparing them with the fundamental data points. If the value of R^2 is deduced to be 1, it denotes that the regression predictions accurately align with the data. Thus, the closer the value of R^2 is to 1, the more influential the model is in predicting trends [19]. The random forest algorithm is a popular unsupervised machine learning algorithm, and it is employed for classification [20, 21]. It is an ensemble machine learning method. The random forest represents a decision tree. N number of outputs are obtained by the N number of the decision tree using this algorithm.

3.4. Forecasting the Trend of Positive COVID-19 Cases across Pakistan Regions. The COVID-19 outbreak has badly affected the essential aspects of life around the world. In order to control this outbreak, smart lockdowns have been imposed all over the country and are highly affected areas of Pakistan. This study will provide an idea about the increase of COVID-19 in Pakistan and its provinces. It will also help Pakistan and its citizens make appropriate decisions to handle the situation by following proper SOP's and guidelines.

3.5. Forecasting the Trend of Positive COVID-19 Cases Using Linear Regression Algorithm. In this study, a detailed description of linear regression is presented. In addition, all the tests performed for the validity of linear pegression are analyzed and discussed. We have used linear regression to forecast the value of a dependent variable by provided independent variable data [22–24]. It was observed that there is a linear relationship between independent variables and dependent variables. In our study, we considered X as an independent variable and Y as a dependent variable, and the value of Y is predicted by using the following equation:

$$Y = f(X), \tag{1}$$

where $X = [x_1, x_2, x_3, \cdots, x_p]$ is a vector of P input parameters and $Y = [y_1, y_2, y_3, \cdots, y_Q]$ is a vector of Q output parameters. X is also called independent variables as response variables. In machine learning regression is a method to find the relation between X and y_i. When the relationship is done using a linear predictor function, assuming a system is linear, equation (1) represented by

$$Y = X_a + \in . \tag{2}$$

Here, a is the vector of coefficients of regression and ϵ represents the vector of model error. If we expand the above equation then the equation would be represented as

$$y_i = a_0 + ax_{i1} + ax_{i2} + ax_{i3} + \cdots + ax_{ip} + \in . \tag{3}$$

In equation (3), i, a, and ϵ are estimated by using standard methods. Let us assume that the estimated coefficients is defined by a and the fitting response is represented in

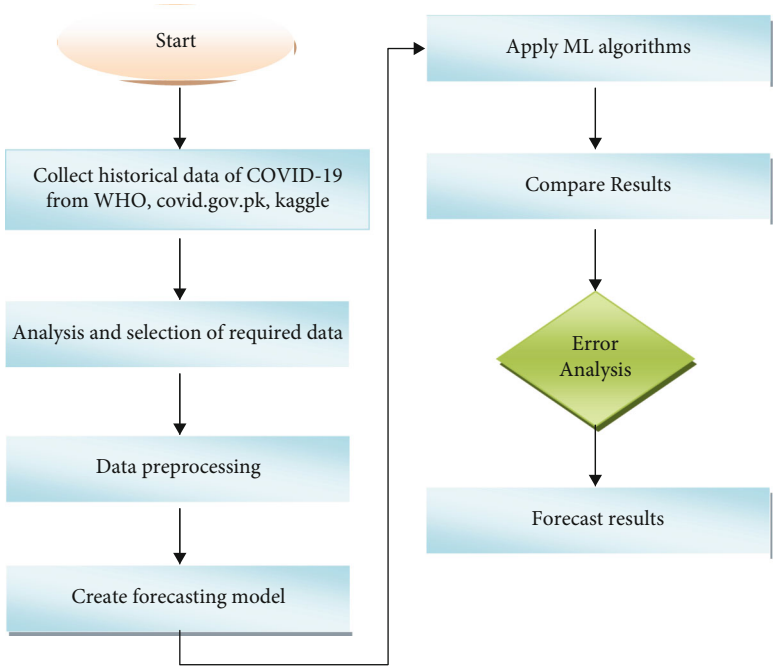

Figure 1: Flow diagram of proposed work.

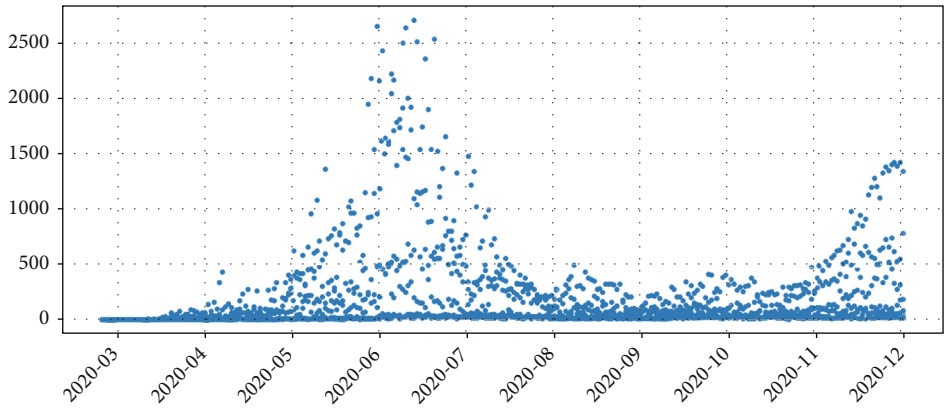

Figure 2: Visualization of daily new positive COVID-19 cases.

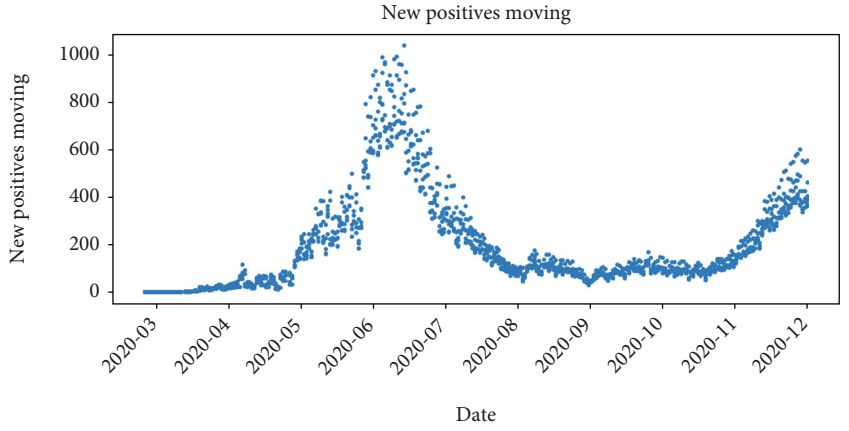

Figure 3: Visualization of new COVID-19 cases averaged over 7-day period.

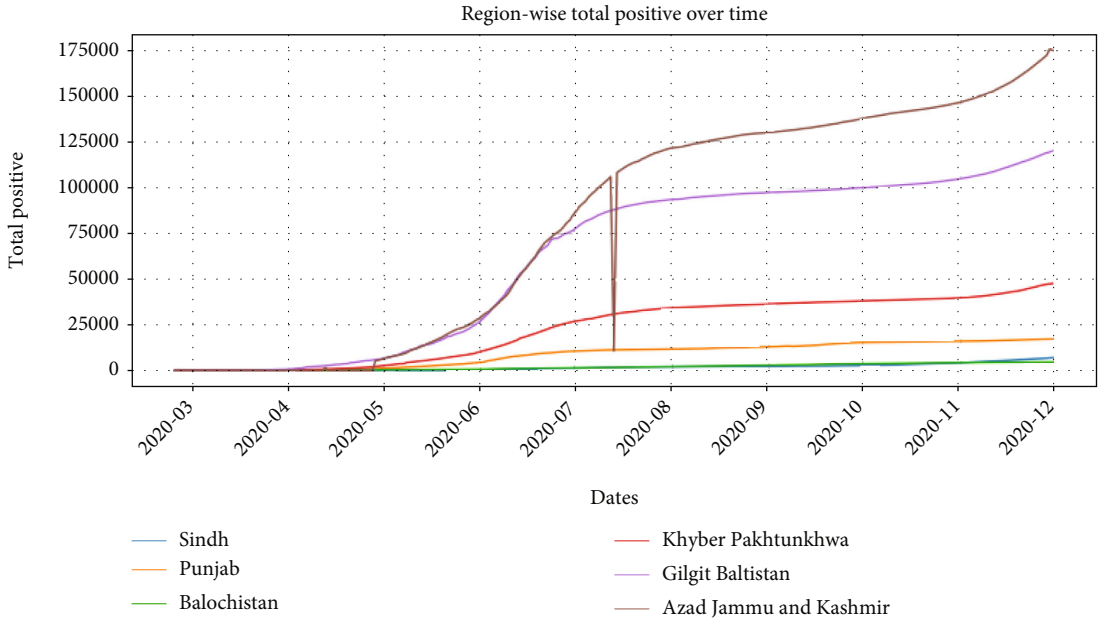

Figure 4: Total COVID-19-positive cases across Pakistan.

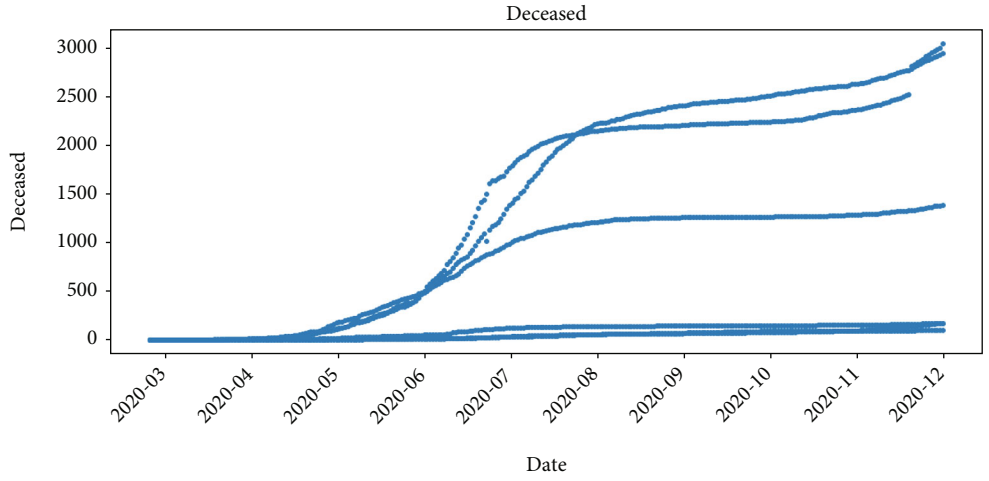

Figure 5: Total COVID-19 deaths across Pakistan.

$$\hat{y}_i = \hat{a}_0 + \hat{a}x_{i1} + \hat{a}x_{i2} + \hat{a}x_{i3} + \cdots + \hat{a}x_{ip}. \quad (4)$$

The R^2 (coefficient of determination) is given by

$$R^2 = \frac{\sum_{i=1}^{N} (y\wedge_i - \bar{y})^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}. \quad (5)$$

Here, Y represents the forecast of total positive cases in Pakistan and X variable represents the date, a_0 denotes the Y-intercept and a_1 indicates the slope. The linear regression model is built by learning the values of a_0 and a_1 from a given dataset, where R^2 is the measure of the proportion of variation in y explained by the P input parameters. In this study, R is used to determine the values of a, R^2, and ϵ, and $\bar{y}$ is the mean of all observations.

3.6. Forecasting the Trend of Positive COVID-19 Cases Using Random Forest Algorithm. To implement the random forest model first, we have taken the COVID-19 dataset of Pakistan as an input. Then, the random forest model was trained on that dataset. Independent variables are considered dependent variables. The actual number of COVID-19 cases is regarded as the dependent variable [25]. The random forest model was used for forecasting the COVID-19-positive cases in Pakistan territories. Implementation of this model is described in the following flowchart.

Random forest consists of many decision trees. The higher the number of decision trees, the more accurate results we will get. There is a direct relation between outcome and number of decision trees in Random Forest. It consists of many decision trees. The higher the number of decision trees, the more accurate results we will get. There

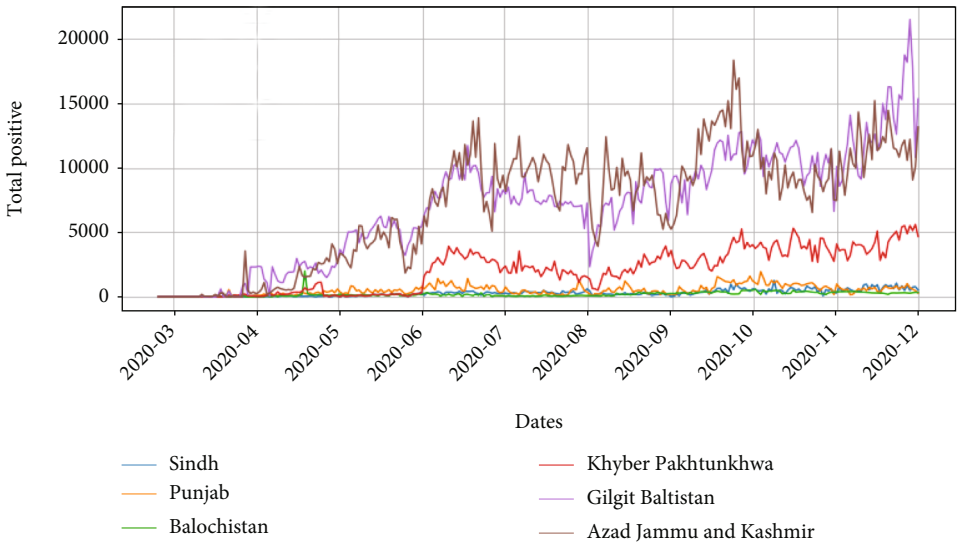

Figure 6: New COVID-19-positive cases in Pakistan regions.

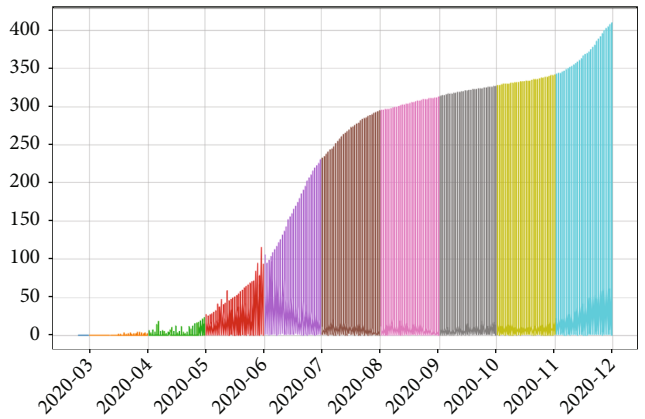

Figure 7: COVID-19 patients who are in serious condition.

is a direct relation between outcome and number of decision trees in random forest [26–28]. The primary purpose of this algorithm is to improve prediction accuracy by aggregating multiple classifiers. The random forest algorithm is widely used for classification and prediction. It can be applied to many fields such as forecasting, data analysis, text classification, and face recognition [29]. This algorithm combines multiple decision trees and classifier models. The construction process of random forest is described in Figure 8. In our study, the prediction process is divided into two significant parts: the first part is the growth of the decision tree, and the second part is the voting process. The growth process is divided into three categories: first is a random selection of training set, second is random forest construction, and third is split node. In the node splitting process, Gini is selected as the smallest coefficient to split the feature. The steps for calculation of coefficient Gini is given as follows:

$$G(K)1 - \sum_{i=0}^{n} .p_i^2, \tag{6}$$

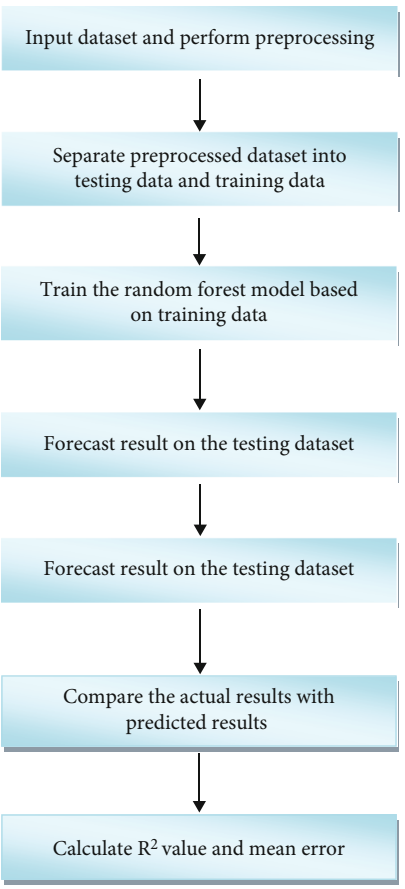

Figure 8: Flow diagram of random forest forecasting.

where p_i^2 represents the probability of category M_j in the sample set K.

$$G_{\text{split}}(k) = \frac{|M_1|}{|K|} G(M_1) + \frac{|M_2|}{|K|} G(M_2), \tag{7}$$

where $|M|$ represents the number of the sample set K and $|M_1|$ and $|M_2|$ represents the samples in subsets M_1 and M_2. Therefore, it was observed that the random forest algorithm provides better performance due to the random selection of the feature set and training set. In this study, the R^2 and mean square error for random forest were calculated using evaluation metrics [30]. The formula for calculating R^2 is given below:

$$R^2(x, \widehat{x}) = 1 - \frac{\sum_{i=0}^{K-1}(x, x\wedge_i)^2}{\sum_{i=0}^{K-1}(x_i, \bar{x})^2}. \tag{8}$$

In the above equation, x_i represents actual values, $\widehat{x}_i$ represents the predicted value, and $\bar{x}_i$ represents the average of all values. If the value R^2 is nearer to 1 then the model is good for forecasting.

The formula for calculation of root mean square (RMSE) is shown below:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{K}(x_i, x\wedge_i)^2}{K}}. \tag{9}$$

Here, $\bar{x}_i$ represents the actual value, $\widehat{x}_i$ represents the predicted value, and K indicates the number of samples, and $i = 1, 2, 3, 4 \cdots n$.

3.7. Forecasting the Trend of Positive COVID-19 Cases Using Extreme Gradient Boost (XGBOOST) Algorithm. The extreme gradient boost (XGBoost) is a widely used and most good machine learning algorithm.

It converts the weak classifier into the robust classifier. The process is repeated according to the needs of the desired model, which in this study is to forecast the positive cases of COVID-19 in Pakistan territories. XGBoost algorithm is a tree learning model which takes the decision tree as its basic unit, and the final learning model of XGBoost consists of many decision trees. It is an impaired algorithm based on a gradient boosting tree. It uses CART or linear classifier as the gradient boosting algorithm. XGBoost algorithm has several advantages for prediction problems.

(i) It supports parallelization

(ii) Used for processing of missing values

(iii) Based on the existing model it supports iteration

(iv) Provide controllable model complexity

(v) Provide high flexibility

(vi) Support shrinkage

The calculation of XGBoost is as follows:

Suppose a sample set $C = \{a_i, b_i)$: $i = 1, 2, 3, \cdots, p, a_i \in p_n, b_i \in p\}$ with sample set p and n-dimensional characteristics a model containing T decision trees can be represented by $\bar{b}_i$:

$$\bar{b}_i = \sum_{m=1}^{T} f_m(a_i), f_m \in A. \tag{10}$$

Here, A represents the function space formed by all tree models and f_m represents regression tree.

$$A = \{f(a) = R_u(a)\}\left(u : P^n \longrightarrow QR \in P^T\right). \tag{11}$$

Here, u shows the mapping relationship from a to the leaf node, and R represents the weight to the leaf node.

The objective function of the defined model is as follows:

$$D = \sum_{i=1}^{e} l\left(b_i, \bar{b}_i\right) + \sum_{t=1}^{T} \Omega(f_m). \tag{12}$$

In the above equation, b_i represents the actual value and $\bar{y}_i$ represents the forecast value where the first part is the learning loss, and the second part represents the sum of complexity of each tree; the complexity of T^{th} is indicated by

$$\Omega(f_m) = \gamma Z + \frac{1}{2}\gamma\|j\|, \tag{13}$$

where Z is the number of leaf nodes, J is the leaf weight, λ is the penalty coefficient of leaf weight, and γ is the penalty coefficient of profit function of segmented leave node.

The gradient boost strategy is used in the XGBoost algorithm to generate a regression tree after every iteration which is added to the existing model.

Assume that the forecasted of i^{th} sample in the o^{th} time of iteration $\bar{b}_i^{(o)}$ and the new added regression tree is $f_c(a_i)$ then we get

$$\begin{aligned}
\bar{b}_i^{(0)} &= 0,\\
\bar{b}_i^{(1)} &= f_1(a_i) = \bar{b}_i^{(0)} + f_1(a_i),\\
\bar{b}_i^{(2)} &= f_1(a_i) + f_2(a_i) = \bar{b}_i^{(1)} + f_2(a_i),\\
\bar{b}_i^{(o)} &= \sum_{m=1}^{0} f_m(a_i) = \bar{b}_i^{(o-1)} + f_c(a_i),
\end{aligned} \tag{14}$$

where $\bar{b}_i^{(o-1)}$ is the forecasted result of the model in $o-1$ round and $f_c(a_i)$ is the new added function in c round.

By combining equations (14), (11), and objective function, we get

$$L^{(g)} = \sum_{i=1}^{e} l\left[\left(b_i, \bar{b}_i^{(o-1)} + f_c(a_i)\right)\right] + \Omega(f_c) + C, \tag{15}$$

where C is the constant term.

Now, apply second-order Taylor expansion on the above equation then we get

$$L^{(g)} = \sum_{i=1}^{e} l\left[\left(b_i, \bar{b}_i^{(o-1)} + d_i f_c(a_i) + \frac{1}{2} e_i f_c^2(a_i)\right)\right] + \Omega(f_c) + C. \tag{16}$$

Here, $d_i = \partial_{\bar{b}_i^{(o-1)}} l(b_i, \bar{b}_i^{(o-1)})$ and $e_i = \partial^2_{\bar{b}_i^{(o-1)}} l(b_i, \bar{b}_i^{(o-1)})$ represent the first and second derivatives of the loss function and C is a constant.

By removing constants, we get

$$\tilde{L}^g \cong \sum_{i=1}^{e}\left[d_i f_c(a_i) + \frac{1}{2} e_i f_c^2(a_i)\right] + \Omega(f_c). \tag{17}$$

4. Results and Discussion

The COVID-19 virus has infected many people, and the number of infected people may increase in the future. The machine learning system approach will show promising results for the forecast of COVID-19-positive cases. Statistical models are essential techniques for evaluating infectious disease data analyses in real-time. In this research, a real-time COVID-19 forecast is built for the regions of Pakistan. Our predicted models performed very well in predicting the daily new confirmed COVID-19-positive cases in the regions of Pakistan. All the steps involved in building the proposed model are implemented in python using Pandas library for data loading and preprocessing of data Matplotlib is used to plot the curves, and Scikit-learn library is also used for implementation of the classifier. This research's experiments are executed on a system with a Dell i7 processor with 64 GB RAM. For further evaluation, metrics (accuracy, precision, support, recall, $F1$-score, and sensitivity) are used to measure the quality of machine learning models. We have proposed a prediction model that works over six months to predict COVID-19 activity by combining the previous incidence of COVID-19. Our proposed model performed well for all regions of Pakistan. The performance of algorithms was evaluated by using mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) [15, 31], and evaluation metric. This proposed model has several advantages compared to other reported works on a similar topic. Our forecasting models performed well for the COVID-19 forecast, but random forest and XGBoost provide better accuracy. We have used a large amount of data which improved the performance of all ML models. Figures 9–14 show results of the COVID-19 trend using linear regression in regions of Pakistan. The red bars are the training data, whereas the blue is the predicted trend with indicated model scores. If the blue bar is increasing, it means positive cases are increasing day by day. In Figure 9, red bars represent the actual COVID-19-positive cases data in Sindh, Pakistan, whereas blue bars represent the predicted COVID-19-positive cases. According to prediction, this figure shows that Sindh may have a higher number of positive cases in May. In Table 1, error metrics shows that the MSE score for Sindh is 0.202, MAE is 2.024, RMSE is 4.859, and MAPE score is 0.011. Furthermore, in Table 2, the linear regression model is evaluated using evaluation metrics on the Sindh region, whose accuracy percentage is 86%, support percentage is 28%, precision percentage is 82%, recall percentage is 1%, $F1$-score percentage is 82%, and sensitivity percentage is 1%. Figure 10 represents the forecast prediction of Punjab, Pakistan. According to prediction, this figure shows that Punjab may have a higher number of positive cases in May. In Table 1, error metrics shows that the MSE score for Punjab is 0.202, MAE is 2.024, RMSE is 4.859, and MAPE score is 0.011. Furthermore, in Table 2, the linear regression model is evaluated using evaluation metrics on the Punjab region, whose accuracy percentage is 82%, support percentage is 27%, precision percentage is 72%, recall percentage is 1%, $F1$-score percentage is 83%, and sensitivity percentage is 1%. Figure 11 represents the forecast prediction of Gilgit Baltistan, Pakistan. According to prediction, this figure shows that in January and February, Gilgit Baltistan may have a higher number of positive cases than cases are slowly decreasing. In Table 1, error metrics shows that the MSE score for Gilgit Baltistan is 0.202, MAE is 2.024, RMSE is 4.859, and MAPE score is 0.011. Furthermore, the linear regression model is evaluated in Table 2. By using evaluation metrics on the Gilgit Baltistan region, it shows accuracy percentage is 84%, support percentage is 94%, precision percentage is 84%, recall percentage is 1%, $F1$-score percentage is 96%, and sensitivity percentage is 1%. Figure 12 represents the forecast prediction of Khyber Pakhtunkhwa, Pakistan. In Table 1, error metrics shows that the MSE score for Khyber Pakhtunkhwa is 0.202, MAE is 2.024, RMSE is 4.859, and MAPE score is 0.011. Furthermore, in Table 2, the linear regression model is evaluated using evaluation metrics on the Khyber Pakhtunkhwa region, whose accuracy percentage is 86%, support percentage is 27%, precision percentage is 76%, recall percentage is 1%, $F1$-score percentage is 82%, and sensitivity percentage is 1%. Figure 13 represents the forecast prediction of Balochistan, Pakistan. In Table 1, error metrics shows that the MSE score for Balochistan is 0.202, MAE is 2.025, RMSE is 4.860, and MAPE score is 0.011. Furthermore, in Table 2, the linear regression model is evaluated by using evaluation metrics on the Balochistan region, whose accuracy percentage is 82%, support percentage is 28%, precision percentage is 71%, recall percentage is 1%, $F1$-score percentage is 82%, and sensitivity percentage is 1%. Figure 14 represents the forecast prediction of Azad Jammu And Kashmir, Pakistan. In Table 1, error metrics shows that the MSE score for Azad Jammu And Kashmir is 0.202, MAE is 2.024, RMSE is 4.859, and MAPE score is 0.011. Furthermore, in Table 2, the linear regression model is evaluated by using evaluation metrics on Azad Jammu And Kashmir region whose accuracy percentage is 74%, support percentage is 30%, precision percentage is 5%, recall percentage is 1%, $F1$-score percentage is 83%, and sensitivity percentage is 1%.

By using the above random forest methodology, a visualization of records in terms of actual versus predicted values

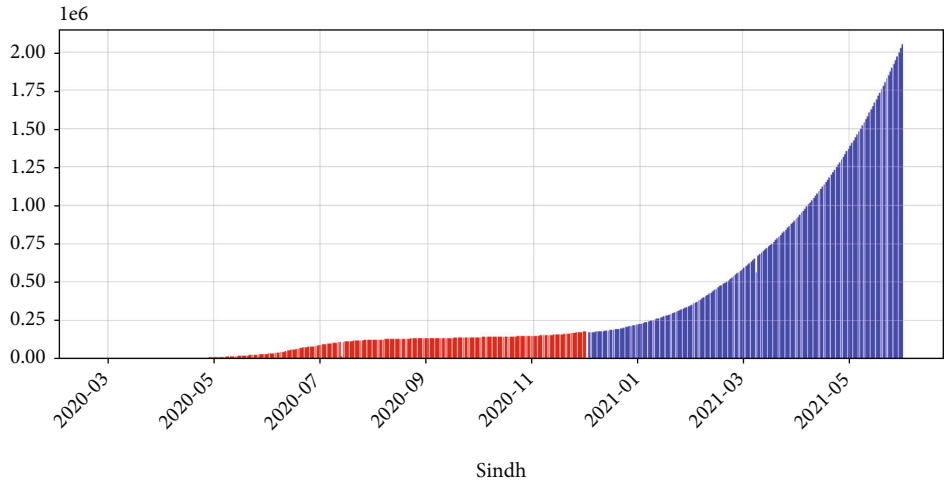

Figure 9: Actual COVID-19 cases and forecasted COVID-19-positive cases in Sindh by employing the linear regression model.

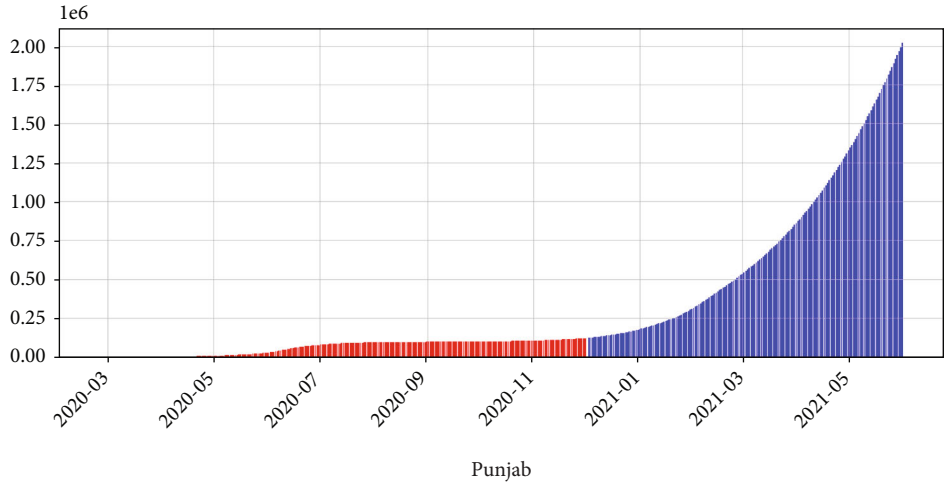

Figure 10: Actual COVID-19 cases and forecasted COVID-19-positive cases in Punjab by employing the linear regression model.

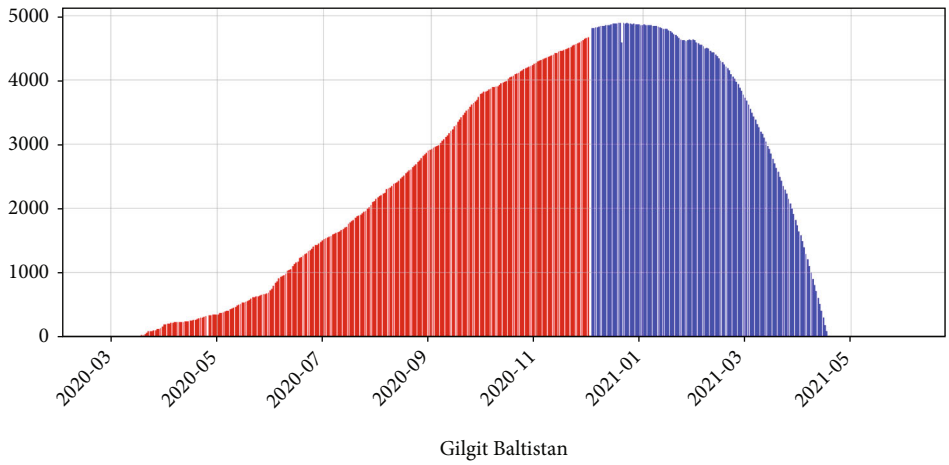

Figure 11: Actual COVID-19 cases and forecasted COVID-19-positive cases in GB by employing the linear regression model.

have shown below in graphs. Figures 15–20 show results of the COVID-19 trend using random forest in regions of Pakistan. The red bars are the training data, whereas the blue is the predicted trend with indicated model scores. In Figure 15, red bars represent the actual COVID-19-positive cases data in Sindh, Pakistan, whereas blue bars represent

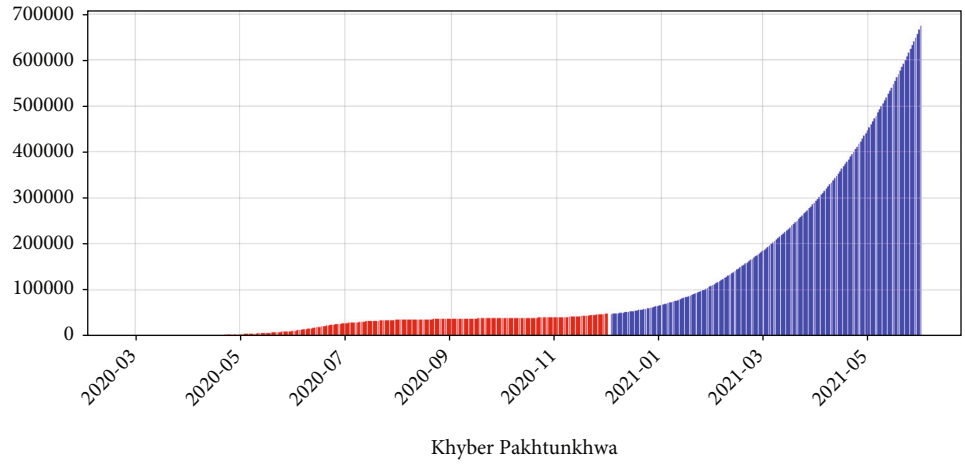

Figure 12: Actual COVID-19 cases forecasted COVID-19-positive cases in KPK by employing the linear regression model.

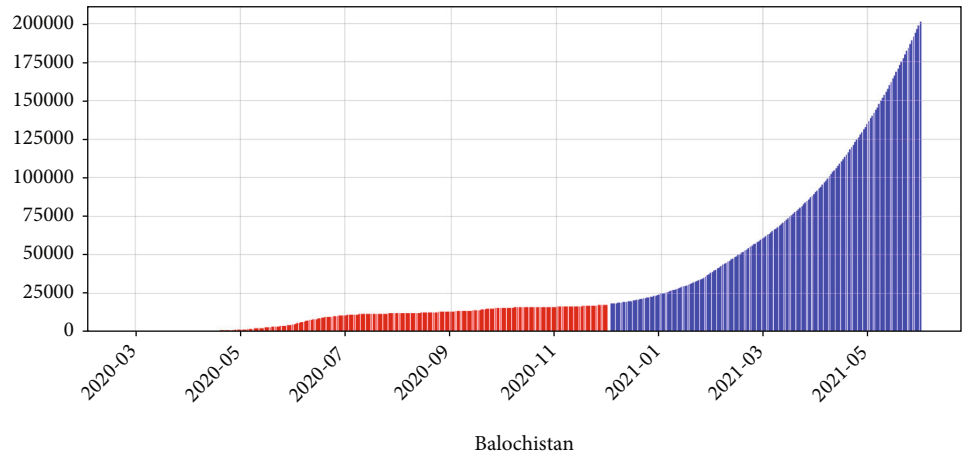

Figure 13: Actual COVID-19 cases forecasted COVID-19-positive cases in Balochistan by employing the linear regression model.

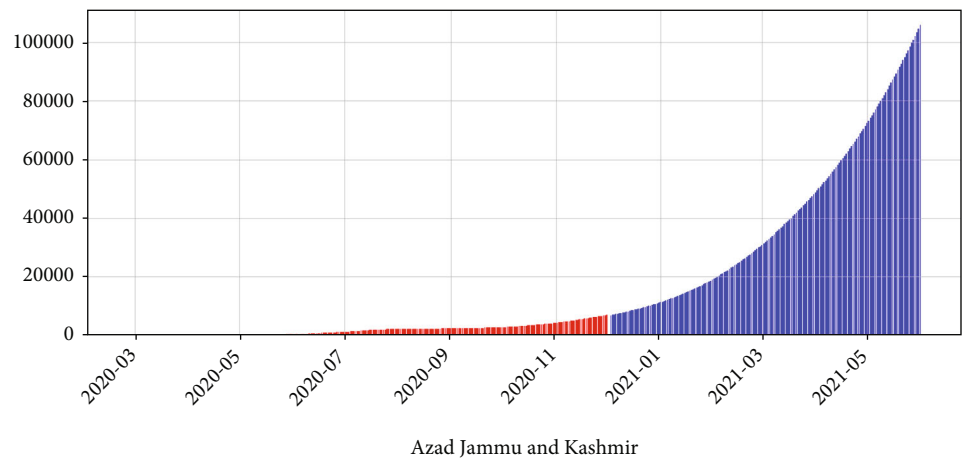

Figure 14: Actual COVID-19 cases and forecasted COVID-19-positive cases in AJK by employing the linear regression model.

the predicted COVID-19-positive cases. According to prediction, this figure shows that Sindh may have a higher number of positive cases in May. In Table 3, error metrics shows that the MSE score for Sindh is 0.006, MAE is 2.035, RMSE is 3.389, and MAPE score is 0.006. Furthermore, in Table 4, the random forest model is evaluated by using evaluation metrics on Sindh region whose accuracy percentage is 93%, support percentage is 136%, precision

Table 1: Performance of the linear regression algorithm.

Method	Error metrics	Sindh	Punjab	KPK	Gilgit Baltistan	Balochistan	AJK
Linear regression	MSE	0.202	0.202	0.202	0.202	0.202	0.202
	MAE	2.024	2.024	2.024	2.024	2.025	2.024
	RMSE	4.859	4.859	4.859	4.859	4.860	4.859
	MAPE	0.011	0.011	0.011	0.011	0.011	0.011

Table 2: Evaluation metrics for the linear regression algorithm.

Method	Evaluation metrics	Sindh	Punjab	KPK	Gilgit Baltistan	Balochistan	AJK
Linear regression	Accuracy	86%	82%	86%	84%	82%	74%
	Support	28%	275	27%	94%	28%	30%
	Precision	82%	72%	76%	84%	71%	5%
	Recall	1%	1%	1%	1%	1%	1%
	*F*1-score	82%	83%	82%	96%	82%	83%
	Sensitivity	1%	1%	1%	1%	1%	1%

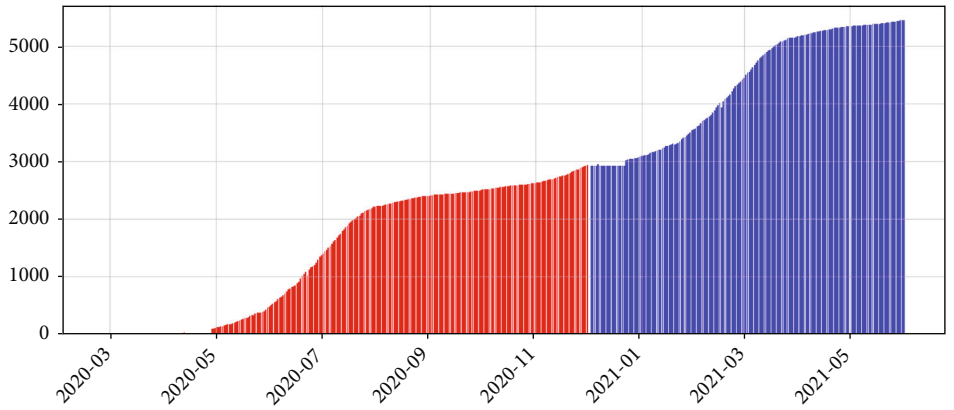

Figure 15: Actual COVID-19 cases and forecasted COVID-19-positive cases in Sindh by employing the random forest model.

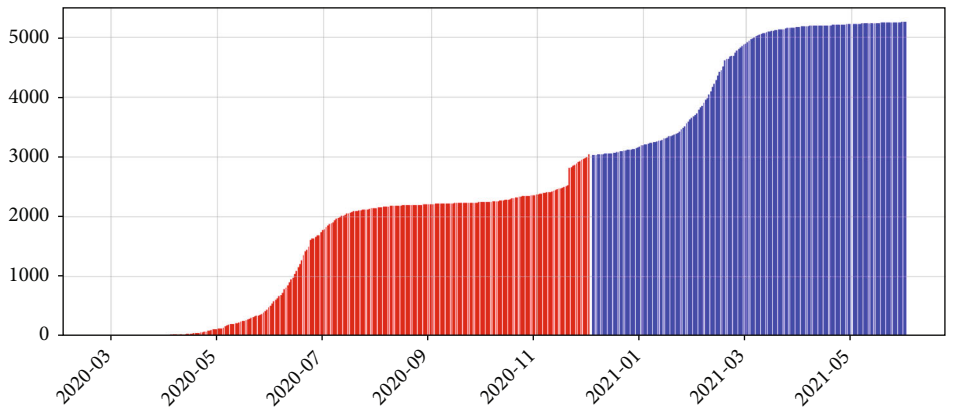

Figure 16: Actual COVID-19 cases and forecasted COVID-19-positive cases in Punjab by using the random forest model.

percentage is 84%, recall percentage is 82%, *F*1-score percentage is 90%, and sensitivity percentage is 92%. Figure 16 shows that in March, April, and May, Punjab may have a higher number of positive cases. In Table 3, error metrics shows that the MSE score for Punjab is 0.149, MAE is 2.035, RMSE is 3.389, and MAPE score is 0.006. Furthermore, in Table 4, the random Forest model is evaluated by using evaluation metrics on Punjab region whose accuracy

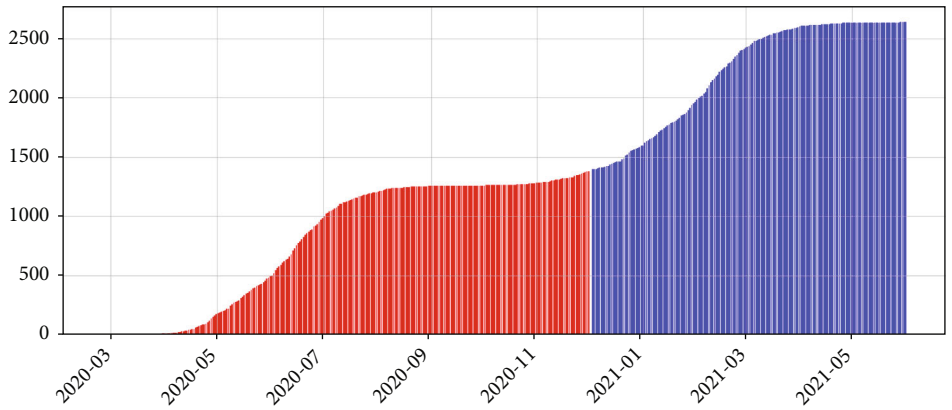

Figure 17: Actual COVID-19 cases and forecasted COVID-19-positive cases in KPK by using the random forest model.

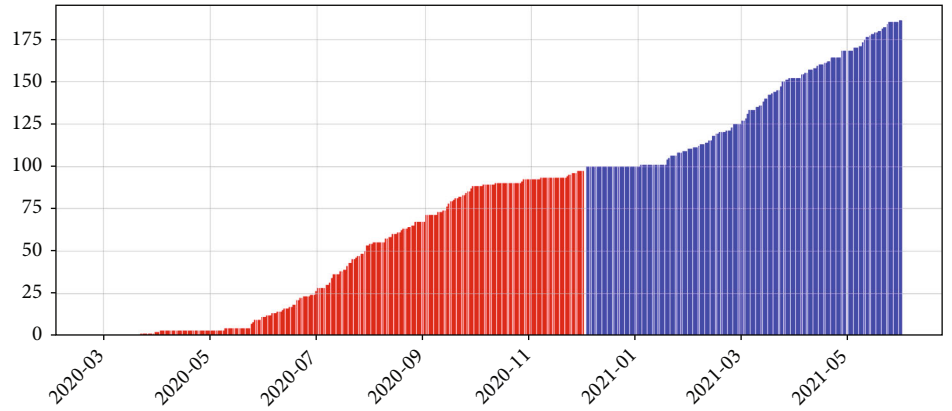

Figure 18: Actual COVID-19 cases and forecasted COVID-19-positive cases in GB by using the random forest model.

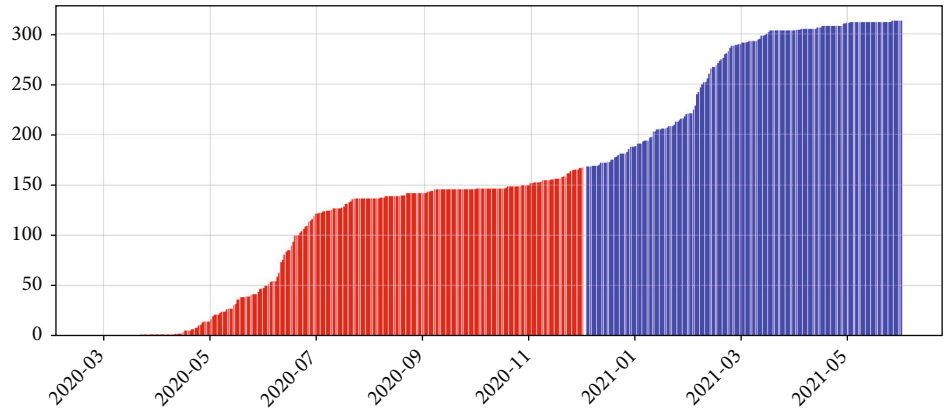

Figure 19: Actual COVID-19 cases and forecasted COVID-19-positive cases in Balochistan by using the random forest model.

percentage is 93%, support percentage is 154%, precision percentage is 85%, recall percentage is 75%, *F*1-score percentage is 88%, and sensitivity percentage is 92%. Figure 17 represents the forecast of Khyber Pakhtunkhwa, and in April and May, Khyber Pakhtunkhwa may have a higher number of Positive cases. In Table 3, error metrics shows that the MSE score for Khyber Pakhtunkhwa is 0.022, MAE is 2.035, RMSE is 3.389, and MAPE score is 0.006. Furthermore, in Table 4, the random forest model is evaluated using evaluation metrics on Khyber Pakhtunkhwa region whose accuracy percentage is 93%, support percentage is 154%, precision percentage is 84%, recall percentage is 84%, *F*1 -score percentage is 89%, and sensitivity percentage is 92%. Figure 18 represents the forecast of Gilgit Baltistan. In Table 3, error metrics shows that the MSE score for Gilgit Baltistan is 0.002, MAE is 2.035, RMSE is 3.389, and MAPE score is 0.006. Furthermore, in Table 4, the random forest model is evaluated using evaluation metrics on the Gilgit Baltistan region, whose accuracy percentage is 95%, support percentage is 117%, precision percentage is 90%, recall

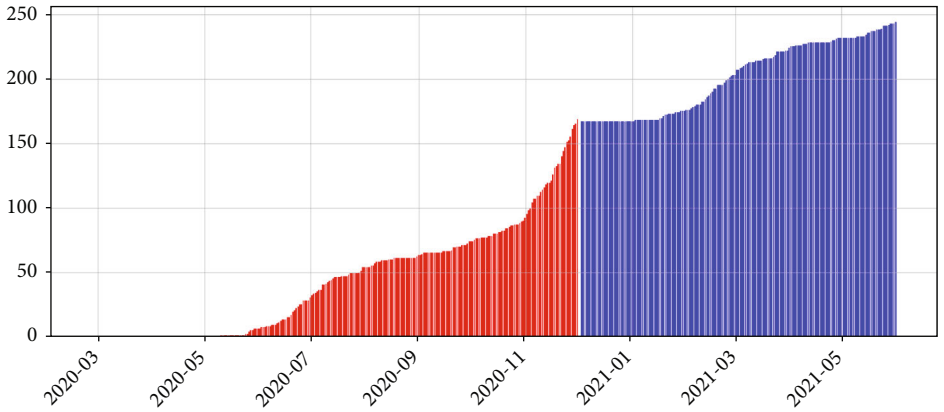

Figure 20: Actual COVID-19 cases and forecasted COVID-19-positive cases in AJK by using the random forest model.

Table 3: Performance of the random forest algorithm.

Method	Error metrics	Sindh	Punjab	KPK	Gilgit Baltistan	Balochistan	AJK
Random forest	MSE	0.006	0.149	0.022	0.112	0.013	0.126
	MAE	2.035	2.035	2.035	2.035	2.035	2.030
	RMSE	3.389	3.389	3.389	3.389	3.389	3.389
	MAPE	0.006	0.006	0.006	0.006	0.006	0.006

Table 4: Evaluation metrics for the random forest algorithm.

Method	Evaluation metrics	Sindh	Punjab	KPK	Gilgit Baltistan	Balochistan	AJK
Random forest	Accuracy	93%	93%	93%	95%	93%	93%
	Support	136%	154%	154%	117%	156%	181%
	Precision	85%	85%	84%	90%	92%	85%
	Recall	82%	79%	84%	76%	79%	74%
	*F*1-score	90%	88%	89%	92%	86%	85%
	Sensitivity	92%	92%	92%	90%	92%	92%

percentage is 76%, *F*1-score percentage is 92%, and sensitivity percentage is 90%. Figure 19 represents the forecast of Balochistan, and blue bars mean that in April, May, and June, Balochistan May have a higher number of COVID-19-positive cases. In Table 3, error metrics shows that the MSE score for Balochistan is 0.013, MAE is 2.035, RMSE is 3.389, and MAPE score is 0.006. Furthermore, in Table 4, the random forest model is evaluated by using evaluation metrics on Balochistan region whose accuracy percentage is 93%, support percentage is 156%, precision percentage is 92%, recall percentage is 79%, *F*1-score percentage is 86%, and sensitivity percentage is 92%. Figure 20 represents the forecast of Azad Jammu And Kashmir forecast, and blue bars represent that in April, May, and June, Azad Jammu And Kashmir May have a higher number of COVID-19-positive cases. In Table 3, error metrics shows that the MSE score for Azad Jammu And Kashmir is 0.126, MAE is 2.030, RMSE is 3.389, and MAPE score is 0.006. Furthermore, in Table 4, the random forest model is evaluated by using evaluation metrics on Azad Jammu And Kashmir region whose accuracy percentage is 93%, support percentage is 181%, precision percentage is 85%, recall percentage is 74%, *F*1-score percentage is 85%, and sensitivity percentage is 92%.

(I) Sindh region

(II) Punjab region

(III) Khyber Pakhtunkhwa region

(IV) Gilgit Baltistan region

(V) Balochistan region

(VI) Azad Jammu And Kashmir region

Using the above XGBoost methodology, a visualization of records in terms of actual versus predicted values is shown below in graphs. Figures 21–26 show results of the COVID-19 trend using the XGBoost model in regions of Pakistan. The red bars are the training data, whereas the blue is the predicted trend. In Figure 21, red bars represent the actual COVID-19-positive cases data in Sindh, Pakistan, whereas blue bars represent the predicted COVID-19-positive cases. According to prediction, this figure shows that in May, Sindh may have a higher number of positive cases. In Table 5 Error Metrics shows MSE score for Sindh is 0.074, MAE is 0.579, RMSE is 1.389, and MAPE score is 0.003. Figure 22 shows that in April and May, Punjab may have a higher number of positive cases. In Table 5, error metrics shows that the MSE score for Punjab is 0.394, MAE is 1.332, RMSE is 3.17, and MAPE score is 0.007. Figure 23 represents the forecast of Balochistan. In Table 5, error metrics shows that the MSE score for Balochistan is 0.304, MAE is 1.169, RMSE is 2.807, and MAPE score is 0.006. Figure 24 represents the forecast of Khyber Pakhtunkhwa. In Table 5, error metrics shows that the MSE score for Khyber Pakhtunkhwa is 0.198, MAE is 0.836, RMSE is 2.008, and MAPE score is 0.004. Figure 25 represents the forecast of Gilgit Baltistan. In Table 5, error metrics shows that the MSE score for Gilgit Baltistan is 0.049, MAE is 0.944, RMSE is 2.266, and MAPE score is 0.005. Figure 26 represents the forecast of Azad Jammu And Kashmir. In Table 5, error metrics shows that the MSE score for Azad Jammu And Kashmir is 0.049, MAE is 0.472, RMSE is 1.135, and MAPE score is 0.002.

4.1. Comparative Analysis. Linear regression, random forest, and XGBoost algorithms are used to predict COVID-19 cases, and it is observed that the random forest algorithm is better than linear regression. The random forest provides high accuracy for the prediction of positive COVID-19 cases in Pakistan. To compare the performance of linear regression, XGBoost, and random forest estimation method, mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are used [32, 33].

4.2. Evaluation Metrics. Since it is an inevitable prediction [34], the accuracy of all algorithms is checked. To identify the model with the best prediction power, we considered six evaluation metrics, including accuracy, precision, sensitivity, recall, support, and *F*1-score [35, 36]. Tables 2 and 4 show the performance results of machine learning algorithms for regions of Pakistan for our proposed model. It is observed that the linear regression and random forest show comparable results. Random forest has comparably better performance than linear regression. However, this paper also proposes using the XGBoost algorithm, which performs better than both ML algorithms.

4.2.1. Accuracy. It is used to check the performance of linear regression and random forest. This study would equate to the correct number of positive cases over the total predictions made by both models.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{18}$$

4.2.2. Precision. It is the ratio of TP (true positive) samples with the sum of false positive (FP) and TP (true positive). It is used to classify the total COVID-19-positive cases by using the Pakistan dataset.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{19}$$

4.2.3. Recall. It is the fraction of TP (true positive) samples with the sum of false negative and true positive.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{20}$$

4.2.4. F1-Score. It is the mean of precision and recall value. It provides a balance between recall and precision by evaluating linear regression and random forest model performance in the classification of COVID-19 patients.

$$F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recal}}. \tag{21}$$

4.2.5. Sensitivity. It is the rate of TP (true positive). It measures the proportion of true positives (TP). In our study, a true positive would be the prediction of positive COVID-19 cases.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{22}$$

4.3. Correlation. It is used to measure the interrelation between two variables and also the direction of their relationship. The value of correlation is always greater than -1 and less than +1. If the coefficient reaches point 0, then the relationship between variables becomes weak. In correlation positive (+) sign indicates a positive relationship between variables, and a negative (-) sign indicates a negative relationship. There are several types of correlation: point-biserial correlation, Kendall rank, Spearman correlation, and Pearson correlation [37, 38].

4.3.1. Pearson Correlation. Through Pearson correlation, we can measure the relationship between linearly related variables. It is a widely used correlation. In this type of correlation, when variables whose correlation is to be found are supposed to be normalized, if they are not normalized, then the first normalization should be performed [39]. The relationship between two variables must be straight, assuming that data is equally distributed about the regression line. Correlation between dataset features provides detailed information about features and the ratio of influence that they have on the target value. The heat map of Pearson correlation between the features of the dataset is shown in Figure 27. It revealed a stronger positive correlation between new positive cases and hospitalized with symptoms. There is also a strong correlation between total cases and deaths. Correlation in Figure 28 reveals a stronger positive

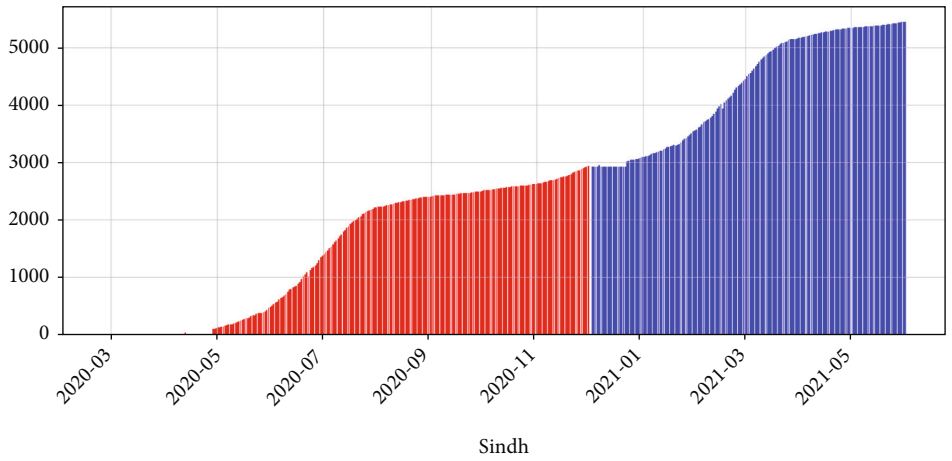

Figure 21: Actual COVID-19 cases and forecasted COVID-19-positive cases in Sindh by using the XGBoost model.

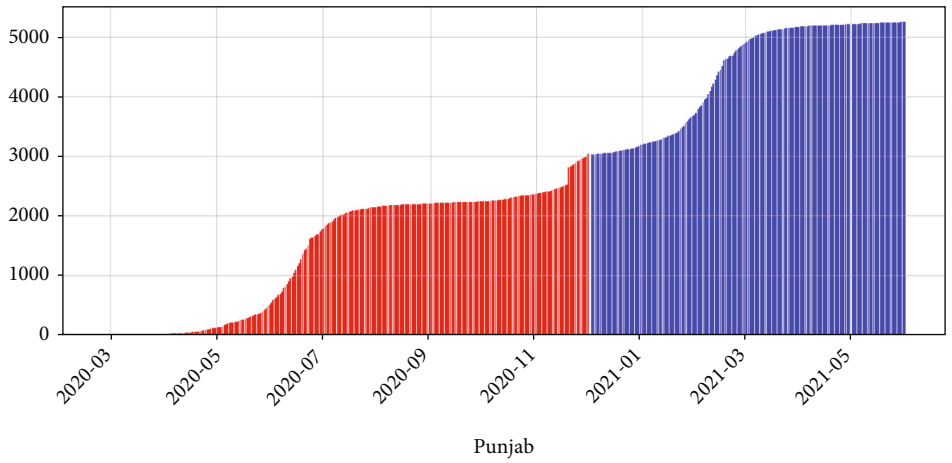

Figure 22: Actual COVID-19 cases and forecasted COVID-19-positive cases in Punjab by using the XGBoost model.

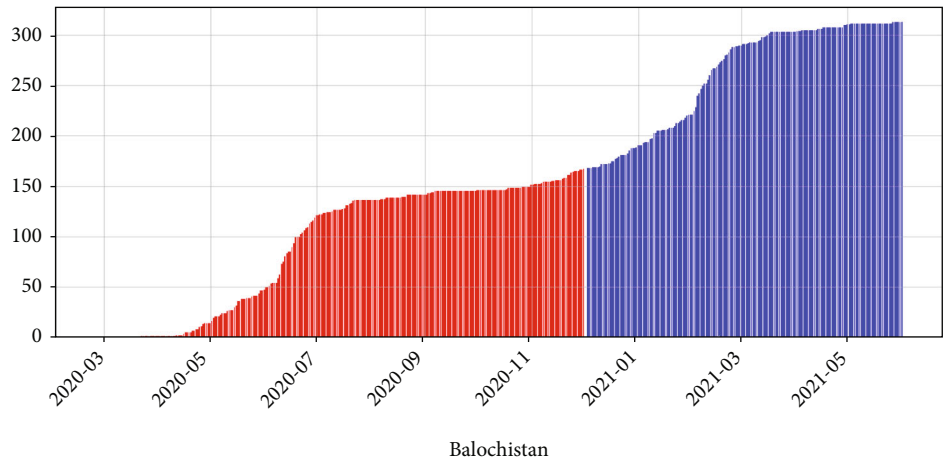

Figure 23: Actual COVID-19 cases and forecasted COVID-19-positive cases in Balochistan by using the XGBoost model.

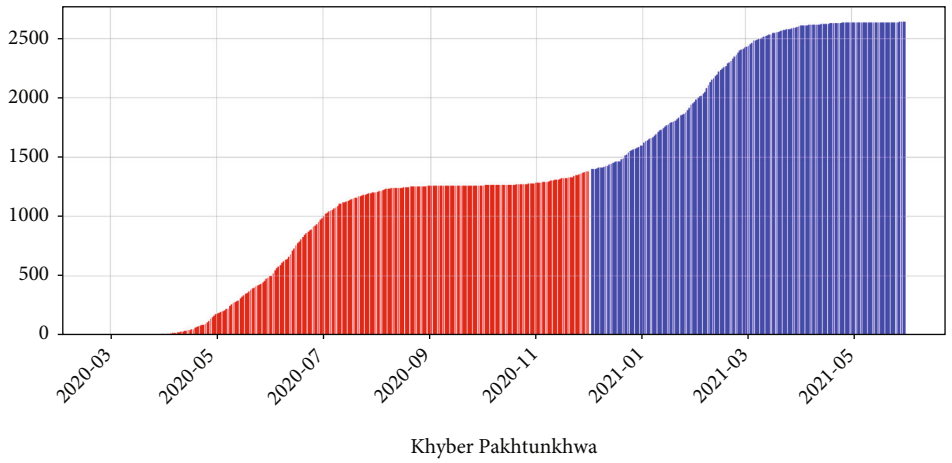

Figure 24: Actual COVID-19 cases and forecasted COVID-19 positive cases in KPK by using XGBoost Model.

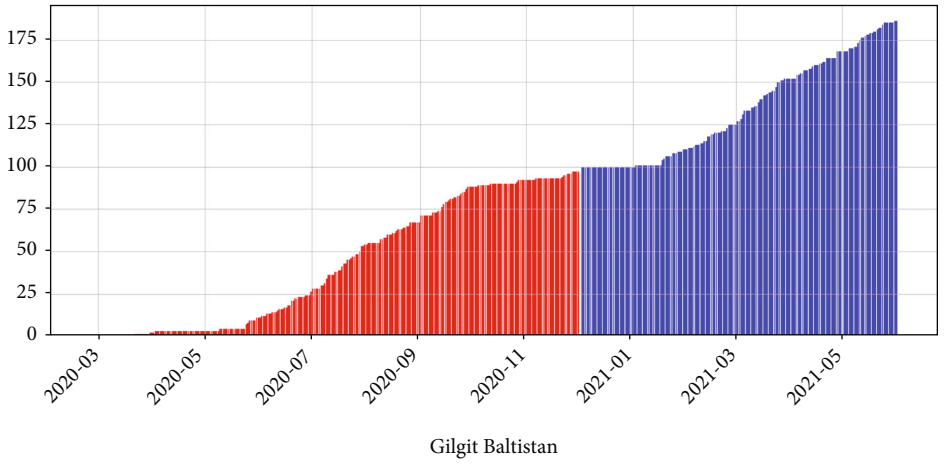

Figure 25: Actual COVID-19 cases and forecasted COVID-19-positive cases in GB by using the XGBoost model.

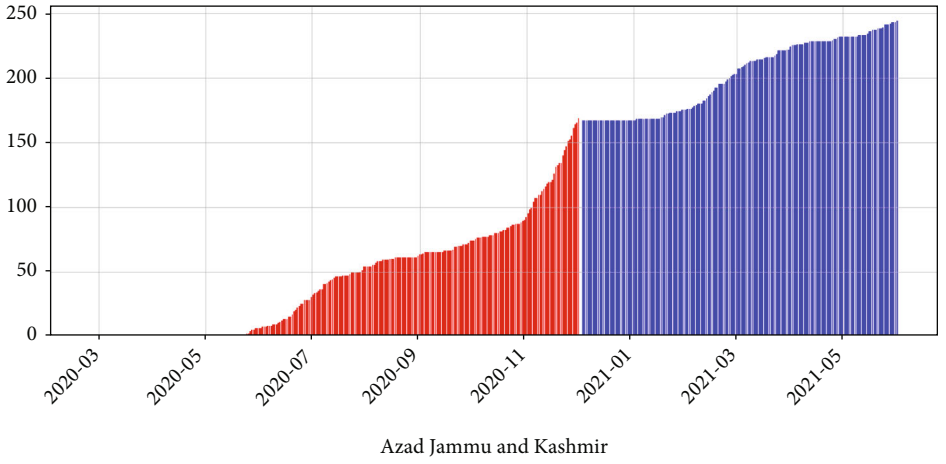

Figure 26: Actual COVID-19 cases and forecasted COVID-19-positive cases in AJK by using the XGBoost model.

Table 5: Performance of the XGBoost algorithm.

Method	Error metrics	Sindh	Punjab	KPK	Gilgit Baltistan	Balochistan	AJK
XGBoost	MSE	0.074	0.394	0.198	0.049	0.304	0.049
	MAE	0.579	1.332	0.836	0.944	1.169	0.472
	RMSE	1.389	3.197	2.008	2.266	2.807	1.135
	MAPE	0.003	0.007	0.004	0.005	0.006	0.002

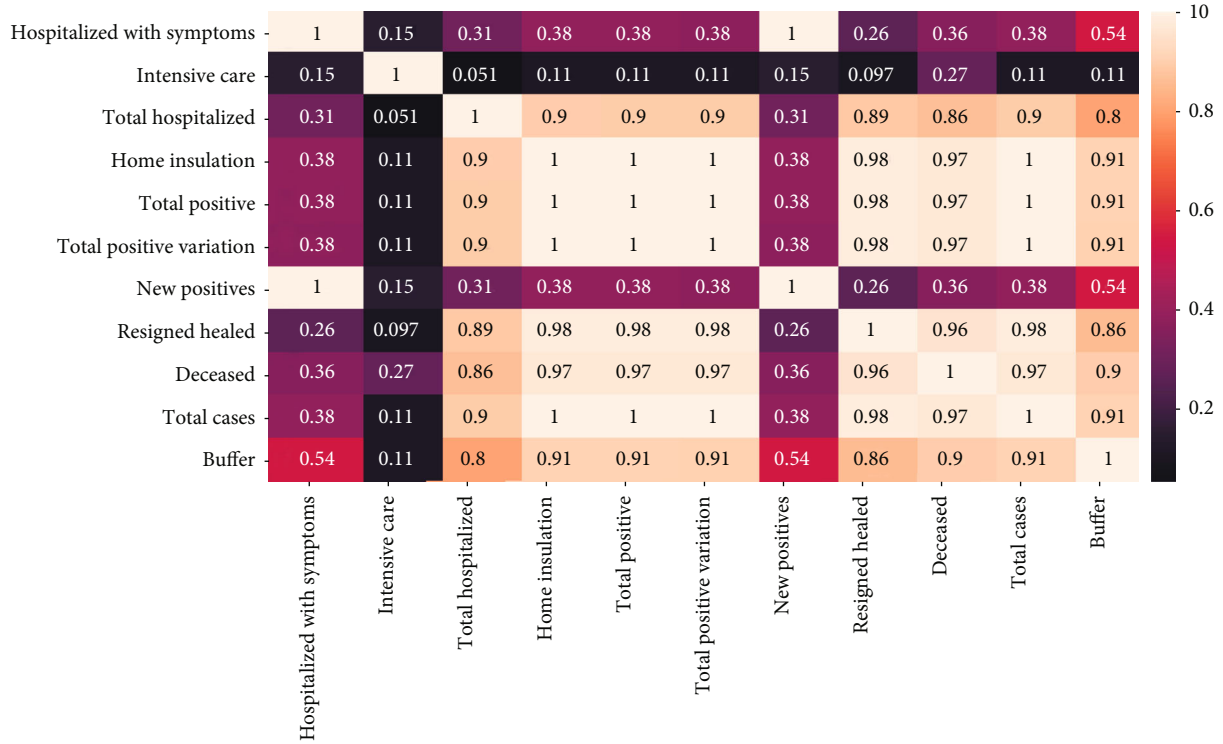

Figure 27: Correlation between data features.

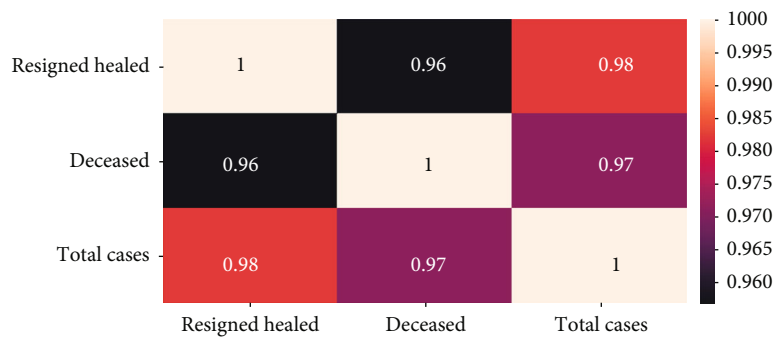

Figure 28: Correlation between data features.

correlation between new positive cases and recoveries, and there is also a strong correlation between total cases and total recoveries.

5. Conclusion

This deadly virus has killed many people all around the world. It is a dangerous disease that transfers from one human to another, and it creates severe damage to the lungs. In this paper, we have proposed machine learning methods for forecasting COVID-19-positive cases in Pakistan regions. Random forest, XGBoost, and linear regression algorithms were used as prediction models. After evaluating these algorithms, it is identified that the random forest and XGBoost algorithm provide better accuracy than linear regression. Random forest and XGBoost algorithms provide a high

prediction rate. The evaluation results of this proposed model prove that using variables as predictors can lead us to high forecasting accuracy. These predictions will be helpful for researchers, government authorities, and health industry planners to manage services and arrange medical infrastructure accordingly. Additionally, the correlation matrix reveals that positive COVID-19 patients and hospitalized patients have a robust correlation. This proposed model is also helpful for other countries for forecasting COVID-19-positive cases.

In the future, this model can be extended to implement various other ML algorithms and prediction methodologies.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] M. D. P. Kavadi, R. Patan, M. Ramachandran, and A. H. Gandomi, "Partial derivative nonlinear global pandemic machine learning prediction of COVID 19," *Chaos, Solitons & Fractals*, vol. 139, no. 2020, article 110056, 2020.

[2] C. Bayes and L. Valdivieso, "Modelling death rates due to COVID-19: a Bayesian approach," 2020, http://arxiv.org/abs/2004.02386.

[3] M. R. Nemati, J. Ansary, and N. Nemati, *Machine-learning approaches in COVID-19 survival analysis and discharge-time likelihood prediction using clinical Data*, 2020, https://www.sciencedirect.com/science/article/pii/S2666389920300945.

[4] S. F. Ardabili, A. Mosavi, P. Ghamisi et al., "COVID-19 outbreak prediction with machine learning," *Algorithms*, vol. 13, no. 10, p. 249, 2020.

[5] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review," *Chaos, Solitons & Fractals*, vol. 139, article 110059, 2020.

[6] A. U. Mandayam, A. C. Rakshith, S. Siddesha, and S. K. Niranjan, "Prediction of Covid-19 pandemic based on regression," in *2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, Bangalore, India, 2020, November.

[7] F. Rustam, A. A. Reshi, A. Mehmood et al., "COVID-19 future forecasting using supervised machine learning models," *IEEE Access*, vol. 8, pp. 101489–101499, 2020.

[8] G. Pandey, P. Chaudhary, R. Gupta, and S. Pal, "SEIR and regression model based COVID-19 outbreak predictions in India," 2020, http://arxiv.org/abs/2004.00958.

[9] Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *Npj digital medicine*, vol. 4, no. 1, pp. 1–5, 2021.

[10] A. N. Roy, J. Jose, A. Sunil, N. Gautam, D. Nathalia, and A. Suresh, "Prediction and spread visualization of Covid-19 pandemic using machine learning," *Preprints*, article 2020050147, 2020.

[11] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R. N. Kalhori, "Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study," *JMIR public health and surveillance*, vol. 6, no. 2, article e18828, 2020.

[12] L. A. Amar, A. A. Taha, and M. Y. Mohamed, "Prediction of the final size for COVID-19 epidemic using machine learning: a case study of Egypt," *Infectious Disease Modelling*, vol. 5, pp. 622–634, 2020.

[13] H. Tyralis and G. Papacharalampous, "Variable selection in time series forecasting using random forests," *Algorithms*, vol. 10, no. 4, p. 114, 2017.

[14] *COVID-19 Government of Pakistan Public Dataset*, 2021, https://covid.gov.pk/.

[15] *Kaggle Machine Learning Dataset*, 2021, https://www.kaggle.com/.

[16] S. S. Aljameel, I. U. Khan, N. Aslam, M. Aljabri, and E. S. Alsulmi, "Machine learning-based model to predict the disease severity and outcome in COVID-19 patients," *Scientific programming*, vol. 2021, Article ID 5587188, 2021.

[17] *COVID-19 World Health Organization Pakistan Related Public Dataset*, 2021, https://www.who.int/emergencies/diseases/novel-coronavirus-2019?gclid=eaiaiqobchmi-lzega3l8aivbj3ch1etq8yeaayasaaegilj_d_bwe.

[18] Z. Ahmad, M. Arif, F. Ali, I. Khan, and K. S. Nisar, "A report on COVID-19 epidemic in Pakistan using SEIR fractional model," *Scientific Reports*, vol. 10, no. 1, pp. 1–14, 2020.

[19] S. Degadwala, B. Patel, and D. Vyas, "A review on Indian state/City Covid-19 cases outbreak forecast utilizing machine learning models," in *In 2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pp. 1001–1005, India, 2021, January.

[20] Y. Li, M. A. Horowitz, J. Liu et al., "Individual-level fatality prediction of COVID-19 patients using AI methods," *Frontiers in Public Health*, vol. 8, p. 566, 2020.

[21] Y. Y. Cheng, P. P. Chan, and Z. W. Qiu, "Random forest based ensemble system for short term load forecasting," in *2012 International Conference on Machine Learning and Cybernetics*, Xi'an, China, 2012, JulyIEEE.

[22] M. A. Zaki, S. Narejo, S. Zai, U. Zaki, Z. Altaf, and N. Din, "Detection of nCoV-19 from hybrid dataset of CXR images using deep convolutional neural network," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, pp. 699–707, 2020.

[23] K. B. Song, Y. S. Baek, D. H. Hong, and G. Jang, "Short-term load forecasting for the holidays using fuzzy linear regression method," *IEEE Transactions on Power Systems*, vol. 20, no. 1, pp. 96–101, 2005.

[24] S. Rath, A. Tripathy, and A. R. Tripathy, "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 5, pp. 1467–1474, 2020.

[25] S. Ballı, "Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods," *Chaos, Solitons & Fractals*, vol. 142, article 110512, 2021.

[26] C. Sohrabi, Z. Alsafi, N. O'Neill et al., "World Health Organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19)," *International Journal of Surgery*, vol. 76, pp. 71–76, 2020.

[27] N. Jain, S. Jhunthra, H. Garg et al., "Prediction modelling of COVID using machine learning methods from B-cell dataset," *Results in Physics*, vol. 21, article 103813, 2021.

[28] S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, "Confusion matrix-based feature selection," *MAICS*, vol. 710, pp. 120–127, 2011.

[29] M. U. Rehman, A. Shafique, S. Khalid, M. Driss, and S. Rubaiee, "Future forecasting of COVID-19: a supervised learning approach," *Sensors*, vol. 21, no. 10, p. 3322, 2021.

[30] C. Iwendi, A. K. Bashir, A. Peshkar et al., "COVID-19 patient health prediction using boosted random forest algorithm," *Frontiers in Public Health*, vol. 8, p. 357, 2020.

[31] R. Sujath, J. M. Chatterjee, and A. E. Hassanien, "A machine learning forecasting model for COVID-19 pandemic in India," *Stochastic Environmental Research and Risk Assessment*, vol. 34, pp. 959–972, 2020.

[32] S. Namasudra, S. Dhamodharavadhani, and R. Rathipriya, "Nonlinear neural network based forecasting model for predicting COVID-19 cases," in *Neural processing letters*, pp. 1–21, Springer, 2021.

[33] I. F. Siddiqui, N. M. F. Qureshi, B. S. Chowdhry, and M. A. Uqaili, "Pseudo-cache-based IoT small files management framework in HDFS cluster," *Wireless Personal Communications*, vol. 113, no. 3, pp. 1495–1522, 2020.

[34] A. A. Khan, F. S. Lodhi, U. Rabbani et al., "Impact of coronavirus disease (COVID-19) pandemic on psychological well-being of the Pakistani general population," *Frontiers in Psychiatry*, vol. 11, 2020.

[35] F. Khan, A. Saeed, and S. Ali, "Modelling and forecasting of new cases, deaths and recover cases of COVID-19 by using vector autoregressive model in Pakistan," *Chaos, Solitons & Fractals*, vol. 140, article 110189, 2020.

[36] M. Yadav, M. Perumal, and M. Srinivas, "Analysis on novel coronavirus (COVID-19) using machine learning methods," *Chaos, Solitons & Fractals*, vol. 139, p. 110050, 2020.

[37] S. K. Bandyopadhyay and S. Dutta, *Machine learning approach for confirmation of covid-19 cases: positive, negative, death and release*, medRxiv, 2020.

[38] E. Shahid and Q. A. Arain, "Indoor positioning: "an image-based crowdsource machine learning approach"," *Multimedia Tools and Application*, vol. 80, no. 17, pp. 26213–26235, 2021.

[39] N. M. F. Qureshi, I. F. Siddiqui, A. Abbas et al., "Stream-based authentication strategy using IoT sensor data in multi-homing sub-aqueous big data network," *Wireless Personal Communications*, vol. 116, no. 2, pp. 1217–1229, 2020.